AFRILEX-reeks
series 8:1998

LEXIKOS 8

BURO VAN DIE WAT

AFRILEX

Lexikos 8

Redakteur

Editor

J.C.M.D. du Plessis

Resensieredakteur

Review Editor

E. Botha

African Association for Lexicography

AFRILEX-REEKS 8:1998

AFRILEX SERIES 8:1998

BURO VAN DIE WAT

STELLENBOSCH

Uitgewer Publisher

BURO VAN DIE WAT
Posbus 245
7599 STELLENBOSCH

Eerste uitgawe 1998

Tipografie en uitleg
deur Tanja Harteveld en Hermien van der Westhuizen
Bandontwerp deur Piet Grobler

Geset in 10 op 12 pt Palatino

Gedruk en gebind deur Rotapress
Stewartstraat 59 Goodwood

ISBN 0 9584120 4 9

Menings wat in artikels en resensies uitgespreek word, is nie noodwendig dié van AFRILEX of die Buro van die WAT nie

Opinions expressed in the articles and reviews are not necessarily those of AFRILEX or of the Bureau of the WAT

Adviesraad / Advisory Board

Redaksiekomitee / Editorial Committee

Inhoud / Contents

Beskouende artikels / Contemplative Articles

Projekte / Projects

Leksikovaria / Lexicovaria

Verslae / Reports

Resensieartikels / Review Articles

Resensies / Reviews

Voorwoord

Leksikograwe en metaleksikograwe se belange sentreer in eenheid en dialoog.

Die tradisies, praktyke en vernuwings in die leksikografie en metaleksikografie is hoogs uiteenlopend. Dit sluit egter nie 'n gevoel van gemeenskaplikheid uit nie. Hierdie gevoel ontstaan uit die gedeelde oortuiging dat die noodsaaklike vereiste vir die leksikografiese en metaleksikografiese beoefening en besinning binne 'n kontinentale en interkontinentale konteks lê. So 'n oortuiging bevorder eenheid.

Lexikos vertoon 'n duidelike leksikografiese en metaleksikografiese gemeenskap. Die artikels in *Lexikos* bied die geleentheid vir die uitruil van gedagtes en intellektuele bespreking deur hierdie gemeenskap op 'n multikulturele en multinasionale vlak. So 'n bespreking lei tot 'n gevoel van gemeenskaplikheid wat gebaseer is op die aanvaarding en erkenning van verskille wat die werklike bron van dialoog vorm.

Lexikos streef daarna om hierdie dialoog te bevorder: die artikels wil verteenwoordigend wees van diegene wat hulle besighou met die leksikografie en metaleksikografie. Hierdie publikasie probeer dus om 'n gemeenskap te verteenwoordig wat op baie plekke, en in baie kulture en tale sy oorsprong het en wie se primêre verantwoordelikheid dit is om die ontwikkelinge en werk op die gebied van die leksikografie en metaleksikografie uiteen te sit en bekend te stel soos dit in verskillende lande en kulture, en by verskillende woordeboekeenhede plaasvind. Op hierdie manier kan daar kennis geneem word van die verskeidenheid teoretiese en praktiese aspekte wat weer as stimulus dien vir verdere ondersoek en dialoog.

Dankbetuiging

Die redakteur van *Lexikos* wil graag 'n dankwoord rig aan

— diegene wat bydraes aan *Lexikos* voorgelê het vir publikasie in hierdie nommer,
— die redaksiekomitee wat verantwoordelik was vir die keuring van die bydraes, en
— die afdeling Redaksionele Steundienste van die Buro van die WAT.

Lexikos se gehalte hang af van die aantal en verskeidenheid van die bydraes wat ontvang word. Aangesien al vanjaar se bydraes nie in hierdie nommer van

Lexikos ingesluit kon word nie, wil die redakteur diegene bedank wat bereid was dat hul bydraes oorstaan tot volgende jaar. Hy wil ook diegene bedank wie se bydraes in hierdie nommer verskyn vir hul samewerking by die persklaarmaking van die artikels.

Die redakteur is dankbaar teenoor die talle keurders wat dikwels op kort kennisgewing bereid was om die artikels te beoordeel en in baie gevalle konstruktiewe voorstelle te maak vir die verbetering daarvan. Hierdie waardevolle bydrae waarborg *Lexikos* se gehalte.

Sonder die lede van die afdeling Redaksionele Steundienste van die Buro van die WAT sou dit onmoontlik gewees het vir hierdie uitgawe van *Lexikos* om te verskyn. Nie alleen het mnr. Etienne Botha as resensieredakteur opgetree nie, maar hy het ook deurgaans gehelp met die taalversorging van die meeste artikels. Mev. Hermien van der Westhuizen en mej. Tanja Harteveld het, benewens die administrasie en korrespondensie ook die elektroniese setwerk met bekwaamheid en ywer behartig. Hulle aldrie wil ek bedank.

J.C.M.D. du Plessis
Buro van die Woordeboek van die Afrikaanse Taal

Foreword

The concern of lexicographers and metalexicographers is in unity and dialogue.

The traditions, practices and innovations in lexicography and metalexicography are highly diversified. However, this does not inhibit a sense of community. This sense originates from the common conviction that the necessary condition for lexicographic and metalexicographic pursuit and reflection lies in the recognition and consideration of the plurality of languages within a continental and intercontinental context. Such a conviction fosters unity.

Lexikos displays a definite lexicographic and metalexicographic community. The articles in *Lexikos* offer the opportunity for the exchange of ideas and intellectual discussion by this community on a multicultural and multinational level. Such a discussion leads to a sense of community based on the acceptance and recognition of differences which forms the real source of dialogue.

Lexikos seeks to further this dialogue: its articles aim to be representative of those who occupy themselves with lexicography and metalexicography. Therefore this publication aims to represent a community originating from many places, societies and languages and whose primary responsibility it is to explain and disseminate the developments and work in the field of lexicography and metalexicography taking place in various countries and cultures and at various dictionary units. In this manner cognizance can be taken of diverse theoretical and practical matters which again serves as a stimulus for further research and dialogue.

Acknowledgements

The editor of *Lexikos* wishes to extend a word of gratitude to

— all those who submitted articles for publication in this edition,
— the Editorial Committee responsible for the evaluation of the contributions, and
— the division Editorial Support Services at the Bureau of the WAT.

The standard of *Lexikos* is dependent on the number and diversity of the contributions received. Since all this year's contributions could not be included in this edition of *Lexikos*, the editor wishes to thank those who were willing to have their contributions stay over until next year. He also wishes to thank those

whose contributions appear in this edition for their cooperation in editing the articles.

The editor is grateful to the many judges who were often on short notice prepared to evaluate articles, in many cases making constructive suggestions for improvement. This valuable contribution ensures the standard of *Lexikos*.

Without the members of the division Editorial Support Services at the Bureau of the WAT it would have been impossible for this edition of *Lexikos* to appear. Not only did Mr Etienne Botha act as review editor, but he also helped throughout with the editing of the language of most of the articles. Mrs Hermien van der Westhuizen and Miss Tanja Harteveld managed, in addition to the administration and correspondence, the electronic typesetting with efficiency and diligence. I wish to thank all three of them.

J.C.M.D. du Plessis
Bureau of the Woordeboek van die Afrikaanse Taal

'n Woord van AFRILEX

Alhoewel *Lexikos* die amptelike mondstuk van AFRILEX is, is *Lexikos* nie die alleenbesit van AFRILEX nie. AFRILEX moedig sy lede wel aan om by te dra tot *Lexikos* maar nooi ook alle leksikografies geïnteresseerdes uit om van hierdie publikasieforum gebruik te maak. Nie alleen die leksikografie in Suid-Afrika en Afrika nie, maar ook die wyer internasionale leksikografie moet baat by 'n tydskrif soos *Lexikos*. *Lexikos* en AFRILEX is altwee aktief by die ontwikkeling van die leksikografiese gesprek in Afrika betrokke maar wil ook saam met ander rolspelers verseker dat dié ontwikkeling op 'n stewige en verteenwoordigende basis voortgesit word.

In die Afrikakonteks neem die behoefte aan leksikografie-opleiding al hoe meer toe. AFRILEX sien dit as deel van sy verantwoordelikheid om deel te neem aan die opleidingsinisiatief en ook om sy lede aan te moedig om betrokke te raak by opleidingsaktiwiteite. Die vereniging was die afgelope twee jaar aktief gemoeid met die opleiding wat tydens SALEX '97 en SALEX '98 aangebied is. Opleiding is nodig op sowel formele as informele vlak. In hierdie verband het *Lexikos* 'n belangrike rol om te speel. Die klem van artikels moet steeds op 'n verskeidenheid aspekte van sowel die teoretiese as die praktiese leksikografie wees, maar in die toekoms behoort meer ruimte geskep te word vir artikels wat fokus op die opleiding van woordeboekskrywers en die opvoeding van woordeboekgebruikers. Hieraan het Afrika 'n behoefte. Alle *Lexikos* lesers word uitgenooi om deur middel van die tydskrif hulle verantwoordelikheid teenoor die opleidingspoging na te kom. Ook deur die skryf van woordeboekresensies kan nuttige leiding aan woordeboekopstellers en woordeboekgebruikers gegee word. Vir die nuwe millennium het Afrika nuwe leksikograwe, woordeboekeenhede en opgeleide woordeboekgebruikers nodig.

Met hierdie uitgawe van *Lexikos* stel die Buro van die Woordeboek van die Afrikaanse Taal nogmaals hulle kundigheid en bystand tot die beskikking van die breër leksikografiese gemeenskap. Hiervoor bedank AFRILEX hulle en by name vir dr. J.C.M.D. du Plessis wat weer eens bereid was om verantwoordelikheid te aanvaar vir die redakteurskap van *Lexikos*.

Rufus Gouws
Voorsitter: AFRILEX

A Few Words from AFRILEX

Although *Lexikos* is the official journal of AFRILEX, it does not exclusively belong to AFRILEX. AFRILEX encourages its members to contribute to *Lexikos*, but also invites all persons interested in lexicography to utilise this publication forum. Not only lexicography in South Africa and Africa, but also international lexicography in general should benefit from a journal such as *Lexikos*. *Lexikos* and AFRILEX are both active in the development of lexicographic debate in Africa but need to co-operate with other role players in order to ensure this development continues on a sound and representative basis.

Within the African context the need for lexicographic training is increasing. AFRILEX considers it part of its responsibility to participate in training initiatives and encourages its members to become involved in training activities. During the past two years the association has been actively involved in the training offered at SALEX '97 and SALEX '98. Training is needed on both the formal and the informal levels. In this regard *Lexikos* has a vital role to play. The emphasis of articles should still be on the theoretical and practical aspects of lexicography, but in future more space should be set aside for articles focusing on the training of dictionary compilers and on the education of dictionary users. Africa needs this. All *Lexikos* readers are invited to use the journal as an instrument to meet their obligations towards this training endeavour. The writing of dictionary reviews can also offer valuable guidance and assistance to lexicographers and dictionary users. Facing the new millennium, Africa needs new lexicographers, dictionary units and trained dictionary users.

With this edition of *Lexikos*, the Bureau of the Woordeboek van die Afrikaanse Taal has once again made their expertise and assistance available to the lexicographic community at large. AFRILEX wishes to thank the WAT for this enterprise and Dr J.C.M.D. du Plessis for his willingness to carry, once again, the responsibility for the editorship of *Lexikos*.

Rufus Gouws
Chairperson: AFRILEX

Redaksionele doelstellings

Lexikos is 'n tydskrif vir die leksikografiese vakspesialis en word in die AFRILEX-reeks uitgegee. "AFRILEX" is 'n akroniem vir "leksikografie in en vir Afrika". Van die sesde uitgawe af dien *Lexikos* as die amptelike mondstuk van die *African Association for Lexicography* (AFRILEX), onder meer omdat die Buro van die WAT juis die uitgesproke doel met die uitgee van die AFRILEX-reeks gehad het om die stigting van so 'n leksikografiese vereniging vir Afrika te bevorder.

Die strewe van die AFRILEX-reeks is:

(1) om 'n kommunikasiekanaal vir die nasionale en internasionale leksikografiese gesprek te skep, en in die besonder die leksikografie in Afrika met sy ryk taleverskeidenheid te dien;
(2) om die gesprek tussen leksikograwe onderling en tussen leksikograwe en taalkundiges te stimuleer;
(3) om kontak met plaaslike en buitelandse leksikografiese projekte te bewerkstellig en te bevorder;
(4) om die interdissiplinêre aard van die leksikografie, wat ook terreine soos die taalkunde, algemene taalwetenskap, leksikologie, rekenaarwetenskap, bestuurskunde, e.d. betrek, onder die algemene aandag te bring;
(5) om beter samewerking op alle terreine van die leksikografie moontlik te maak en te koördineer, en
(6) om die doelstellings van die *African Association for Lexicography* (AFRILEX) te bevorder.

Hierdie strewe van die AFRILEX-reeks sal deur die volgende gedien word:

(1) Bydraes tot die leksikografiese gesprek word in die vaktydskrif *Lexikos* in die AFRILEX-reeks gepubliseer.
(2) Monografiese en ander studies op hierdie terrein verskyn as afsonderlike publikasies in die AFRILEX-reeks.
(3) Slegs bydraes wat streng vakgerig is en wat oor die suiwer leksikografie of die raakvlak tussen die leksikografie en ander verwante terreine handel, sal vir opname in die AFRILEX-reeks kwalifiseer.
(4) Die wetenskaplike standaard van die bydraes sal gewaarborg word deur hulle aan 'n komitee van vakspesialiste van hoë akademiese aansien voor te lê vir anonieme keuring.

Lexikos sal jaarliks verskyn, terwyl verdienstelike monografiese studies sporadies en onder hulle eie titels in die AFRILEX-reeks uitgegee sal word.

Editorial Objectives

Lexikos is a journal for the lexicographic specialist and is published in the AFRILEX Series. "AFRILEX" is an acronym for "lexicography in and for Africa". From the sixth issue, ***Lexikos*** serves as the official mouthpiece of the *African Association for Lexicography* (AFRILEX), amongst other reasons because the Bureau of the WAT had the express aim of promoting the establishment of such a lexicographic association for Africa with the publication of the AFRILEX Series.

The objectives of the AFRILEX Series are:

(1) to create a vehicle for national and international discussion of lexicography, and in particular to serve lexicography in Africa with its rich variety of languages;
(2) to stimulate discourse between lexicographers as well as between lexicographers and linguists;
(3) to establish and promote contact with local and foreign lexicographic projects;
(4) to focus general attention on the interdisciplinary nature of lexicography, which also involves fields such as linguistics, general linguistics, lexicology, computer science, management, etc.;
(5) to further and coordinate cooperation in all fields of lexicography; and
(6) to promote the aims of the *African Association for Lexicography* (AFRILEX).

These objectives of the AFRILEX Series will be served by the following:

(1) Contributions to the lexicographic discussion will be published in the specialist journal ***Lexikos*** in the AFRILEX Series.
(2) Monographic and other studies in this field will appear as separate publications in the AFRILEX Series.
(3) Only subject-related contributions will qualify for publication in the AFRILEX Series. They can deal with pure lexicography or with the intersection between lexicography and other related fields.
(4) Contributions are judged anonymously by a panel of highly-rated experts to guarantee their academic standard.

Lexikos will be published annually, but meritorious monographic studies will appear as separate publications in the AFRILEX Series.

Redaktionelle Ziele

Lexikos ist eine Zeitschrift für Fachleute der Lexikographie, die in der AFRILEX-Serie erscheint. "AFRILEX" ist ein Akronym für "Lexikographie in und für Afrika". Von der sechsten Ausgabe dient ***Lexikos*** als amtliches Mundstück des *African Association for Lexicography* (AFRILEX), u.a. weil das Büro des WAT gerade das ausgesprochene Ziel mit der Ausgabe der AFRILEX-Serie hatte, die Gründung solches lexikographischen Vereins für Afrika zu fördern.

Die folgenden Ziele werden mit den Publikationen der AFRILEX-Serie verfolgt: Man möchte:

(1) ein Medium schaffen für die nationale und internationale Diskussion, besonders aber der Lexikographie in Afrika mit seinen zahlreichen Sprachen dienen;
(2) die Diskussion fördern, unter Lexikographen als auch zwischen Lexikographen und Linguisten;
(3) Kontakt herstellen und fördern zwischen südafrikanischen und ausländischen lexikographischen Projekten;
(4) die Aufmerksamkeit lenken auf die interdisziplinäre wissenschaftliche Praxis der Lexikographie, die Beziehung aufweist zur Linguistik, allgemeinen Sprachwissenschaft, Lexikologie, Computerwissenschaft, zum Management und zu anderen Bereichen;
(5) die Zusammenarbeit auf allen Gebieten der Lexikographie fördern und koordinieren;
(6) die Ziele der *African Association for Lexicography* (AFRILEX) fördern.

Gemäß den Zielsetzungen der AFRILEX-Serie werden:

(1) Beiträge zum lexikographischen Gespräch in der Fachzeitschrift ***Lexikos*** veröffentlicht;
(2) monographische und andere Studien auf diesem Gebiet als getrennte Publikationen in der AFRILEX-Serie erscheinen;
(3) nur einschlägige Beiträge, die sich ausschließlich mit Lexikographie oder mit fachverwandten Gebieten befassen, für Aufnahme in der AFRILEX-Serie in Betracht gezogen;
(4) Beiträge anonym von einem aus Spezialisten des Faches von hohem akademischen Ansehen bestehenden Ausschuß beurteilt.

Lexikos erscheint jährlich. Ausgewählte monographische Studien dagegen erscheinen gelegentlich als getrennte Publikationen in der AFRILEX-Serie.

Die teenstellingsdefinisie in Afrikaanse verklarende woordeboeke

Herman L. Beyer, Departement Germaanse en Romaanse Tale,
Universiteit van Namibië, Windhoek, Namibië

Abstract: The Contrastive Definition in Afrikaans Explanatory Dictionaries. This paper deals with the contrastive definition which is readily applied in Afrikaans standard explanatory dictionaries to define the female member of a gender opposition pair. The aim of the paper is to determine whether the contrastive definition is a full lexicographic definition, and whether its application is lexicographically justifiable. The terms *gender opposition, gender opposition pairs, male and female members of gender opposition pairs* and *contrastive definition* are explained at the outset. Arguments in favour of the application of the contrastive definition found in Afrikaans metalexicographic literature, are presented, followed by arguments opposing its use. The absence of an evaluation of the contrastive definition as lexicographic definition is confirmed, and such an evaluation, based on further research into the use of the contrastive definition in the *Verklarende Handwoordeboek van die Afrikaanse Taal*, is consequently presented. The implications of the use of the contrastive definition in terms of current lexicographic conventions and the employment of a model in terms of which lexicographic definitions can be classified, lead to the conclusion that the contrastive definition cannot be regarded as a lexicographic definition. Comparable defining strategies in non-Afrikaans dictionaries are briefly highlighted, followed by proposals for the alternative treatment of (female members of) morphologically marked gender opposition pairs.

Keywords: CONTRASTIVE DEFINITION, CROSS REFERENCE, FEMALE MEMBERS, GENDER OPPOSITION, GENDER OPPOSITION PAIRS, LEXICOGRAPHER, LEXICOGRAPHIC TREATMENT, LEXICOGRAPHIC DEFINITION, MALE MEMBERS, MORPHOLOGICALLY MARKED GENDER OPPOSITION PAIRS, REFERENCE, SEMANTIC AND SYNTACTIC INFORMATION, SEMANTIC EXPLANATION, STANDARD EXPLANATORY DICTIONARY

Opsomming: Hierdie artikel behandel die teenstellingsdefinisie wat geredelik in Afrikaanse standaard verklarende woordeboeke gebruik word om die vroulike lid van 'n geslagsopposisiepaar te definieer. Die doel van die artikel is om te bepaal of die teenstellingsdefinisie 'n volwaardige leksikografiese definisie is, en of die gebruik daarvan leksikografies geregverdig is. Die terme *geslagsopposisie, geslagsopposisiepare, manlike en vroulike lede van geslagsopposisiepare* en *teenstellingsdefinisie* word ten aanvang verduidelik. Argumente ten gunste van die gebruik van die teenstellingsdefinisie aangetref in die Afrikaanse metaleksikografiese literatuur, word weergegee, gevolg deur argumente wat die gebruik van dié definisie teenstaan. Die afwesigheid van 'n evaluering van die teenstellingsdefinisie as leksikografiese definisie word bevestig, en sodanige evaluering word vervolgens aangebied na aanleiding van verdere ondersoek na die optrede van dié definisie in die *Verklarende Handwoordeboek van die Afrikaanse Taal*. Die implikasies van die optrede van die teenstellingsdefinisie in terme van geldende leksikografiese konvensies en die aanwending van 'n

model in terme waarvan leksikografiese definisies geklassifiseer kan word, lei tot die gevolgtrekking dat die teenstellingsdefinisie nie as leksikografiese definisie erken kan word nie. Vergelykbare definiëringstrategieë in nie-Afrikaanse woordeboeke word kortliks uitgewys, gevolg deur voorstelle vir die alternatiewe hantering van (vroulike lede van) morfologies gemerkte geslagsopposisiepare.

Sleutelwoorde: BETEKENISVERKLARING, GESLAGSOPPOSISIE, GESLAGSOPPOSISIEPARE, KRUISVERWYSING, LEKSIKOGRAAF, LEKSIKOGRAFIESE DEFINISIE, LEKSIKOGRAFIESE HANTERING, MANLIKE LEDE, MORFOLOGIES GEMERKTE GESLAGSOPPOSISIEPARE, SEMANTIESE EN SINTAKTIESE INLIGTING, STANDAARD VERKLARENDE WOORDEBOEK, TEENSTELLINGSDEFINISIE, VERWYSING, VROULIKE LEDE

1. Inleiding

Hierdie artikel behandel die teenstellingsdefinisie, wat geredelik in Afrikaanse standaard verklarende woordeboeke gebruik word om die vroulike lid van 'n geslagsopposisiepaar te definieer. Hierdie manier van definiëring is nie uniek aan die Afrikaanse leksikografie nie: in ten minste twee Duitse verklarende woordeboeke word die vroulike lede van geslagsopposisiepare op soortgelyke wyse gedefinieer.

Die doel van hierdie artikel is om te bepaal of die teenstellingsdefinisie 'n volwaardige leksikografiese definisie is, en of die gebruik daarvan leksikografies geregverdig is.

Die *Verklarende Handwoordeboek van die Afrikaanse Taal* (Odendal, Schoonees, Swanepoel, Du Toit en Booysen 1994 — voortaan *HAT*) is as verteenwoordiger van die versameling Afrikaanse standaard verklarende woordeboeke as databron gebruik. Tensy anders vermeld, kom alle aangehaalde voorbeelde uit dié woordeboek. Die teenstellingsdefinisie kom ook dikwels voor in *Verklarende Afrikaanse Woordeboek* (Labuschagne en Eksteen 1993).

2. Geslagsopposisie, geslagsopposisiepare en die teenstellingsdefinisie

Ten aanvang word die volgende terme omskryf: geslagsopposisie, geslagsopposisiepare, manlike en vroulike lede van geslagsopposisiepare en die teenstellingsdefinisie.

2.1 Geslagsopposisie en geslagsopposisiepare

'n Semantiese verhouding van geslagsopposisie bestaan tussen twee leksikale items wat in die eerste plek ten opsigte van die semantiese kategorie [geslag] in opposisie tot mekaar staan, dit wil sê die een item bevat die semantiese komponent [+ manlik], terwyl die ander item die semantiese komponent [+ vroulik] bevat. Behalwe vir die verskil in hierdie semantiese kategorie het die twee leksikale items identiese semantiese waardes, byvoorbeeld *koning* x *koningin* (vir

die denotasie "regeerder oor 'n koninkryk"). Die items *koning* en *koningin* word dus beskou as onderskeidelik die manlike en vroulike lede van die geslagsopposisiepaar *koning* x *koningin*. Volgens Beyer (1997: 114) bestaan 'n geslagsopposisiepaar "slegs uit 'n [+ manlik]-gemerkte leksikale item en die teenoorstaande [+ vroulik]-gemerkte leksikale item. Geslagsopposisiepare bevat dus nié geslagtelik neutrale (dit is [± manlik]-gemerkte) leksikale items nie." In 'n geslagsopposisiepaar soos *onderwyser* x *onderwyseres* is die [± manlik]-gemerkte waarde van die leksikale item *onderwyser* dus nie ter sprake nie; dié waarde geld as aparte polisemiese waarde van die leksikale item *onderwyser* as superordinaat van beide lede van die geslagsopposisiepaar, hoewel hierdie feit nie in Afrikaanse verklarende woordeboeke verantwoord word nie (vgl. Beyer 1997). Dit is dus duidelik dat dit ook by die semantiese verhouding van geslagsopposisie nie "'n leksikale item is wat in een of ander verhouding staan tot 'n ander leksikale item nie, maar eerder die een of ander polisemiese waarde van die betrokke leksikale item (poliseem) wat in 'n bepaalde verhouding staan tot die een of ander polisemiese waarde van 'n ander leksikale item" (De Stadler 1989: 85).

Oor die aard van die semantiese verhouding van geslagsopposisie bestaan uiteenlopende standpunte. Fouché (1990: 91) benader geslagsopposisie as 'n verhouding van komplementariteit, terwyl Combrink (1990: 108) betoog "[+ manlik] en [+ vroulik] is nie binêr teenoorgestelde eienskappe nie — hulle is punte op 'n skaal". Combrink se siening maak egter slegs voorsiening vir die polisemiese waarde "met tipiese manlike/vroulike eienskappe" van 'n leksikale item (geslag as 'n karakteristieke kenmerk van die referent), maar nie vir die polisemiese waarde "behorende tot die manlike/vroulike geslag" (geslag as 'n inherente kenmerk van die referent) nie. Die geslag waaraan 'n referent (inherent) behoort, kan nie op 'n skaal geleë wees nie — dit is 'n vaste gegewe (kenmerk), maar die karakteristieke wat 'n referent openbaar, kan die referent wel op 'n skaal met die punte (kenmerke) "manlikheid" (Eng. "masculinity") en "vroulikheid" (Eng. "femininity") plaas (vgl. De Stadler 1989: 85). Die semantiese verhouding van geslagsopposisie slaan slegs op geslag as inherente kenmerk van die betrokke referente.

Die siening dat geslagsopposisie 'n verhouding van komplementariteit verteenwoordig, is ook nie sonder meer geldig nie. Volgens Beyer (1997: 108) behoort daar eerder sprake te wees van 'n verhouding van gedeeltelik komplementêre opposisies: "'n Komplementêre geslagsopposisiepaar ontstaan [...] eintlik slegs wanneer 'n ooreenkomstige komplementêre opposisieverhouding deur 'n bepaalde konteks geaktiveer word." So byvoorbeeld kan *onderwyser* óf die komplement *onderwyseres* óf die komplement *leerling* neem, na gelang van die konteks waarin *onderwyser* optree. Beide die leksikale items is gedeeltelike komplemente van die item *onderwyser*.

Vir die doel van hierdie artikel word geslagsopposisie dus as 'n semantiese verhouding van gedeeltelike komplementariteit met die semantiese kategorie [geslag] as diagnostiese kenmerk beskou.

2.2 Die teenstellingsdefinisie

Met die term *teenstellingsdefinisie* word verwys na dié leksikografiese meganisme wat blykbaar eksklusief aangewend word in die beskrywing van morfologies gemerkte vroulike lede van geslagsopposisiepare, dit wil sê vroulike lede wat van [+ manlike] of [– animate][1] stamme afgelei word deur die aanvoeging van sogenaamde vervroulikingsuffikse. Vergelyk die geslagsopposisiepaar *baron* x *barones* in (1)(a) as tipiese voorbeeld waar die vroulike lid van 'n [+ manlike] stam afgelei is, en (1)(b) as tipiese voorbeeld waar die vroulike lid van 'n [– animate] stam afgelei is:

(1) (a) *baron* (stam, manlike lid) + *-es* (vervroulikingsuffiks) → *barones* (vroulike lid)
(b) *bak* ([– animate] stam) + *-ster* (vervroulikingsuffiks) → *bakster* (vroulike lid)

In *HAT* neem die teenstellingsdefinisie die vorm van 'n formule aan wat geïdentifiseer is deur die vorm van die meerderheid van die opgetekende teenstellingsdefinisies, naamlik

(2) [**vroulike lid van geslagsopposisiepaar**] Vr. vorm van [*manlike lid van geslagsopposisiepaar*]

byvoorbeeld

(3) (a) **barones** Vr. vorm van *baron*.
(b) **bakster** Vr. vorm van *bakker*.

In die Afrikaanse metaleksikografiese literatuur word hierdie soort definisie aanvaar (vgl. Fouché 1990 en Gouws 1989), maar nie verduidelik of gemotiveer nie. Fouché (1990: 149) plaas die definisie saam met die sogenaamde verkleiningsdefinisie in 'n vae kategorie wat genoem word "definisies wat semantiese pluswaardes ekspliseer" en wat blykbaar op dieselfde vlak as die ander hoofsoorte leksikografiese definisies funksioneer. Hierdie stand van sake vra om nadere ondersoek, waarvan in die volgende paragrawe verslag gedoen word.

3. Argumente ten gunste van die gebruik van die teenstellingsdefinisie

Die bestaande argumente ten gunste van die aanwending van die teenstellingsdefinisie word vervolgens slegs weergegee; relevante kritiek op dié argumente vloei uit die res van die bespreking voort.

3.1 Eksplisering van semantiese pluswaarde

Volgens Zgusta (1971: 253) is die funksie van die leksikografiese definisie om die gedefinieerde leksikale item te differensieer van die ander leksikale items: "The lexicographic definition enumerates only the most important semantic features of the defined lexical unit, which suffice to differentiate it from other units." Gouws (1989: 159) ondersteun ook die aanwending van die teenstellingsdefinisie, byvoorbeeld by die lemma **onderwyseres**: "Semanties is dit bevredigend aangesien die betekenisverklaring by die lemma **onderwyser** gegee word en die definiens by **onderwyseres** die onderskeidende semantiese waarde tussen dié twee woorde duidelik stel." Volgens Fouché (1990: 151) besit die morfologies gemerkte vroulike vorm 'n semantiese pluswaarde, naamlik dié van vroulik, wat die aanwending van die teenstellingsdefinisie regverdig, aangesien dié semantiese pluswaarde sodoende geëkspliseer word. Dit is dan die eksplisering van hierdie semantiese pluswaarde wat voldoende is om die betekenis van die vroulike vorm van dié van die manlike vorm te onderskei.

3.2 Funksie van betekenisverklaring en verwysing

Die teenstellingsdefinisie vervul 'n dualistiese funksie daarin dat dit 'n betekenisverklaringsaspek sowel as 'n verwysingsaspek bevat. Volgens Beyer (1995: 53) het die teenstellingsdefinisie die volgende twee funksies:

(4) (a) Die teenstellingsdefinisie ekspliseer 'n bepaalde opposisieverhouding tussen twee leksikale items, naamlik dié leksikale item wat as definiendum van die teenstellingsdefiniens optree en dié leksikale item in terme waarvan die lemma verklaar word.
(b) Die teenstellingsdefinisie dien terselfdertyd as ('n vorm van onregstreekse) betekenisverklaring van die lemma waar dit aangewend word.

Fouché (1990: 150) ondersteun hierdie standpunt: "Wanneer 'n teenstellingsdefinisie gebruik word, word die definiendum se semantiese betrekkinge met 'n ander leksikale item duidelik" sonder dat dit nodig sou wees om bykomende semantiese inligting deur verdere inskrywings te voorsien.

Vanuit 'n praktiese oogpunt werk die teenstellingsdefinisie dus ruimtebesparing in die hand.

3.3 Taalkundige inligting

Die aanwending van die teenstellingsdefinisie bevorder die akkuraatheid van taalkundige inligting wat in die woordeboek weergegee word.

Volgens Beyer (1995: 22-23) is die semantiese verskil tussen die manlike en vroulike lede van 'n morfologies gemerkte geslagsopposisiepaar kongruent aan die morfologiese verskil tussen hulle. Hierdie kongruensie word akkuraat deur die teenstellingsdefinisie weergegee. Ilson (1987: 63) wys ook op "the usefulness of formulaic definitions for giving morphological information: they incorporate words linked morphologically to their definienda. In a very real sense the information they give is etymological in nature [...]."

Die teenstellingsdefinisie is ook 'n akkurate beskrywing van die aard van die semantiese verhouding ter sprake (nl. gedeeltelike komplementariteit, vgl. 2.1) deur die formulering "Vr. vorm van ...", en nie bloot "teenoor" (wat volledige komplementariteit sou suggereer) nie, hetsy in die definiens, hetsy deur middel van 'n addisionele inskrywing.

4. Argumente teen die gebruik van die teenstellingsdefinisie

Argumente in die Afrikaanse metaleksikografiese literatuur wat die gebruik van die teenstellingsdefinisie teenstaan, kom hoofsaaklik vanuit 'n sosiolinguistiese perspektief.

4.1 Opname van vroulike vorme

Volgens Beylefeld en Van Jaarsveld (1994: 46-47) is die lemmastatus van vroulike lede van morfologies gemerkte geslagsopposisiepare "twyfelagtig aangesien die betekenisleiding slegs per kruisverwysing as 'vroulik van...' aangegee word — selfs in gevalle waar die vroulike vorm volgens alfabetiese rangorde voor die manlike vorm gelys is [...]. In die lig hiervan, is die eksplisiete [+ vroulik]-gemerktheid van beroepsbenaminge oorbodig en 'n vermorsing van ruimte." Dit is duidelik dat Beylefeld en Van Jaarsveld nie die teenstellingsdefinisie as leksikografiese definisie erken nie — hulle beskou dit as 'n blote kruisverwysingsmiddel. Afgesien hiervan is hulle kritiek wat die lemmastatus van die vroulike lede betref, leksikografies ongegrond. Die doel van enige standaardwoordeboek is om die leksikale items van die standaardtaal te verklaar. Indien die vroulike lede van morfologies gemerkte geslagsopposisiepare die standaardtaal verteenwoordig, moet hulle as lemmas in die woordeboek opgeneem word. Gouws (1989: 159) stel dit ook kategories dat beide lede van 'n geslagsopposisiepaar as volle lemmas hanteer behoort te word. Dié stelling kan wel gekwalifiseer word deur die voorwaarde dat beide lede verteenwoordigend van die standaardtaal moet wees. Dit sal dus byvoorbeeld nie nodig wees om die volle geslagsopposisiepaar *beer* x *berin* of *sot* x *sottin* in die makrostruktuur op te neem nie, aangesien die items *berin* en *sottin* verouderde taalgebruik verteenwoordig, wat in elk geval impliseer dat die items *beer* en *sot* volledig geslagtelik geneutraliseer is. Van der Merwe (1994: 232) stel dit ook "dat 'n handwoordeboek nie historiese vorme, dialektiese vorme, sleng en tegniese vorme [moet] opneem nie. 'n Gebruiker van 'n handwoordeboek [as tipiese ge-

bruiker — HLB] stel ook nie in etimologiese inligting en verouderde taalgebruik belang nie." Hierdie kriteria behoort die enigste maatstaf te wees waarvolgens 'n item se lemmastatus in 'n standaardwoordeboek bepaal word.

Beylefeld en Van Jaarsveld (1994: 44) is egter van mening dat "indien dit so is dat die meerderheid Afrikaanse moedertaalsprekers nog vassteek by konvensioneel seksistiese taalgebruik, behoort woordeboekmakers dit hul plig te ag om alternatiewe, nieseksistiese taalvorme en -gebruik te vestig." Hierdie siening is onaanvaarbaar, aangesien dit die taak van die verklarende leksikograaf is om die inhoud van die leksikon te beskryf, en nie om preskriptief op te tree nie. Volgens Spender (1980: 29-30) is daar boonop "fundamental problems with the creation of new words because while they are also subjected to the existing semantic rule that male is positive and minus male is negative, there is reason to believe that when consigned to negative semantic space they too will become pejorated and sexist. It is the semantic rule which needs to change, not the words themselves, yet this suggestion has rarely arisen in language/sex research." Vanuit 'n leksikografiese oogpunt moet Romaine (1994: 127) se standpunt geld "that society's perceptions of men and women must change in order for linguistic reform to be successful." Uiteindelik sal die taalgebruiker bepaal of sekere items langer in die verklarende woordeboek opgeneem sal word.

Die aanwending van die teenstellingsdefinisie by morfologies gemerkte geslagsopposisiepare is ook nie beperk tot beroepsbenaminge nie: vergelyk die gebruik van dié definisie by die vroulike lede van die geslagsopposisiepare *eienaar* x *eienares*, *speler* x *speelster*, *voorsitter* x *voorsitster*, *beer* x *berin* en *leeu* x *leeuin* (hoewel laasgenoemde twee se vroulike lede verouderde taalgebruik verteenwoordig).

4.2 Vroulike vorme se betekenisbeskrywing

Beylefeld en Van Jaarsveld (1994: 47) lewer verdere kritiek op die gebruik van die teenstellingsdefinisie vanuit 'n sosiolinguistiese oogpunt: "Op die semantiese vlak vorm manlikheid dus die domein waarbinne die vroulike beroepsbekleder [of vroulike lid van die geslagsopposisiepaar —HLB] betekenis moet verkry." Hierdie punt van kritiek blyk geldig te wees, aangesien dit die primêre rede is vir die huidige niegebruik van die teenstellingsdefinisie in die *Woordeboek van die Afrikaanse Taal* (*WAT*) (W.F. Botha en A.E. Cloete (Buro van die WAT) — persoonlike mededeling). Fouché (1990: 151) se siening dat die vroulike lid van 'n morfologies gemerkte geslagsopposisiepaar 'n semantiese pluswaarde besit (vgl. 3.1), word ook deur hierdie kritiek onder verdenking geplaas.

5. Die teenstellingsdefinisie as leksikografiese definisie

Geeneen van die bestaande argumente raak die geldigheid al dan nie van die teenstellingsdefinisie as leksikografiese definisie aan nie. Die feit dat daar

onder teoretici blykbaar nie eenstemmigheid is oor die blote bestáán van dié definisie nie (vgl. Beylefeld en Van Jaarsveld 1994 se siening in 4.1) is rede genoeg om verdere ondersoek te regverdig. Vervolgens word die argumente voortspruitend uit sodanige ondersoek aangebied.

5.1 Beperkte semantiese gebruik

Die aanwending van die teenstellingsdefinisie blyk beperk te wees tot 'n bepaalde soort opposisieverhouding, naamlik geslagsopposisie. Dit is hierdie beperkte aanwending wat dit verdag maak. Ander soorte opposisieverhoudings word gewoonlik deur verwysings naas die betekenisverklaring (wat deur middel van enige ander leksikografiese definisie gedoen word) gemaak deur byvoorbeeld die merkers *teenoor* en *vergelyk* te gebruik. Daar bestaan geen ander leksikografiese definisie wat eksklusief aangewend word in die eksplisering van 'n bepaalde leksikale opposisieverhouding nie.

Die teenstellingsdefinisie is ook nie noodsaaklik in die betekenisverklaring van die vroulike lid van 'n geslagsopposisiepaar nie; dié definisie kan maklik deur enige ander leksikografiese definisie vervang word sonder om suksesvolle betekenisverklaring in die gedrang te bring. Addisioneel tot die aanwending van enige ander leksikografiese definisie kan 'n blote verwysing dan 'n leksikale opposisieverhouding eksplíseer. Beylefeld en Van Jaarsveld (1994: 46) beskou die teenstellingsdefinisie inderdaad slegs as 'n verwysing (vgl. 4.1).

5.2 Beperkte leksikale bestek

Na aanleiding van Combrink (1990: 105-106) word die vroulike lede van geslagsopposisiepare op drie maniere geleksikaliseer. Eerstens bestaan daar afsonderlike leksikale items (stamme) vir vroulike referente, byvoorbeeld *dogter, vrou* en *hen*. Tweedens bestaan daar komposita met [+ vroulike] stamme soos *dames-, -vrou* en *wyfie-/-wyfie*. Derdens bestaan daar afleidings met [+ manlike] stamme en [+ vroulike] suffikse, byvoorbeeld *barones, koningin* en *onderwyseres*, en afleidings met [– animate] stamme en [+ vroulike] suffikse, byvoorbeeld *bakster, helpster* en *speelster*.

Daar is reeds melding gemaak dat die teenstellingsdefinisie slegs by morfologies gemerkte geslagsopposisiepare aangewend word. Eersgenoemde kategorie vroulike lede van geslagsopposisiepare (vroulike lede as afsonderlike leksikale items), wat nie produkte van morfologiese prosesse is nie, word dus nie deur hierdie definisie gedek nie. Dit veroorsaak 'n diskrepansie, aangesien daar geen semantiese verskil tussen leksikale en morfologies gemerkte geslagsopposisiepare bestaan nie.

Dié diskrepansie word vergroot deur die feit dat die teenstellingsdefinisie slegs optree by vroulike lede van morfologies gemerkte geslagsopposisiepare wat produkte van suffigering is. Kompositumprodukte word ook nie deur die teenstellingsdefinisie gedek nie.

Slegs een van die drie maniere waarop die semantiese verhouding van geslagsopposisie geleksikaliseer word, word deur die teenstellingsdefinisie verantwoord. Dit blyk dus duidelik dat die aanwending van die teenstellingsdefinisie morfologies gemotiveer is, en nie semanties nie. Hierdie stand van sake strook nie met die primêre motivering vir die gebruik van die leksikografiese definisie nie, naamlik die weergee van semantiese inligting.

5.3 Beperkte optrede

Die aanwending van die teenstellingsdefinisie by suffigaal gemerkte vroulike lede van geslagsopposisiepare veronderstel oënskynlik nie die konsekwente optrede van dié definisie by sulke lemmas nie.[2] Vergelyk die *HAT*-uittreksels in (5) en (6):

(5) (a) **tikker** Manlike persoon wat (as beroep) met 'n tikmasjien tik.
(b) **tikster** Vroulike persoon wat as beroep met 'n tikmasjien tik.
(6) (a) **verpleër** Man wat spesiaal opgelei is om siekes te verpleeg.
(b) **verpleegster** Vroulike persoon wat spesiaal opgelei is om siekes te verpleeg.

Teen die agtergrond van die inkonsekwente toepassing van leksikografiese beginsels in Afrikaanse woordeboeke bestaan daar twee perspektiewe waaruit die hantering van die geslagsopposisiepare in (5) en (6) verklaar kan word.

Die eerste verklaring steun op die feit dat die vroulike lede ter sprake nie afgelei is van die manlike lede nie, maar van dieselfde stamme as waarvan die manlike lede sélf afgelei is. Die morfologiese verskil tussen die manlike en vroulike lede is dus nie kongruent aan die semantiese verskil tussen hulle nie; derhalwe sou die optrede van 'n teenstellingsdefinisie by die vroulike lede nie 'n akkurate weergawe van taalkundige inligting verteenwoordig nie (vgl. 3.3). Indien hierdie verklaring die geldende een is, word dit nie konsekwent toegepas nie: die lemmas **bakster, helpster, waarsegster** en **skryfster** word byvoorbeeld wel deur teenstellingsdefinisies verklaar.

Die tweede verklaring steun op die waarskynlikheid (na aanleiding van die uitsonderings by die eerste verklaring) dat die leksikograaf nie die semanties-morfologiese kongruensie (vgl. 3.3) in ag geneem het by die hantering van die lemmas in (5)(b) en (6)(b) nie, maar dat die vroulike lede telkens volledig verklaar is omdat hulle die mees gebruiklike items in algemene taalgebruik is, terwyl die manlike lede selde voorkom. Beide lede van elke geslagsopposisiepaar word dus van volledige betekenisverklarings voorsien. Die vraag wat dan tereg gevra mag word, is waarom die manlike lid nie deur 'n teenstellingsdefinisie soos

(7) **tikker** Manlike vorm van *tikster*

verklaar word nie.

Dit is duidelik dat beide verklarings nie houdbaar is nie, en verdere meriete word aan Beylefeld en Van Jaarsveld (1994: 46) se sosiolinguistiese kritiek in 4.2 verleen.

Ten spyte van bogenoemde verklarings vir die nieoptrede van die teenstellingsdefinisie by sommige suffigaal gemerkte geslagsopposisiepare het die semantiese verhouding van geslagsopposisie tussen die betrokke lede onveranderd gebly. Weer eens word dit duidelik dat die optrede van die teenstellingsdefinisie eerder morfologies as semanties gemotiveer is (hoewel selfs die morfologiese motivering blykbaar nie konsekwent toegepas word nie).

5.4 Beperkende optrede

Die optrede van die teenstellingsdefinisie kan die implisiete of eksplisiete stel van ander leksikale betrekkinge waarin die definiendum tot ander leksikale items in die makrostruktuur mag staan, verhinder. Vergelyk die hantering van die lemma **regisseuse**:

(8) **regisseuse** Vroulike vorm van *regisseur*; spelleidster.

Indien *spelleidster* hipoteties gesproke die mees gebruiklike lid van die sinoniemparadigma [*regisseuse, spelleidster*] is (soos wat die bewerking in (8) inderdaad wil suggereer), behoort die lemma **regisseuse** slegs deur 'n sinoniemdefinisie verklaar te word. Die teenstellingsdefinisie is dus onnodig.

Indien die manlike lid *regisseur* verder hipoteties gesproke die mins gebruiklike lid van die sinoniemparadigma [*regisseur, spelleier*] sou wees, sou die lemma **regisseur** deur 'n sinoniemdefinisie ("spelleier") verklaar word. Die woordeboekgebruiker wat die betekenis van *regisseuse* sou soek, word dus deur die teenstellingsdefinisie (as primêre definisie) verwys na die lemma **regisseur**, en dan weer verder verwys na die lemma **spelleier**. Dit is duidelik dat hierdie verwysingsnetwerk nie gebruikersvriendelik is nie. Volgens Fouché (1990: 105) is dit wenslik dat die woordeboekgebruiker nie meer as een keer na 'n ander woordeboekartikel verwys moet word nie.

Die feit dat die teenstellingsdefinisie slegs een bepaalde leksikale betrekking weergee, veroorsaak dat sy optrede beperkend inwerk op die weergee van ander semantiese inligting. Die omgekeerde is ook waar: die weergee van ander semantiese inligting in die mikrostruktuur werk beperkend op die konsekwente toepasbaarheid van die teenstellingsdefinisie in.

5.5 Semantiese en sintaktiese inligting

Beyer (1995: 50-63) het die teenstellingsdefinisie as leksikografiese definisie ondersoek deur dit te vergelyk met die deskriptiewe definisie en die sinoniemdefinisie, en kom tot die gevolgtrekking dat "uit die aard van die semantiese funksies wat die teenstellingsdefinisie in die verklarende woordeboek vervul,

kan die teenstellingsdefinisie nie onvoorwaardelik as leksikografiese definisie beskou word nie". Hierdie gevolgtrekking word gebaseer op die feit dat die teenstellingsdefinisie die aanduiding van 'n leksikale betrekking voorop stel, en nie in die eerste plek betekenisbeskrywend optree nie (vgl. (4) in 3.2).

Die toepassing van Ilson (1987) se ontleding van leksikografiese definisies lewer die deurslaggewende kommentaar op die teenstellingsdefinisie as leksikografiese definisie. Volgens Ilson (1987: 72) is daar "a list of types of information provided by [lexicographic] definitions. Definitions can then be classified by the ways this information is distributed in them. "

'n "Vervangbare" leksikografiese definisie gee volgens Ilson (1987: 61) vier soorte inligting oor die definiendum:

SIN 1: sintaktiese hoofkategorie, byvoorbeeld s.nw., ww.
2: sintaktiese subkategorie, byvoorbeeld indien s.nw.: soortnaam, massanaam, ens.

SEM 1: semantiese kategorie: Januarie is 'n *maand*; om pragtig te wees, is om *mooi* te wees.
2: semantiese subkategorie: Januarie is die *eerste* maand; om pragtig te wees, is om *baie* mooi te wees.

Die leksikografiese definisie wat aangewend word by die lemma **onderwyser** kan nou in terme van Ilson (1987) ontleed word:

(9) **onderwyser** Persoon [SIN 1, 2; SEM 1] wat na afgelegde eksamen bevoeg verklaar is om les te gee [SEM 2].

Dié definisie is in terme van die analise suksesvol, aangesien al vier genoemde tipes inligting daarin verantwoord word. Vergelyk nou die analise van die teenstellingsdefinisie by die lemma **onderwyseres**:

(10) **onderwyseres** Vr. vorm van *onderwyser* [SIN 1, 2].

Dit blyk dat die teenstellingsdefinisie nie ewe suksesvol is in die weergee van die vereiste inligting nie. SEM 2 sou nie na die woorde "vr. vorm" geplaas kon word nie, aangesien die formulering slaan op vorm, en nie op betekenis nie. SEM 1 sou ook nie na "*onderwyser*" in die definiens geplaas kon word nie, aangesien die formulering van die teenstellingsdefinisie 'n geslagsopposisieverhouding daarstel (vgl. 3.3), wat veronderstel dat die item *onderwyser* in die definiens slaan op die [+ manlike] referent (die komplement of kohiponiem), en nie op die [± manlike] referent (die superordinaat) nie (vgl. die aard van geslagsopposisie in 2.1).

'n Genus-differentia definiens sou die vereiste inligting suksesvoller weergee, soos uit die volgende analise duidelik is:

(11) vroulike [SEM 2] onderwyser [SIN 1, 2; SEM 1].

'n Voorwaarde vir dié definiens om suksesvol te wees, is dat die verwysing na *onderwyser* nié verwysing na die manlike lid van die geslagsopposisiepaar moet wees nie, maar na die [± manlik]-gemerkte superordinaat *onderwyser* (vgl. 2.1). Die relevante poliseem van die lemma **onderwyser** moet dus in die definiens aangedui word, byvoorbeeld

(12) vroulike onderwyser (*onderwyser* bet. 1).

Die geldigheid van hierdie voorwaarde word ondersteun deur die analise van die teenstellingsdefinisie by 'n lemma soos **barones**, waar daar geen geslagtelik neutrale superordinaat geld nie:

(13) **barones** Vr. vorm van *baron* [SIN 1, 2].

In (13) is daar geen sprake van enige semantiese inligting wat oorgedra word nie. Die leksikografiese hantering in (13) is op semantiese vlak vergelykbaar met die onsuksesvolle definisie

(14) **driehoek** Driehoekige vorm van 'n vierkant [SIN 1, 2]

vir 'n konteks waarin *vierkant* en *driehoek* komplemente van mekaar is.[3]

'n Driehoek kan nie as 'n *soort* vierkant gedefinieer word nie, aangesien die twee konsepte mekaar wedersyds uitsluit; insgelyks kan 'n barones nie as 'n *soort* baron gedefinieer word nie, ensovoorts.

Hierdie stand van sake onderstreep die onsuksesvolle hantering van die lemma **beskermvrou** in (15):

(15) **beskermvrou** Beskermheer wat 'n vrou is.

Die leksikografiese definisie is volgens Ilson (1987: 71) "the de-lexicalisation of a lexical unit into semantic and syntactic components which are then presented in a single phrase whose content characterises the definiendum semantically and whose form characterises the definiendum syntactically". Dit word duidelik dat die teenstellingsdefinisie op sy beste slegs aan die helfte van Ilson se beskrywing van 'n leksikografiese definisie voldoen. Verder is die metataal wat in die teenstellingsdefinisie geld, nie die gewenste register vir suksesvolle definiëring nie: "If definitions succeed at all, it is because they are couched in *natural* language — typically, though not necessarily, in the same language as the definiendum. [...] It is probably a better metalanguage than any artificially constructed language can be [...]" (Ilson 1987: 71-72; my kursivering — HLB).

5.6 Gevolgtrekking

Dit is duidelik dat die teenstellingsdefinisie nie as 'n leksikografiese definisie beskou kan word nie. Sy aanwending as sodanig kan ook nie leksikografies

geregverdig word nie. Leksikograwe wat akkurate taalkundige inligting in hulle woordeboeke wil oordra, sal moet afsien van die gebruik van die teenstellingsdefinisie, en moet terugkeer na die bestaande en beproefde, volwaardige leksikografiese definisies.

6. Ander woordeboeke

Daar bestaan ook nie-Afrikaanse woordeboeke waarvan die hantering van die vroulike lede van geslagsopposisiepare vergelykbaar is met die aanwending van die teenstellingsdefinisie in Afrikaanse woordeboeke. Hier word volstaan met enkele sodanige voorbeelde en kriptiese opmerkings.

Vergelyk die hantering van die geslagsopposisiepare *König* x *Königin* en *Lehrer* x *Lehrerin* in *Deutsches Wörterbuch* (Wahrig 1987: 771, 825):

(16) (a) **König** höchster Herrscher eines Staates [...]
(b) **Königin** weibl. König [...] (*weibl.* "weiblich" — HLB)
(17) (a) **Lehrer** jmd., der beruflich lehrt, unterrichtet [...]
(b) **Lehrerin** weibl. Lehrer

Vergelyk bogenoemde hantering met die hantering van dieselfde lemmas in 'n ander Duitse woordeboek, *Deutsches Wörterbuch* (Mackensen 1977: 609, 659):

(18) (a) **König** Beherrscher eines Königreichs [...]
(b) **Königin** [...] Herrscherin eines Königreichs [...]
(19) (a) **Lehrer** Berufserzieher [...]
(b) **Lehrerin** w (*w* "weiblich" — HLB)

Dit blyk dat verskillende woordeboeke binne een taal se hantering van geslagsopposisiepare kan verskil, en dat verskillende metodes ook in een woordeboek gebruik kan word. Dit geld ook Afrikaanse woordeboeke. Die hantering van die lemmas in (16)(b), (17)(b) en (19)(b) is aan dieselfde kritiek onderworpe as wat op die teenstellingsdefinisie in Afrikaanse woordeboeke van toepassing is.

Die *Oxford Advanced Learner's Dictionary* (Cowie 1989: 1431) hanteer die morfologies gemerkte geslagsopposisiepaar *waiter* x *waitress* as volg:

(20) **waiter** (*fem.* **wait.ress** [...]) *n* person employed to take customers' orders [...]

Die item *waitress* beskik nie oor volle lemmastatus nie, en word opgeneem in die gleuf vir grammatiese (morfologiese) inligting. Hierteenoor hanteer *The Concise Oxford Dictionary of Current English* (Allen 1990: 1379) dié geslagsopposisiepaar soos volg:

(21) (a) **waiter 1** a man who serves at table in a hotel or restaurant etc. **2** [...]
(b) **waitress** a woman who serves at table in a hotel or restaurant etc.

Die *Collins Cobuild English Dictionary* (Sinclair 1995: 1876) en *Chambers-Macmillan South African Student's Dictionary* (Grearson en Higgleton 1996: 1113) hanteer die geslagsopposisiepaar soortgelyk aan die hantering in (21). Wat wel telkens ontbreek, is kruisverwysings tussen die manlike en vroulike lede.

7. Voorgestelde hantering

Beide lede van 'n geslagsopposisiepaar wat die standaardtaal verteenwoordig, behoort volle lemmastatus in 'n standaard verklarende woordeboek te hê, en elke lemma moet volledig verklaar word. 'n Verwysing na die teenoorstaande lid van die geslagsopposisiepaar behoort as bykomende inskrywing by die mikrostruktuur ingesluit te wees. Vergelyk die voorgestelde hantering van die geslagsopposisiepare *onderwyser* x *onderwyseres* en *koning* x *koningin*:

(22) (a) **onderwyser 1.** Persoon wat na afgelegde eksamen bevoeg verklaar is om les te gee, [ens.].
2. Manlike onderwyser (*onderwyser* bet. 1); vr. *onderwyseres*.
(b) **onderwyseres** Vroulike onderwyser (*onderwyser* bet.1); ml. *onderwyser* (bet.2).

(23) (a) **koning 1.** Manlike regeerder oor 'n koninkryk; vr. *koningin* (bet.1). **2.** [...]
(b) **koningin 1.** Vroulike regeerder oor 'n koninkryk; ml. *koning* (bet.1). **2.** [...]

Dit is belangrik dat die manlike en vroulike lede van 'n bepaalde geslagsopposisiepaar verbaal so identies moontlik verklaar moet word. Sodoende kan die leksikograaf en sy/haar woordeboek nie van geslagsvoorkeur verdink word nie, en dra die bewerking by tot die sosiale gelykstel van die twee geslagte.

Die implementering van hierdie hanteringswyse behoort nie ingrypend ruimte in die woordeboek in beslag te neem nie, aangesien die minderheid vroulike lede van geslagsopposisiepare volgens Combrink (1990: 106) deur suffigering gevorm word; vroulike lede wat nie deur suffigering gevorm is nie, word buitendien tans oorwegend volledig in verklarende woordeboeke beskryf. Die nodige kruisverwysings kan slegs ingevoeg word.

8. Slot

Hoewel in hierdie artikel hoofsaaklik op morfologies gemerkte geslagsopposisiepare gekonsentreer is, word die hoop uitgespreek dat leksikograwe die algemene hantering van geslagsopposisie in (Afrikaanse) verklarende woordeboeke sal herevalueer.

Aantekeninge

1. Die semantiese komponent [animaat] word hier gebruik soos dit deur David Crystal (*A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell) gedefinieer word: "A term used in the

grammatical classification of words (especially nouns) to refer to a subclass whose reference is to persons and animals, as opposed to inanimate entities and concepts" (1991[3]: 19).

2. Vir die doel van hierdie artikel word daar 'n onderskeid getref tussen die terme *aanwending* en *optrede*. Met *aanwending* word bedoel dat dit geredelik moontlik is (in terme van geldende leksikografiese konvensies) vir 'n definisie om as 'n bepaalde inskrywing op te tree, terwyl met *optrede* verwys word na die fisiese verskyning of optrede van 'n definisie waar dit aangewend kan word. Die moontlikheid van aanwending veronderstel dus nie noodwendig optrede nie: die teenstellingsdefinisie word so byvoorbeeld aangewend in die beskrywing van vroulike lede van morfologies gemerkte geslagsopposisiepare, maar tree nie noodwendig op by elke sodanige vroulike lid nie.

3. 'n Algemene komponensiële analise van die betrokke items vir die gegewe konteks verduidelik die vergelykbaarheid tussen hulle:

baron	*barones*	*vierkant*	*driehoek*
[+ menslik]	[+ menslik]	[+ figuur]	[+ figuur]
[+ adellik]	[+ adellik]	[+ geometries]	[+ geometries]
[+ manlik]	**[+ vroulik]**	**[+ vierkantig]**	**[+ driehoekig]**

In beide pare items verskil die lede slegs ten opsigte van een semantiese komponent. Soos wat die teenstellingsdefinisie slegs hierdie verskil tussen *baron* en *barones* weergee, dui die definiens in (14) slegs die enkele verskil tussen *vierkant* en *driehoek* aan. Daar sou wel geargumenteer kon word dat die semantiese komponente [+ vierkantig] en [+ driehoekig] verder onderverdeel kan word in onderskeidelik die komponente [+ 4 sye], [+ 90°-hoeke] en [+ 3 sye], [– 90°-hoeke], en dat hierdie komponente die eintlike definiense van die betrokke items sou konstitueer. Die semantiese komponente [+ manlik] en [+ vroulik] sou egter elk ook in verdere komponente onderverdeel kon word om meer tegniese definiense vir die items *baron* en *barones* daar te stel!

In aansluiting by Beylefeld en Van Jaarsveld (1994: 46) se sosiolinguistiese kritiek in 4.2 sou tereg gevra kon word na die rede waarom die leksikograaf besluit het om die item *driehoek* in terme van die item *vierkant* te definieer, en nie byvoorbeeld andersom nie.

Vergelyk hierteenoor die suksesvoller definisies vir dieselfde konteks:

vierkant Vierkantige geometriese figuur; vgl. *driehoek*

driehoek Driehoekige geometriese figuur; vgl. *vierkant*

Die items *vierkantig* en *driehoekig* (soos *manlik* en *vroulik*) word elders in die makrostruktuur opgeneem en verklaar in terme van die genoemde komponente waarin hulle onderverdeel kan word.

Verwysings

Woordeboeke

Allen, R.E. (Red.). 1990[8]. *The Concise Oxford Dictionary of Current English*. Oxford: Oxford University Press.

Cowie, A.P. (Red.). 1989[4]. *Oxford Advanced Learner's Dictionary*. Oxford: Oxford University Press.

Grearson, Penny en Elaine Higgleton (Reds.). 1996. *Chambers-Macmillan South African Student's Dictionary*. Manzini: Macmillan Boleswa.

Labuschagne, F.J. en L.C. Eksteen. 1993[8]. *Verklarende Afrikaanse Woordeboek*. Pretoria: J.L. van Schaik.

Mackensen, L. (Red.). 1977[3]. *Deutsches Wörterbuch*. München: Südwest Verlag.

Odendal, F.F., P.C. Schoonees, C.J. Swanepoel, S.J. du Toit en C.M. Booysen. 1994[3]. *Verklarende Handwoordeboek van die Afrikaanse Taal*. Midrand: Perskor.

Schoonees, P.C. – D.J. van Schalkwyk. (Reds.) 1970-1996. *Woordeboek van die Afrikaanse Taal*. Pretoria: Die Staatsdrukker/Stellenbosch: Buro van die Woordeboek van die Afrikaanse Taal.

Sinclair, John (Red.). 1995. *Collins Cobuild English Dictionary*. Londen: HarperCollins.

Wahrig, Gerhard (Red.). 1987[7]. *Deutsches Wörterbuch*. München: Mosaik Verlag.

Woordeboek van die Afrikaanse Taal. Kyk Schoonees – Van Schalkwyk (1970-1996).

Ander bronne

Beyer, Herman Louis. 1995. *Die leksikografiese hantering van morfologies gemerkte geslagsopposisiepare in Afrikaanse woordeboeke, met spesifieke verwysing na die* Verklarende Handwoordeboek van die Afrikaanse Taal. Ongepubliseerde M.A.-skripsie. Universiteit van Stellenbosch.

Beyer, Herman L. 1997. Aard en leksikografiese hantering van sogenaamde geslagtelik neutrale lede van Afrikaanse geslagsopposisiepare. *Suid-Afrikaanse Tydskrif vir Taalkunde* 15(4): 107-115.

Beylefeld, Adri en Gert van Jaarsveld. 1994. Is die wêreld manlik, tensy anders vermeld? *Suid-Afrikaanse Tydskrif vir Taalkunde* 12(2): 43-49.

Combrink, J.G.H. 1990. *Afrikaanse morfologie. Capita exemplaria*. Pretoria: Academica.

De Stadler, L.G. 1989. *Afrikaanse semantiek*. Johannesburg: Southern.

Fouché, Michele. 1990. *'n Evaluering van die semantiese inligting in die* Verklarende Handwoordeboek van die Afrikaanse Taal. Ongepubliseerde M.A.-skripsie. Universiteit van Stellenbosch.

Gouws, R.H. 1989. *Leksikografie*. Pretoria/Kaapstad: Academica.

Ilson, Robert F. 1987. Towards a Taxonomy of Dictionary Definitions. Ilson, Robert (Red.) 1987. *A Spectrum of Lexicography*: 61-73. Amsterdam: John Benjamins.

Romaine, Suzanne. 1994. *Language in Society. An Introduction to Sociolinguistics*. New York: Oxford University Press.

Spender, Dale. 1980. *Man Made Language*. Londen: Routledge & Kegan Paul.

Van der Merwe, Michele. 1994. 'n Evaluering van die *Verklarende Handwoordeboek van die Afrikaanse Taal* as standaard verklarende woordeboek. *Tydskrif vir Geesteswetenskappe* (36)4: 231-236.

Zgusta, Ladislav. 1971. *Manual of Lexicography*. Praag: Academia/Den Haag: Mouton.

Cross-Referencing as a Lexicographic Device

R.H. Gouws, *Department of Afrikaans and Dutch, University of Stellenbosch, South Africa* and
D.J. Prinsloo, *Department of African Languages, University of Pretoria, South Africa*

Abstract: The mediostructure, that is the system of cross-referencing, is a lexicographic device that can be used to establish relations among different components of a dictionary. This paper focuses on different mediostructural strategies and their practical application in general synchronic dictionaries. The structure of dictionaries is discussed from a metalexicographic perspective in order to explain the system of cross-referencing. It is shown how textual cohesion, achieved by the interaction of the various structural components, is promoted by the use of a system of cross-referencing and improved by an innovative approach towards a mediostructure-orientated lexicography.

Keywords: CROSS-REFERENCING, MEDIOSTRUCTURE, LEXICOGRAPHY, DICTIONARY, METALEXICOGRAPHY, REFERENCE ENTRY, REFERENCE RELATION, REFERENCE ADDRESS, AFRICAN LANGUAGES

Opsomming: Kruisverwysing as 'n leksikografiese tegniek. Die mediostruktuur, d.w.s. die stelsel van kruisverwysing, is 'n leksikografiese tegniek wat gebruik kan word om verbande te lê tussen verskillende komponente van 'n woordeboek. In hierdie artikel word aandag gegee aan verskillende mediostrukturele strategieë en hul praktiese toepassing in algemene sinchroniese woordeboeke. Woordeboekstruktuur word vanuit 'n metaleksikografiese perspektief bespreek ten einde die kruisverwysingstelsel te verduidelik. Daar word aangetoon hoe tekskohesie, verkry deur die interaksie van die onderskeie struktuurkomponente, bevorder word deur die gebruik van 'n kruisverwysingstelsel en verbeter word deur 'n vernuwende benadering ten opsigte van mediostruktureel-georienteërde leksikografie.

Sleutelwoorde: KRUISVERWYSING, MEDIOSTRUKTUUR, LEKSIKOGRAFIE, WOORDEBOEK, METALEKSIKOGRAFIE, VERWYSINGSINSKRYWING, VERWYSINGSVERBAND, VERWYSINGSADRES, AFRIKATALE

Introduction

In spite of the fact that lexicography has been practised for centuries, metalexicography, that is the theory of lexicography, is a relative new subdiscipline within the broader field of linguistics. Dictionaries existed and functioned quite well long before theoreticians, critical analysis and theoretical frameworks. Today, however, it is widely accepted that there is a strong interplay between

metalexicography and the lexicographic practice. The metalexicographical influence has transformed lexicography into a scientific practice with a very definite purpose, viz. the production of dictionaries. The production of dictionaries as a result of the scientific practice of lexicography should lead to the establishment of the cultural practice of dictionary use. The user-perspective, so prevalent in modern-day metalexicography, compels lexicographers to compile their dictionaries according to the needs and research skills of well-defined target user groups. The dominant role of the user has had a definite effect on the compilation of dictionaries as well as on the evaluation of their quality. Good dictionaries do not only display a linguistically sound treatment of a specific selection of lexical items. They are also products that can be used as linguistic instruments by their respective target user groups. The better they can be used, the better dictionaries they are.

The quality of dictionary use, that is the degree of success a user experiences when consulting a dictionary and employing the retrieved information, is determined by a variety of features, but one of the most important characteristics of a good dictionary is its accessibility. Accessibility leads to an unambiguous retrieval of the information presented on both the macro- and microstructural levels. Any theory of lexicography should present strategies to improve the linguistic quality of dictionaries. However, this should be preceded by strategies to enhance the way in which the target user can identify the data he/she is looking for in order to retrieve the necessary information and to utilise it for decoding or encoding purposes.

Dictionary research has lead to the establishment of different structures of printed dictionaries, e.g. in addition to the macrostructure, microstructure and access structure also the mediostructure. The mediostructure, that is the system of cross-referencing, is a lexicographic device that can be used to establish relations between different components of a dictionary. According to Wiegand (1996: 11) it interconnects the knowledge elements represented in different sectors of the dictionary on several levels of lexicographic description to form a network. Working with a dictionary as a carrier of texts, the mediostructural entries can guide the user between different texts, e.g. between the central text and any text in the front or back matter or between various articles functioning as subtexts in the central word-list. An article-internal mediostructural relation assists the user to relate various microstructural entries employed in the same article.

This paper focuses on different mediostructural strategies and their practical application in general synchronic dictionaries. The structure of dictionaries is discussed from a metalexicographic perspective in order to explain the domain of application of a system of cross-referencing. It is shown how textual cohesion, achieved by the interaction of the various structural components, is promoted by the use of a system of cross-referencing and improved by an innovative approach towards a mediostructure-orientated lexicography. Although the mediostructure of dictionaries is a central topic of this paper, references to

the theory of mediostructures will only cover a small segment of this structural component. A detailed discussion can be found in Wiegand (1996).

Some basic terms relating to a theory of mediostructures

Wiegand (1996) gives an exposition of the fundamental terms employed in a theory of mediostructures. According to his theory, a lexicographer refers the dictionary user from a reference position to a reference address. This is usually done by means of a reference entry and gives the user access to additional relevant lexicographic data. A reference relation is established between the reference entry and the reference address. In *Webster's Ninth New Collegiate Dictionary* (W9) the article of the lemma sign *frog* contains the following entries:

> any of various smooth-skinned web-footed largely aquatic tailless agile leaping amphibians ... — compare TOAD

In this example the specific slot in the article of the lemma sign *frog* is the reference position, and the lemma sign *toad*, the separate macrostructural entry to which the user is referred, is the reference address. Here the reference entry consists of two separate text segments, i.e. the entry marking the reference relation (*compare*), henceforth referred to as the reference marker, and the entry indicating the reference address (*toad*).

A variety of reference markers is used in different dictionaries and often also in one dictionary, e.g. text segments like *see, compare,* →, ⇒, etc. In the English-Dutch translation dictionary *Van Dale Groot Woordenboek Engels-Nederlands* a single arrow is used as one of the reference markers. In the article of the lemma sign **track system** the reference entry "→ tracking" consists of the reference marker "→" and the entry *tracking*, indicating the reference address.

A rather interesting example of cross-reference is found in *Dictionary of Lexicography* (DL) in its treatment of the entries **circular reference** and **reference circularity**. The first is referred to the second, and the second again to the first:

circular reference
⇒ REFERENCE CIRCULARITY

reference circularity
⇒ CIRCULAR REFERENCE

After having fallen victim to this cunning technique by which the user is put into an unending loop he/she will understand exactly what idea is conveyed by "circular reference"!

For the purpose of this article it is important to pay attention to one further

aspect of the theory of mediostructures, i.e. three important types of reference addresses. The first category is that of the internal reference address. With an internal reference address the mediostructural relation does not exceed the boundaries of the article. This type of cross-referencing is used to establish cohesion between different microstructural entries in one article. *Basiswoordeboek van Afrikaans* contains the following article for the lemma sign *frats* (trick/freak):

(1) **frats (fratse)** *nw.* ① Iemand haal fratse uit om mense te vermaak of te laat lag. 'n Mens moet soms vaardigheid hê om dit te kan doen; toertjie, kunsie. *Die kinders het baie gelag vir die fratse van die hanswors by die sirkus. Die aap haal baie fratse uit met die hoop dat die mense vir hom sal grondboontjies gee.* ... ② 'n Frats is iets wat baie ongewoon of onverwags is. Fratse kan ook iets wees wat skielik afwyk van die gewone. *'n Man wat na die kinders kyk terwyl sy vrou werk, sal deur baie mense as 'n frats beskou word. Hier is laasweek 'n fratskalf gebore met twee koppe en vyf bene.* ... **fratsvlieër** (by 1); **fratsbrander, fratsgolf, fratsongeluk** (by 2)

The niched lemmata, included as run-on entries, consist of the unexplained compounds *fratsvlieër* (aerobat), *fratsbrander, fratsgolf* (freak wave) and *fratsongeluk* (freak accident), with the lexical item *frats-* as word initial stem. However, it is not the same sense of the lexical item *frats* that functions in these self-explanatory compounds. Two sense discriminations occur in the treatment of the lemma *frats*. To assist the user in achieving the correct semantic interpretation a relation has to be established between the different niched lemmata and the relevant sense discriminations. One of the ways in which this can be done, is by means of a mediostructural procedure that is aimed at an article-internal address. The lemma sign *fratsvlieënier* is followed by the entry *(by 1)* and the other compounds by the entry *(by 2)*. These are cross-references with the text segment *by* functioning as a reference marker, and the text segments 1 and 2 functioning as reference address indicators. These reference entries guide the user from the reference position to an address in the same article, i.e. the meaning paraphrases presented for the respective sense discriminations.

A second type of reference address is the external reference address. The cross-reference exceeds the boundaries of the article. Two search domains can be identified for external reference addresses. Dictionary articles are texts but they also function as subtexts of the central word-list which is the dominating lexicographic text. The external address can be located either elsewhere in the central word-list, e.g. another lemma sign or a specified microstructural element in another article, or in a separate text outside the central word-list. Compare the articles of *gyro* and *stow* in *Collins Dictionary of the English Language* (CED) and *Longman Dictionary of Contemporary English* (LDOCE) respectively.

(2) **gy·ro** ('dʒaɪrəʊ) *n., pl.* **·ros. 1.** See **gyrocompass. 2.** See **gyroscope.**

(3) **stow** /stəʊ/ *v* [X9 (AWAY)] **1** to put away or pack, esp. for some time: *to stow goods (away) in boxes* **2 stow it!** *sl* Be quiet!

The lexical item *gyro* is polysemous and has two different senses. The article of this lemma sign displays no meaning paraphrase for either of the polysemous senses but cross-refers the user instead to the treatment presented for two other lemma signs, i.e. *gyrocompass* and *gyroscope*. These lemma signs are the external reference addresses located elsewhere in the central word-list. In the article of the lemma sign *stow* LDOCE includes the text segment "[X9(AWAY)]". The X9 refers the user to a text in the back matter of the dictionary which contains a table of codes indicating a variety of grammatical values. X9 is explained in this table as a verb with one object as well as an additional descriptive word or phrase, e.g. *put + it + in the box*. The text element *away* in the quoted text segment is the additional word to be used with the verb *stow*. In this example the text segment X9 is a reference entry indicating an external address located in another text of the dictionary.

Quite often a combination of external and internal reference addresses are given in one reference entry. In *The Concise Oxford Dictionary* the article of the lemma sign *ghosting* contains the following entries:

> the appearance of a 'ghost' (see GHOST *n.* 4) or secondary image in a television picture.

In this example the meaning paraphrase of the lemma sign *ghosting* is the reference position containing a triple address which consists, as the main address, of an external reference address located in the central word-list, i.e. the lemma sign *ghost*, as well as two additional internal addresses, i.e. a secondary address, the nominal function of this lexical item, and a tertiary address, the fourth polysemous sense of this item. The last two reference addresses identify text segments in the article of the lemma sign *ghost*.

The use of a mediostructural strategy of external reference addresses endeavours to enhance the functionality of a dictionary as a source reflecting aspects of the linguistic reality. One of the real problems experienced by the users of alphabetically ordered dictionaries is the decontextualisation of lexical items. Bolinger (1985: 69) maintains that lexicography is an unnatural occupation: "It consists in tearing words from their mother context and setting them in rows — carrots and onions and beetroot and salsify next to one another — with roots shorn like those of celery to make them fit side by side, in an order determined not by nature but by some obscure Phoenician sailors who traded with Greeks in the long ago." He continues this argument by saying that "half of the lexicographer's labor is spent repairing this damage to an infinitude of natural connections that every word in any language contracts with every other word, in a complex neural web knit densely at the center but ever more diffusely as it spreads outward". According to him "a bit of context, a synonym, a grammatical category, ... and a cross-reference or two" are "the additives that accomplish the repair".

From both a semantic and a pragmatic perspective the lexicon has to be regarded as an ordered set of lexical entries. However, the *alphabetical ordering* of a dictionary defies the network of semantic relations existing between this set of lexical entries. The mediostructure of a dictionary is a powerful mechanism to re-establish some of the lexical relations. Dictionaries employ the mediostructure to refer the user to external addresses which are linked with the lemma sign of the reference position article in relations such as synonymy, oppositeness of meaning, hyponomy, dialectal, stylistic, chronolectic and other forms of variation, etc. For the language learner as well as the seasoned native speaker of any given language these cross-references represent an added value which assists them in improving their communicative potential. South African dictionaries should employ external reference addresses in a more general and consistent way. However, it is of extreme importance that these strategies be explained comprehensively in the front matter of the dictionary.

The third category of reference address is the dictionary external reference address. This mediostructural procedure links a text segment in a dictionary to a source outside the dictionary. In *A Dictionary of Language Planning Terms* Cluver (1993) puts the strategy of dictionary external reference to good use. The back matter of the dictionary contains a bibliography of sources in which more information regarding the terminology treated in the dictionary can be found. Many articles contain condensed bibliographical references which leads the user to the bibliography in the back matter. This is the reference position from where the user is guided by means of a complete reference to the specific source. The condensed bibliographical references in the articles are clearly indicated by the reference marker "**Bibl.**". In the article of the lemma sign *primary language* the following text segment is found: "**Bibl.** Mühlhäusler 1986: 9". The bibliography gives the full reference, i.e. "Mühlhäusler, P. 1986. *Pidgin and creole linguistics.* Oxford: Basil Blackwell." By means of the dictionary external reference address the lemma sign is linked to this external source. A variety of other reference addresses can also be identified but they are not relevant for the present discussion.

For the African languages, apart from the disruption of semantic relations, *alphabetical ordering* has serious detrimental consequences for *grammatical relations*. Many traditional compilers, although following an alphabetical ordering in principle, regard the importance of combined semantic and grammatical cohesion as too important to break.

This view implies that in the case of African languages the mediostructure is incapable of re-establishing the most relevant lexical relations. In most dictionaries this results in a hybrid approach where different derivations, sometimes a hundred or more, of a single word are treated within the article of a nominal or especially verbal stem in a complex article with numerous sublemmas and sublemmatic addresses, in addition to being entered as separate lemmas in their appropriate alphabetical positions.

(4) **RÊKA** (–rêka, –rêkilê, –rêkwa, –rêkilwê) koop, aankoop, ruil // buy, purchase, barter; ~ *polasa* in weelde lewe // live in comfort/luxury; ~ *o lebêlêtše godimo* kat in die sak koop // buy a pig in a poke; *nku e rêkwa mosela* 'n mooi geboude dame is 'n aantrekkingskrag vir jongmans // a lady with a good figure easily attracts young men; **dirékárekáne (dirêkarêkane)** verskeidenheid gekoopte goedere // variety of things bought; **leréko, ma– (lerêkô)** gewoonte/neiging om te koop // habit of buying, inclination to buy; **morêki, ba– (morêki)** pers. dev.; koper // buyer, purchaser; **seréki, di– (serêki)** pers. dev.; lustige koper // keen buyer; **seréko, di– (serêkô)** impers. dev.; wat gekoop word, aankope // purchase(s); **théko, (n–)/di– (thêkô)** man. dev.; koopwyse, prys // manner of buying, price; *RÊKÁNA* (–rêkana, –rêkane, –rêkanwa, –rêkanwe) rec.; ruil met mekaar // exchange with one another; *a re rêkanê, wêna o mphê hêmpê êla, nna ke go fê dîêta tšê* laat ons met mekaar ruil, jy gee my daardie hemp en ek gee jou hierdie skoene // let us exchange. you give me that shirt, I will give you these shoes; **barékáni (barêkani)** pers. dev.; **thékáno, (n–)/di– (thêkanô)** man. dev.; *RÊKÁNTŠHA* (–rêkantšha, –rêkantšhitšê, –rêkantšhwa, –rêkantšhitšwê) caus. < *RÊKÁNA*; (om)ruil, wissel (geld), inruil // exchange, barter, trade in, swop; **morékántšhi, ba– (morêkantšhi)** pers. dev.; **serékántšhwá, di– (serêkantšhwa)** impers. pass. dev.; **thékántšho, (n–)/di– (thêkantšhô)** man. dev.; omruiling, inruiling, wisseling // exchange, bartering, swopping; *RÊKÁNYA* (–rêkanya, –rêkantšê, –rêkanywa, –rêkantšwê) caus. < *RÊKÁNA*; (om)ruil, wissel (geld) // exchange, barter, swop; **morékányi, ba– (morêkanyi)** pers. dev.; **serékanywá, di– (serêkanywa)** impers. pass. dev.; **thékányo, (n–)/di– (thêkanyô)** man. dev.; v. **thékántšho**; *RÊKÉGA* (–rêkêga, –rêkêgilê) neutr.; koopbaar w. // b. purchasable; *RÊKÉLA* (–rêkêla, –rêkêtše, –rêkêlwa, –rêkêtšwe) appl.; koop vir // buy for; ~ *kolobê kgetsing* (< Afr.) kat in die sak koop // buy a pig in a poke; **borékélo (borêkêlô)** lo. dev.; koopplek // place where things are bought; **morékédi, ba– (morêkêdi)** pers. dev.; **morékelwá, ba– (morêkêlwa)** pers. pass. dev.; **serékélo, di– (serêkêlô)** impers. dev.; iets waarin jy koop // that into which one buys; **thékélo, (n–)/di– (thêkêlô)** man. dev.; maat, skaal (waarin bv. bier gekoop word) // measurement, bowl (one used for buying beer); *RÊKÉLANA* (–rêkêlana, –rêkêlane, –rêkêlanwa, –rêkêlanwe) appl. rec.; **barékélani (barêkêlani)** pers. dev.; **thékélano, (n–)/di– (thêkêlanô)** man. dev.; *RÊKÍŠA* (–rêkiša, –rêkišitšê, –rêkišwa, –rêkišitšwê) caus.; laat/help koop, verkoop, van die hand sit // cause/help buy, sell; ~ *ka leleme* kul, mislei, verdraai // deceive, mislead, pervert; ~ *leleme* praatsiek w., skinder // gossip, b. loquacious, b. garrulous; ~ *motho a sa phela* iemand kul // deceive someone; ~ *motho lebake* iemand kul, 'n tevergeefse belofte maak, iemand verag weens sy slegte gedrag // deceive someone, give a vain promise, despise someone because of his bad conduct; ~ *segaé* iets aan iemand so verkoop dat hy 'n goeie slag slaan omdat jy sy vriend of familielid is, afslag gee // sell to someone at bargain price because he is your friend/relative, give discount; **morékíši, ba– (morêkiši)** pers. dev.; verkoper, verkoopsman, winkelier // seller, salesman, storekeeper; **serékišwá, di– (serêkišwa)** impers. pass. dev.; **thékišo, (n–)/di– (thêkišô)** man. dev.; verkoping, uitverkoping, afset, bemarking // sale, selling, market, marketing; *RÉKÍŠANA* (–rêkišana, –rêkišane, –rêkišanwa, –rêkišanwe) caus. rec.; ruil met mekaar // exchange with one another; **barékíšani (barêkišani)** pers. dev.; **thékíšano, (n–)/di– (thêkišanô)** man. dev.; *RÉKÍŠEGA* (–rêkišêga, –rêkišêgilê) neutr. < *RÉKÍŠA*: verkoopbaar w. // b. sellable; *RÉKÍŠETŠA* (–rêkišêtša, –rêkišêditšê, –rêkišêtšwa, –rêkišêditšwê) caus. appl.; verkoop vir // sell for; **borékíšetšo (borêkišêtšô)** lo. dev.; verkoopplek // selling place; **morékíšetši, ba– (morêkišêtši)** pers. dev.; †agent // †(business) agent; **thékíšetšo, (n–)/di– (thêkišêtšô)** man. dev.; *RÉKÍŠETŠANA* (–rêkišêtšana, –rêkišêtšane, –rêkišêtšanwa, –rêkišêtšanwe) caus. appl. rec.; sake verrig // transact business; **barékíšetšani (barêkišêtšani)** pers. dev.; **thékíšetšano, (n–)/di– (thêkišêtšanô)** man. dev.; besigheidstransaksie // business transaction; *RÉKÓLLA* (–rêkolla, –rêkolotše, –rêkollwa, –rêkolotšwe) rev. tr.; terugkoop, terugruil, geld terugvra, los // buy back, exchange back, ask for a refund, redeem; **morékólli, ba– (morêkolli)** pers. dev.; **serékóllwá, di– (serêkollwa)** impers. pass. dev.; 'n ding wat teruggekoop word // that which is bought back; **thékóllo, (n–)/di– (thekollô)** man. dev.; (Bl.) lossing // (Bl.) redemption; *RÉKÓLLANA* (–rêkollana, –rêkollane, –rêkollanwa, –rêkollanwe) rev. rec.; **barékóllani (barêkollani)** pers. dev.; **thékóllano, (n–)/di– (thêkollanô)** man. dev.; *RÉKÓLLELA* (–rêkollêla, –rêkollêtše, –rêkollêlwa, –rêkollêtšwe) rev. appl.; **morékólledi, ba– (morêkollêdi)** pers. dev.; **thékóllelo, (n–)/di– (thêkollêlô)** man. dev.; *RÉKÓLLELANA* (–rêkollêlana, –rêkollêlane, –rêkollêlanwa, –rêkollêlanwe) rev. appl. rec.; **barékóllelani (barêkollêlani)** pers. dev.; **thékóllelano, (n–)/di– (thêkollêlanô)** man. dev.; *RÉKÓLLIŠA* (–rêkolliša, –rêkollišitšê, –rêkollišwa, –rêkollišitšwê) rev. caus.; **morékólliši, ba– (morêkolliši)** pers. dev.; **thékólliše, (n–)/di– (thêkolliše)** man. dev.; *RÉKÓLLIŠANA* (–rêkollišana, –rêkollišane, –rêkollišanwa, –rêkollišanwe) rev. caus. rec.; **barékóllišani (barêkollišani)** pers. dev.; **thékóllišano, (n–)/di– (thêkollišanô)** man. dev.

Thus in dictionaries such as *Groot Noord-Sotho Woordeboek* (GN) word stems and their derivations are clustered together in one huge entry with the noun or verbal root as the lemma often containing up to eighteen levels of sublemmas. Compare example (4). Where derivations are entered separately in their appropriate alphabetical positions in GN, only minimal grammatical information is given and a reference back to the cluster.

(5) **thékóllano**, (n–)/**di**– v. **RÉKA**
thékóllelano, (n–)/**di**– v. **RÉKA**
thékóllelo, (n–)/**di**– v. **RÉKA**
thékóllišano, (n -)/**di**– v. **RÉKA**
thékóllišo, (n–)/**di**– v. **RÉKA**
thékóllo, (n–)/**di**– v. **RÉKA**

In this way mediostructure is exhausted/overused for the sole purpose of maintaining structural links. Little or no realization of mediostructure as a powerful access structure is achieved. Once referred back to the main cluster (4), it is unlikely that the user will be able to work out the meaning, especially for those cases which lie relatively deep in the modular structure as in the case of **dithekollišano**. The user has to look up this word under the singular **thekollišano** in (5) and is then referred to **reka** in (4) and eventually, after having struggled through this lengthy article, he finds **thekollišano** at the end of (4) with no translation equivalents given. (Compare Prinsloo (1994) for similar examples and a detailed discussion on problematic aspects of the lemmatization of verbs.)

This obsession with keeping together what in their view semantically and grammatically "belong together" thus results in extremely user-unfriendly entries in which successful retrieval of information virtually becomes impossible. It could be argued that the utilization of cross-references simply for the sake of grammatical binding is nonfunctional.

It was stated in the introduction that one factor in the evaluation of a dictionary is the extent to which it is useful to the user. Dictionaries such as these fail in this main criterion. Students consequently opt for less sophisticated dictionaries with less information categories and less exhaustively treated lemmas, i.e. a lower density of information.

Cross-referencing has not been employed to its full potential in dictionaries for most African languages. Typical errors and shortcomings will be briefly outlined below.

Consider the treatment of **molelo** versus **mollo** in *New English Northern Sotho Dictionary* (NEN):

(6) (a) **mo'lelo**, see: **mollo**.
(b) **'mol'lo**, n., fire, witch-weed, principal wife; ...
(c) **mol'lo**, n., cry, (manner of) crying.

In (6) the exact reference address to which the user is referred from (6)(a) is uncertain. The second entry for **mollo**, (6)(c), has no relation to **molelo** whatsoever. To make things worse, the reference in (6)(a) is to **mollo** instead of **'mollo**. The convention " ' " is also not explained in the dictionary. Since **mollo** in (6)(b) and **mollo** in (6)(c) represent a fairly rare situation where two words in Sepedi can neither be distinguished phonetically (both **mollô**) and having an identical *tonal pattern* (low-low-low), the compiler should therefore distinguish by means of homonym numbers, i.e. **mollo**[1] and **mollo**[2]. The decision of the compiler to refer the user who looks up **molelo** to **mollo** without treating **molelo**, is however acceptable in terms of frequency-of-use criteria since **mollo** is frequently used and **molelo** not. Thus no cross-reference from **mollo** to the less frequently used **molelo** is necessary or appropriate because the target user of this dictionary is looking for translation equivalents in the target language and is not interested in (more) information in the source language. However, within the article of **mollo**[1], reference should be made to **molelo** but then labeled as *dialectical* or treated by means of *inserted text*. Compare the additional information given by means of inserted text in *Reader's Digest English–Afrikaans Dictionary* (RD) in the case of **rekenaar** versus **komper**:

(7) **kom'per** =s computer; *vid.* **rekenaar, rekenoutomaat.**

> WORDS IN ACTION
> ***komper, rekenaar, rekenoutomaat***
>
> *Komper* (computer) is the word used by some speakers and writers in the Western Cape. Other people there and most people elsewhere in South Africa use *rekenaar*. *Rekenaar* is also the word preferred by people in the computer profession. ...

The treatment of **bracket** versus **brackets** in *Northern Sotho Terminology and Orthography* (NTO) can now be considered:

(8)	(a) bracket (symbol) (see: brackets)	hakie	lešakana
	(b) brackets	vierkantige hakies	mašakanakhutlwana

In (8)(a) translation equivalents in Afrikaans and Sepedi are given for **bracket**. The cross-reference to **brackets** is quite appropriate since the latter is more frequently used. Also, due to considerations regarding frequency of use, no reference from **brackets** to **bracket** is necessary. However, in looking up **brackets**, the user does not get any additional information, e.g. in respect of types and use of brackets. On the contrary, he/she is misguided by the additional information given at the reference address namely that the lemma **brackets** is

translated in Afrikaans and Sepedi as necessarily *square*. Thus, in contrast to the singular **bracket**, the plural form excludes other types of brackets.

It is important that the user should find more information at the reference address, otherwise the value of cross-referencing is devalued. Cross-reference, or more specifically, the position of a cross-reference entry, indicates to the user that this is the starting-point in the process of information retrieval. The usage frequency of the item which stands in the cross-reference position, is lower than the reference address. The lexicographer may never utilize the system of cross-referencing simply because he/she does not want to give proper treatment to the items in question. If it is in the interest of the target user, a specific lemma should be entered and treated. Cross-references such as those attempted in (6) and (8), will have a negative effect on the target user. Once disappointed, it will discourage him/her from following up cross-references since it is impossible to distinguish between functional and nonfunctional references in the dictionary.

Consider also NTO's treatment of **complainant** versus **plaintiff**:

(9)	(a) complainant (see: plaintiff)	klaer	mmelaedi, molli, mmegi
	(b) plaintiff	eiser, klaer	mmelaedi, molli, mmegi, motlalei

It is unclear why no cross-reference from **plaintiff** to **complainant** is given. Such a reference is necessary because equivalents in both target languages are given under **complainant**. The addition of the translation equivalents **eiser** in the Afrikaans column and **motlalei** in the Sepedi column also raise a few questions. Firstly, it implies that **eiser** and **motlalei** are suitable equivalents for **plaintiff** but not for **complainant**. Secondly, to give **eiser** as the first translation equivalent for **plaintiff** suggests that it is the best option. However, although it is added in the case of **motlalei** to the translation equivalent paradigm in (9)(b) for the sake of **eiser**, it is given at the end. Thus the entire relationship between **complainant** and **plaintiff** becomes unclear. The user cannot determine in which relation they stand to each other. Central text-internal reference should strengthen the cohesion, as is correctly done in the case of **molelo** versus **mollo** in (6) above. In the case of (9) this cohesion is actually broken off. The user who wants to find translation equivalents in Afrikaans and Sepedi is referred to another word where the same treatment is given for no reason.

An even more confusing example of cross-reference in NTO is its treatment of **Brave West Wind** versus **anti-trade wind**:

(10)	(a) Brave West Wind (see: Anti-trade wind)	Antipassaatwind	Phefomadibakgwebo
	(b) anti-trade wind	antipassaatwind	phefomadibakgwebo, diphefo tša bodikela

Cross-reference is given from (10)(a) to (10)(b) but not vice versa. The compilers are not consistent in the use of capital versus lower case letters in Anti-/anti- and Phefo-/phefo. Ironically reference to the "West" in Sepedi, namely, **tša bodikela**, is added to the translation equivalent paradigm of (10)(b) instead of (10)(a).

In addition to implicit cross-referencing, two types of explicit cross-referencing are used in *Thanodi ya Setswana* (TS) namely:

¦ what follows differs from the explained.
BÔNA ("SEE") the following is related to the explained

Compare the entries for **kgarebê** and **lekgarebê** in TS and *Thanodi ya Setswana ya Dikole* (TSD).

(11) **kgarebê** TTT ¦lekgarebê *ln./9.* ma-. mosetsana yo o godileng mme a ise a nyalwe
lek·garebê TTTT *ln./5.* ma-. 1. mosetsana yo o lekaneng go nyalwa 2. mosetsana yo o itlhôkômêlang a apara sentlê

The examples under (11) from TS, explicitly referring the user from **kgarebê** to **lekgarebê** is sensible since apart from the meaning "girl who can be married" which is similar to that given in sense 1 of **lekgarebê**, an *extended* meaning "a neatly, well-dressed girl" is given as sense 2. The fact that no explicit reference from **lekgarebê** to **kgarebê** is given, is also quite acceptable since the user who looks up **lekgarebê** will not find any new information under **kgarebê**. However **kgarebê** must be given as a synonym directly following the sense 1 definition. TS's treatment of **kgarebê** can also be improved in respect of the *position* allocated to the reference entry. The explicit cross-reference ¦**lekgarebê** should not be given in the focus position of the article. It can be regarded as an unnecessary or even nonfunctional cross-reference interfering with the user's information retrieval process. Formulated differently, the information primarily needed by the user who looks up **kgarebê** is that given in the *definition*. Once given the definition, he/she might be interested to consult the reference address for additional information. It can also be argued that the wrong reference symbol is used in the case of **kgarebê** versus **lekgarebê**. The relation is one of relatedness rather than difference — thus in terms of TS's conventions "BÔNA" rather than "¦".

(12) **kgarebê(ma)** mosetsana yo o godileng mme a ise a tsewe (nyalwe).
lekgarebê(ma) kgarebê; mosetsana yo o ka tšewang.

TSD's treatment of the same words are shown under (12). In the case of **kgarebê** only a definition is offered, while a synonym as well as a definition is given for **lekgarebê**. Since no cross-reference is made from **kgarebê** to **lekgarebê**, it

suggests that **kgarebê** is the entry with the higher usage frequency. However, the user gets more information from **lekgarebê**, namely a synonym as well as a definition, than from **kgarebê**. This is confusing. In terms of cross-reference it can be said that the article of the lemma sign **lekgarebê** is a reference position of the reference entry **kgarebê**. Normally, for economical reasons, the same definition is not given in two places. Two definitions and the lack of cross-reference has a negative effect on cohesion. Here the user cannot establish which one is the more frequently used. The more frequently used word is the one likely to be treated. This in itself is an indication of higher frequency of use. It would thus be better to enter **kgarebê** with a definition, adding **lekgarebê** as a synonym. It is normal practice to give a list of synonyms after the definition since they meet the criteria to be lemmatized themselves. Such synonyms can be listed in order of frequency of use if such criteria is available or otherwise alphabetically. Thus, since all synonyms have to be entered as lemmas, **lekgarebê** will be entered as a lemma sign but only with a cross-reference to **kgarebê**.

It is also not clear why in both TS and TSD the definitions differ in respect of the concept "grown up". In the case of **lekgarebê** "a girl who can be married" and in the case of **kgarebê** "a grown-up girl, one who is not yet taken/married". When comparing the two, the user can get the wrong impression that **kgarebê** implies an adult and **lekgarebê** not.

Cross-references from the front matter, especially from the user's guidelines to the central text are crucial to the user for successful or optimal retrieval of information. *Dead references*, especially in the *guidelines* of a dictionary are defects which undermine the trust of the user in the dictionary as a reliable source of information, and in the value of the cross-referencing system as a whole. Such dead references often do not effect only one reference address, a key to a whole section can be lost. Consider the following example: In the guidelines to *A Learner's Chichewa and English Dictionary* (LCE), the compilers explain the policy not to lemmatize derived forms when the meaning is readily ascertainable from the root plus suffix combination. In support of this far-reaching decision for lemmatization of an African language, they include cross-references in the central text: "Thus, both **-mva** 'hear, understand' and its derived form **-mvana** 'get along together' are listed". However, the *very examples* that they quote to illustrate their policy, are not treated as such: **-mva** is listed but not **-mvana**. This dead reference to **-mvana** can cause the user to doubt the treatment policy not only in respect of a single entry but a whole category of entries. A similar dead reference occurs in the next sentence: "the derived verb **-mverana** 'listen to each other' is not listed because its meaning is readily determined from the root **-mvera** 'listen to' plus an affix". However, again the root **-mvera** is not listed, clearly violating the claim "verbs are entered according to their root forms".

The treatment of cross-references in the *Dictionary of Northern Sotho Grammatical Terms* (NGT) can now be considered. This dictionary is a pioneering first for Sepedi and very popular among its target users.

(13) **tone** *(segalô, toon)*
Tone can be defined as *pitch variations* which affect the meaning and function of words. *Tone* is one of the distinctive features of the Bantu language family (see *Bantu languages*), and in these languages differences in *tone* between words which have exactly the same shape, result in a difference in meaning. Two basic tones (also called *tonemes*) are usually distinguished, namely a *high tone* and a *low tone*, although more detailed distinctions are often drawn between, for example *rising* and *falling tones*, *mid*, *mid-high* and *mid-low* tones, etc. A *tone* (or *toneme*) is always associated with a particular *syllable*, i.e. there are as many *tones* in a word as there are syllables since *tones* realise on *vowels*. This is one of the reasons why vowels are often referred to as *syllable nuclei*. (See: *nucleus.*) ...

In this article of **tone** explicit reference is made to **Bantu languages** and **nucleus**. At the reference address, **Bantu languages**, the user finds more useful information on tone in the African languages. Likewise, the user who consults the entry **Bantu languages** first will find, in addition to other useful information given there, "tone plays a distinctive role. See **tone**". This is good lexicographical practice since for the user who consults the entry **Bantu languages**, as well as for the user who looks up **tone**, the cross-references are useful. Both contain more information at the respective reference addresses with regard to two important and closely related issues such as **African languages** and **tone**. The same holds true for the explicit reference made to *tone* in the article of **syllabic nasal**. The treatment of **syllable** and **tonal pattern** as reference positions of explicit reference to the addresses **nucleus** and **tone** respectively, can however be improved.

(14) **syllable** *(noko, sillabe/lettergreep)*
See *nucleus.*

(15) **tonal pattern** *(*patrone ya segalô, toonpatroon)*
See *tone*

Firstly, **syllable** in (14) deserves full treatment, especially in a dictionary of grammatical terms. Apart from translation equivalents in Sepedi and Afrikaans, no definition is given, only an explicit reference to **nucleus**. In the article of **nucleus**, many references are once again made to **syllable**, such as "[a nucleus] is used to characterize the nature of a *syllable*". As for **syllable**, it is maintained that "vowels form the nuclei of syllables", etc. However, **syllable** itself remains undefined. Thus **syllable** deserves a definition and treatment as for example in (16) from *South African Student's Dictionary* (SSD) and in (17) from *New Student's Dictionary* (NSD):

(16) **syllable** ... *noun*: **syllables**
A **syllable** is any of the parts, consisting of one or more sounds and usually including a vowel or a consonant acting like a vowel, that a spoken word can be divided into: *The word 'telephone' has three syllables, 'te', 'le', and 'phone', and 'tiger' has two, 'ti' and 'ger'.*

(17) **syllable** ... **syllables**. N-C A **syllable** is a part of a word that contains a single vowel-sound and that is pronounced as a unit. For example, 'book' has one syllable, and 'reading' has two syllables.

(18) **toneme** *(*segalwana, toneem/toonfoneem)*
See *tone.*

An important statement in the article of **tone** in (13) reads: "Tone is always associated with a particular *syllable.*" The user of NGT consulting **tone** could easily perceive the italicized word *syllable* as an implicit reference entry but find it to be nonfunctional since in looking up **syllable** in (14), he/she is referred to *another* address namely **nucleus**. Furthermore, although **tone** is one of the key issues discussed in the article of **nucleus**, no *explicit* reference is given to **tone**.

In the case of **tonal pattern**, the user is referred to **tone** but the distinction between **tone** and **tonal pattern** is unclear. From phrases such as "depending on its tone or tonal pattern", it is not clear whether *or* means "equal to" or "in contrast to". The user who wishes to know the meaning of **tonal pattern** is referred to **tone** but will not know for sure after having studied the treatment of **tone** whether **tone** and **tonal pattern** is synonymic or not. In the case of **tone** versus **toneme** in (13) and (18), *or* in the phrase "a tone or toneme" means **tone** is equal to **toneme**. It should rather be clearly stated that **tonal pattern** is a series of tones/tonemes. This could be explained by using **mosadi** as an example, where **mo-** has a low tone, **-sa-** a high tone and **-di** again a low tone. The tonal pattern of these three tones/tonemes is therefore low-high-low, often indicated as LHL. This suggests that **tonal pattern** deserves to be treated on its own. In Wiegand's (1996) terms, it means that if the user is referred from **tonal pattern**, which is the reference position, to the article of the lemma sign **tone**, the reference address, more information on/a fuller treatment of **tonal pattern** must be given. Thus the cross-reference from **tonal pattern** to **tone** is not observed in the sense that **tonal pattern** is not really treated within the article of **tone**. The purpose and value of the cross-reference is lost.

Finally, key terms used in the treatment of the lemma **tone** which are italicized such as *pitch variations*, *tonemes*, and especially *syllable* are not treated in the dictionary. The user expects a clearer distinction between *implicit reference* to a different reference address, on the one hand, and mere instances of *emphasis* on the other.

This does not mean that the lexicographer should solely utilize explicit references to distinguish between emphasis and cross-referencing, since there is

no fundamental value difference between explicit and implicit reference systems. The former is only more obvious than the latter. Thus it is suggested that the lexicographer should utilize both as long as implicit references can be clearly distinguished from mere emphasis. The implicit cross-reference strategy must however be clearly apparent. This means that terms used within the articles of entries which are themselves lemmatized and treated elsewhere in the dictionary, must stand out and be treated consistently.

The mediostructure has not in all instances been employed to its full potential. In a dictionary of grammatical terms, the mediostructure could be employed as a powerful access structure by ensuring that *at least all keywords* used within the treatment of a specific lemma *which are themselves entered as lemmas in the same dictionary,* are marked for cross-reference. *Dictionary of Lexicography* (DL) can serve as an excellent example in this regard:

(19) **lexicography**
The professional activity and academic field concerned with DICTIONARIES and other REFERENCE WORKS. It has two basic divisions: lexicographic practice, or DICTIONARY-MAKING, and lexicographic theory, or DICTIONARY RESEARCH. ...

It can rightfully be argued that the lexicographer should guard against excessive text condensation. However, opportunities should be utilized to strengthen the cohesion of the dictionary by optimal organization of the mediostructure as an access structure.

An excellent example in African language lexicography where mediostructure has been employed as a powerful access structure is the *Lexicon Cilubà–Nederlands* (LCN) compiled by De Schryver and Kabuta. This dictionary is highly successful in interconnecting the knowledge elements represented in different sectors of the dictionary on several levels of lexicographic description to form a network.

In contrast to GN, for example, the compilers of LCN are aware of the benefits of "keeping together what semantically and grammatically belong together" but also of the need (a) to avoid extremely long entries and (b) to ensure proper treatment of each derivation in terms of grammatical, tonal and lexical information. Compare the entries in LCN for **-funda** and its derivations:

(20) **-fùnda I** [tww] **1** schrijven; aantekenen; **2** aanklagen; **II** [adj ◄ **I**] 't geschrevene; ***mwakù mu~*** *'t geschreven woord*
► -dìfùndisha; -fùndangana; -fùndangeena; -fùndiibwa; -fùndika; -fùndila; -fùndilangana; -fùndisha; -fùndishangana; -fùndishiibwa; -fùndishila; -fùndishilangana; -fùndishisha; -fùndulula; kafùndilà

(21) **-fùndangana** [tww, ass **-fùnda**] **1** elkaar schrijven; **2** elkaar, iemand aanklagen
-fùndangeena [tww, ass app **-fùnda**] elkaar aanklagen ... < + plaatsbepaling>

-fùndiibwa [tww, pas **-fùnda**] geschreven z/w
-fùndika [tww, imp **-fùnda**] schrijfbaar z
-fùndila [tww, app **-fùnda**] schrijven voor, op, aan, om, naar; toeschrijven
-fùndilangana [tww, app ass **-fùnda**] aan elkaar schrijven
-fùndisha [tww, cau **-fùnda**] doen schrijven; inschrijven

In contrast to cases such as **dithekollišano** above, the user can find even complicated derived words such as **fundilangana** firstly lemmatised separately in its proper alphabetical position and secondly fully treated. The user does not have to refer back to the stem entry **funda** to find the necessary information. An implicit cross-reference is nevertheless given to the root **-funda** where all the relevant derivations are listed. The compilers of LCN thus succeeded in harmonising lumping and splitting, capturing the advantages of both these approaches. It can, of course, be argued that the listing of the different derivations occupies precious space in the dictionary. However, by substantially reducing the font size, this redundancy is diminished.

Thus the compilers not only succeeded in linking stems and derivations and treating both stems and derivations satisfactorily, but they also employed a complex system of cross-referencing:

> De klassieke opdeling macro- vs. micro-structuur wordt helemaal opengebroken door een doorgedreven netwerk van verwijzingen. Inderdaad, zowel vanuit een slot tussen **rechte haken**, als vanuit een slot van **vertalingen/omschrijvingen**, als vanuit een slot van **commentaar**, als vanuit een slot van **voorbeelden**, als vanuit het slot van een **samenstelling met lemmastatus**, als vanuit een slot van **versteende uitdrukkingen met lemmastatus**, als vanuit het slot van de **staart**, kan men verwijzingen vinden naar of elementen uit de **macro-**, of elementen uit de **micro-**structuur van een ander artikel! De meeste van deze verwijzingen werden ofwel reeds impliciet behandeld, of zijn zo evident dat ze geen verdere uitweiding vereisen. Het zij voldoende te vermelden dat hiervoor de volgende "SYMBOLEN & VERWIJZINGEN" worden gebruikt: ~, ◄, ►, □, ⌂, afk X, ant X, cf X, syn X, var X, vgl X; en ook Romeinse en/of Arabische cijfers. Op enkele details na, zijn AL deze verwijzingen ook **kruisverwijzingen**!

Cross-references, whether explicit or implicit, text-internal or text-external, are given from all possible slots of an article. See the following *seven* typical reference positions:

(22) (a) A [...] slot
Compare **-fùnda** in the article of **-fùndisha** in (21) above.
(b) A translation/description slot
Compare **nswà** in the article of **ciswà (-munène)**:
ciswà (-munène) ... **2** maanmaand gedurende dewelke **nswà** uitvlie gen ...

(c) A comment slot
Compare **lupòse** in the article of **kabangu**:
kabangu ... < ... de larve v deze kever heet **lupòse>**.

(d) An example slot
Compare **-dìngisha** in the article of **-enza**:
-enza ... ~ **bu** [ud; syn **-dìngisha**] ...

(e) A "compound with lemma-status" slot
Compare **-kàjì** in the article of **bakàjì**:
bakàjì ... □ **cn** ~ [cn adj; var **-kàjì**] ...

(f) A "fossilised expression with lemma-status" slot
Compare **à.n.** in the article of **ànu**:
ànu ... ◊ **1** ~ **nànku** [ud; afk **à.n.**] enzovoort; ...

(g) A tail slot
Compare ▶ -dìfùndisha; ... in (20) above.

The endeavour to achieve an optimal transfer of information

Dictionaries are containers of knowledge (cf. McArthur 1986). Although lexicographers have to take this into account, they should also be alert to the fact that a dictionary has to be compiled in such a way that the intended target user can employ it as a practical linguistic instrument. One of the components of a dictionary aimed at a better retrieval of information by the target user is the access structure. The access structure can be regarded as the search route of the user on his way to the lexicographic data needed. The internal access structure, that is the search route followed within the article, can display a variety of so-called structural markers. These markers signpost various microstructural data categories. Because reference markers indicate the reference entry, it can be argued that they are also part of the access structure, functioning as nontypographical structural markers. A reference marker does not only indicate the fact that a specific text segment relates to another text segment but it sometimes also explicates the type of relation that holds between the two segments. In the W9 different strategies, elucidated in the explanatory notes in the front matter, are used to accomplish successful cross-referencing. One kind of cross-reference used is the reference to a variant of the lemma. The reference position in this mediostructural category accommodates the marker *var. of*. In the article of the lemma-sign *inclose* a cross-reference "*var. of* ENCLOSE, ENCLOSURE" explicates the kind of mediostructural relation between the lemma-sign and the reference address. This lexicographic procedure does not only constitute a valuable type of cross-referencing but it also assists in presenting the lexicon as a structured collection characterised by a network of internal relations.

Within a multilingual and multicultural society dictionaries have an important role to play as instruments to promote mutual understanding and communicative competence. South African lexicographers should employ all available strategies to create a dictionary culture and to enhance the dictionary-

using skills of their intended target users. This approach compels lexicographers to structure their dictionaries in such a way that the retrieval of information exceeds the traditional domains. It will always be important to find comments on specific lexical items in a dictionary and it will always be important to find information linking a specific lemma or a microstructural entry like a sense discrimination, to other text segments in the dictionary. However, dictionaries compiled for the South African linguistic environment should go further than this. Besides information regarding a specific lemma or article component as treatment unit, it is of vital importance that dictionaries should expose the underlying system by focusing on a lexical item as part of an overall linguistic or grammatical pattern.

According to Jackson (1985: 53) grammar and dictionary are complementary parts of the overall description of language. However, the average member of a speech community uses a dictionary much more than a grammar. One of the assignments of the lexicographer in a multilingual society is to make his/her target user aware of aspects regarding both the lexicon and the grammar of the specific language. The first step to achieve this goal is to include a mini-grammar as a separate text in the front or back matter of the dictionary (cf. Gouws 1989). The fact that the average dictionary user focuses his attention exclusively on the data presented in the central word-list, compels the lexicographer to employ innovative strategies to ensure a successful utilisation of the grammar as one of the other texts in the dictionary. The most obvious strategy would be the establishment of text-external mediostructural relations between the central list and the mini-grammar. Dictionaries compiled for use in South Africa should be text carriers that include, among others, separate texts in which the grammatical system of the treated language is explained. Lexicographers have to employ an extended mediostructural application to guide their users from a variety of reference positions in the central list to specific reference addresses in the mini-grammar.

The front matter texts should also include a systematic exposition of other language-specific characteristics and these texts have to be addressed from the central list by means of a well-developed mediostructural network.

In conclusion

There is nothing as practical as a good theory. Therefore the success of a dictionary as a practical instrument depends on its theoretical basis. If a lexicographer cannot base his practical applications on sound theoretical principles, the dictionary is bound to be of a lesser quality. Knowledge of the structural components of a dictionary as a carrier of texts equips the lexicographer with the expertise to produce a better dictionary. Understanding the importance of cross-referencing as a functional lexicographic device enables the lexicographer to compile a dictionary which offers the target user friendly access to participa-

tion in the language game. This is desperately needed in the multilingual and multicultural South Africa.

References

Allen, R.E. (Ed.). 1990[8] *The Concise Oxford Dictionary*. Oxford: Clarendon Press.

Bolinger, D. 1985. Defining the Indefinable. Ilson, R. (Ed.). 1985: 69-73.

Botne R. and A.T. Kulemeka. 1995 *A Learner's Chichewa and English Dictionary*. Cologne: Rüdiger Köppe.

Chambers–Macmillan. 1996. *South African Student's Dictionary*. Manzini: Macmillan Boleswa.

Collins Cobuild. 1997. *New Student's Dictionary*. Birmingham: HarperCollins.

Cluver, A.D. de V. 1993. *A Dictionary of Language Planning Terms*. Pretoria: University of South Africa.

Department of Education and Training. 1988. *Northern Sotho Terminology and Orthography*. Pretoria: Government Printer.

De Schryver, G.-M. and N.S. Kabuta. 1997. *Lexicon Cilubà–Nederlands*. Ghent: Recall.

Gouws, R., I. Feinauer and F. Ponelis. 1994. *Basiswoordeboek van Afrikaans*. Pretoria: J.L. van Schaik.

Gouws, R.H. 1989. *Leksikografie*. Cape Town: Academica.

Grobbelaar, P. (Ed.). 1987. *Reader's Digest English–Afrikaans Dictionary*. Cape Town: Reader's Digest Association South Africa.

Hanks, P. (Ed.). 1979. *Collins Dictionary of the English Language*. London: Collins.

Hartmann, R.R.K. and G. James. 1998. *Dictionary of Lexicography*. London/New York: Routledge.

Ilson, R. (Ed.). 1985. *Dictionaries, Lexicography and Language Learning*. Oxford: Pergamon Press.

Jackson, H. 1985. Grammar in the Dictionary. Ilson, R. (Ed.). 1985: 53-59.

Kgasa, M.L.A. (Ed.). 1976. *Thanodi ya Setswana ya Dikole*. Cape Town: Longman Penguin Southern Africa.

Kgasa, M.L.A. and J. Tsonope. 1995. *Thanodi ya Setswana*. Botswana: Longman.

Kriel, T.J. 1985. *New English Northern Sotho Dictionary*. Johannesburg: Educum.

Louwrens, L.J. 1994. *Dictionary of Northern Sotho Grammatical Terms*. Pretoria: Via Afrika.

Martin, W. and G.A.J. Tops (Eds.). 1989[2]. *Van Dale Groot Woordenboek Engels-Nederlands*. Utrecht: Van Dale.

McArthur, T. 1985. *Worlds of Reference*. Cambridge: Cambridge University Press.

Mish, F.C. (Ed.). 1987. *Webster's Ninth New Collegiate Dictionary*. Springfield, Massachusetts: Merriam-Webster.

Prinsloo, D.J. 1994. Lemmatization of Verbs in Northern Sotho. *S.A. Journal of African Languages* 14(2): 93-102.

Procter, P. (Ed.). 1978. *Longman Dictionary of Contemporary English*. Harlow: Longman.

Wiegand, H.E. 1996. Über die Mediostrukturen bei gedruckten Wörterbüchern. Zettersten, A. and V.H. Pedersen (Eds.). 1996: 11-43.

Zettersten, A. and V.H. Pedersen (Eds.). 1996. *Symposium on Lexicography VII*. Tübingen: Max Niemeyer.

Ziervogel, D. and P.C. Mokgokong. 1975 *Groot Noord-Sotho Woordeboek*. Pretoria: J.L. van Schaik.

Abbreviations used in reference to dictionaries

A Learner's Chichewa and English Dictionary (LCE)
Collins Dictionary of the English Language (CED)
Dictionary of Lexicography (DL)
Dictionary of Northern Sotho Grammatical Terms (NGT).
Groot Noord-Sotho Woordeboek (GN)
Lexicon Cilubà–Nederlands (LCN)
Longman Dictionary of Contemporary English (LDOCE)
New English Northern Sotho Dictionary (NEN)
New Student's Dictionary (NSD)
Northern Sotho Terminology and Orthography (NTO)
Reader's Digest English–Afrikaans Dictionary (RD)
South African Student's Dictionary (SSD)
Thanodi ya Setswana (TS)
Thanodi ya Setswana ya Dikole (TSD)
Webster's Ninth New Collegiate Dictionary (W9)

Loanwords in Cilubà*

Ngo Semzara Kabuta, *Vakgroep Afrikaanse Talen en Culturen, University of Ghent, Belgium*

Abstract: The present study examines loanwords in Cilubà from both a phonological and a morphological point of view. Two large categories of loanwords can be distinguished: on the one hand those which are entirely integrated and on the other hand more recent loanwords which retain a large number of their original phonological features. On the phonological level, loanwords (1) introduce new phonemes such as [R] and [g], (2) increase the proportion of low tones, and (3) introduce new combinations of phonemes (e.g. in the sequence C_1C_2V, in which consonants C_1 and C_2 are respectively a nasal and a semivowel, loanwords allow the presence of any consonant). On the morphological level, one notices the appearance not only of forms whose plural is no longer predictable, but also of forms whose plural can be realized in different classes. This phenomenon has important implications in lexicography. As a matter of fact, it is no longer possible to mention in a Lubà dictionary only the singular form and let the reader infer the plural. For nouns the concept of "gender" must therefore be introduced. Gender is defined as a pair of classes whose left and right poles which generally represent the singular and the plural respectively, are chosen in relation to the syntactic concords for the different class affixes (nominal, pronominal, verbal and object prefixes; enclitics), the possessive and the demonstratives, and no longer only in relation to the nominal prefix. Thus, the gender of a noun appears to play a fundamental role in the macrostructure of a noun lemma. Finally, the study of the processes which are intuitively applied by the speakers to integrate foreign words will be a useful source of stimulation for the coinage of neologisms.

Keywords: CLASS, DICTIONARY, LOANWORD, GENDER, LEXICOGRAPHY, LEXICOLOGY, MORPHOLOGY, PHONOLOGY, PREFIX

Abstrait: La présente étude examine les mots d'emprunt en cilubà du double point de vue phonologique et morphologique. On reconnaît deux grandes catégories d'emprunts: d'une part ceux qui sont entièrement intégrés et, d'autre part, ceux qui, plus récents, retiennent un grand nombre de leurs traits phonologiques originels. Sur le plan phonologique, l'emprunt (1) introduit des phonèmes nouveaux tels que [R] et [g], (2) augmente la proportion des tons bas, et (3) introduit de nouvelles combinaisons de phonèmes (par exemple, dans la syllabe de type C_1C_2V, où les consonnes C_1 et C_2 doivent être respectivement une nasale et une semi-voyelle, les emprunts permettent la présence de consonnes quelconques). Sur le plan morphologique, on observe non seulement l'apparition de formes dont le pluriel n'est plus prévisible, mais aussi de formes qui peuvent former leur pluriel dans différentes classes. Ce phénomène a des implications importantes sur le

* An earlier version of this article was read at the First International Conference of the African Association for Lexicography, held at the Rand Afrikaans University, Johannesburg, 1-2 July 1996.

plan lexicographique. En effet, il ne suffira plus désormais de mentionner dans un dictionnaire lubà la seule forme du singulier et de laisser au lecteur le soin d'en deviner la forme du pluriel. On est ainsi amené à développer pour les substantifs la notion de "genre". Celui-ci est défini comme une paire de classes dont les pôles gauche et droit, qui représentent généralement le singulier et le pluriel, sont choisis en fonction de leurs accords syntaxiques pour les différents affixes de classe (préfixes nominal, pronominal, verbal et objet; enclitiques), le possessif et les démonstratifs, et non plus seulement en fonction de la forme du préfixe nominal. Ainsi le genre d'un substantif s'avère être une donnée fondamentale dans la macrostructure d'un lemme substantival. Enfin, l'étude des procédés appliqués intuitivement par les locuteurs pour l'intégration de mots étrangers sera une source d'inspiration utile pour la création de néologismes.

Mots-clefs: CLASSE, DICTIONNAIRE, EMPRUNT, GENRE, LEXICOGRAPHIE, LEXICOLOGIE, MORPHOLOGIE, PHONOLOGIE, PRÉFIXE

Abbreviations[1]

The following abbreviations are used in this article:

": a vowel preceded by this sign is syllabic
$: syllable boundary
#: word boundary
=: exactly the same as adjacent word on the left
+: this sign means that an np is secondary
±: this sign after the monomoraic locative np means that this prefix can precede a stem or a noun
ad: anaphoric distributive
Ar: Arabic
C: consonant
cc: cardinal concord (used in cardinal numbers 1-6)
cl: class
dd1: deictic demonstrative 1 (this, these)
dd2: deictic demonstrative 2 (that, those)
Du: Dutch
Eng: English
F: falling tone
Fr: French
G: glide
gen: gender
Gr: Greek
H: high tone
Kswa: Kiswahili
L: low tone
Lat: Latin
M: middle tone
N: nasal
np: nominal prefix
npq: np used in qualificatives (adjectives, ordinals 1-6 and past participles)
oc: object concord
pe: pronominal enclitics
pl: plural
po: possessive morpheme (à + affix except in cl 1)
pp: pronominal pronoun
`pp: pronominal prefix with L and floating tone
Port: Portuguese
R: rising tone
sc: subject concord
sing: singular
V: vowel

1. Introduction

Cilubà[2] is one of the four national languages of the Congo (formerly Zaïre[3]), the other three being Kiswahili, Lingala and Kikongo. It is in direct contact with French (the official language) as well as with these three languages. It is spoken

in two of the eight provinces: in Western Kàsaayì (capital: Kanàngà) by the Beena-Luluwà and Bakwà-Luntu, and in Eastern Kàsaayì (capital: Mbùjimâyi) by the Balubà proper. However, it extends far beyond these provinces, with many speakers in the other provinces, particularly in Shaba and Kinshasa (Kalonji 1993: 346). There are at least five million active Cilubà speakers (Kalonji 1993: 26).[4]

Studies have been devoted to the phonology, morphology, dialectology and syntax of Cilubà in the past, although most of these need updating (e.g. Gabriel 1921, Burssens 1946b, Stappers 1949, Coupez 1954, Meeussen 1944-59 and 1962, Mutombo 1977, Kabuta 1995, 1996). However, no research was done on lexicology, while lexicography was left to the missionaries (e.g. Morrison 1906, 1939, De Clercq 1914, 1936, Gabriel 1922 and 1925, De Clercq and Willems 1960, and Willems 1986). It is only recently that some linguists have compiled word lists and lexicons (e.g. Yukawa 1992, Kadima et al. 1995, and especially ACCT 1983 which e.g. contains hundreds of neologisms coined among others by borrowing from the field of economic and social activities, as well as Bunduki 1975, a terminology of linguistics). A theoretical work giving guidelines for the compilation of a modern dictionary was also published a few years ago (Kalonji 1993). The present article is part of a preliminary study on some important issues to be taken into account in any modern monolingual or bilingual Cilubà dictionary project. It describes on the one hand the strategies used to nativize words, and on the other hand the changes which borrowing introduces into the phonology and the morphology.

Sociolinguistically, French has always enjoyed a prestigious position in the Congo, since it was the language of the colonizer. Even after independence (1960), it remained the obligatory passage to social promotion. In 1962 when it became the official language, it was constitutionally given a predominant role in different spheres of activities, namely in education and administration. Consequently, many Congolese are in a situation of diglossia, which explains the importance of borrowing from French. Before the colonization, contacts with Portugal started as early as 1482, when the first Portuguese, led by Diego Cão, arrived in the kingdom of the Kongo which spread along the Atlantic Ocean. The arrival of the Portuguese was followed by at least two centuries of intense political and commercial activity. In the second half of the 19th century, the country of the Luluwà was visited by Cokwe hunters and traders from Angola.[5] During this period, new products from Europe and the Americas were introduced by the traders, and these products generally came with their foreign names. There were also commercial exchanges with East Africa, which resulted in the introduction of new products and their names, generally from Arabic. As a rule, the source languages are either coastal, trade or administrative languages. Not surprisingly, the main source languages for Cilubà are Portuguese, Kiswahili and especially French.

Loanwords will be understood here as "those words which were not in the vocabulary at one period and are in it at a subsequent one, without having been

made up from the existing lexical stock of the language or invented as entirely new creations, as for example, certain names for products are (kodak, etc.)" (Robins 1975: 324). Words sometimes travel a long way from one language to another, passing through other languages. For example, Cilubà has a few words from Arabic, although it was never exposed to the direct influence of this language. Other languages have indeed served as "carriers", e.g. Kiswahili in the case of Arab words. The aim of this article is not to discuss this issue and trace the history of the loanwords, although such a study would certainly be of great interest for the cultural history of the Balubà. The source languages mentioned in the examples are therefore just meant to show the foreign origin of the words, and not necessarily their original forms. Furthermore, there is a fair amount of loanwords in the field of Christian religion which have different forms according to whether they were introduced by Protestant or Catholic missionaries. As a rule, "Protestant" loans are closer to Lubà phonology than "Catholic" ones and will therefore preferably be referred to.

To study borrowing implies answering at least the following questions: What is borrowed and how does it happen? Who borrows? Why and when does one borrow? The answers to the first two questions are of a linguistic nature, whereas the answers to the others are sociolinguistic. The data at our disposal allow us to focus only on the linguistic questions.

Analyzing current conversations with different social groups as well as written material,[6] we noticed that besides inter- or intrasentencial code-switching, loanwords are used extensively. A list of about 600 loanwords was drawn up. This list is insignificant compared to the whole Lubà lexicon, but, interestingly enough, it belongs to everyday vocabulary[7] which generally does not exceed 3 000 words (the COBUILD English Dictionary 1995, e.g. uses a vocabulary of 2 500 words to define all the lemmatized words). As is the case with other languages (cf. e.g. Bader and Mahadin 1996: 39), most of the words (over 90%) are nouns.[8] The remainder are verbs, adjectives (mostly used with a connective pronoun) and adverbs. There are a few phrases which are borrowed as one word.

All the words have been spelt uniformly, irrespective of their spelling in the source material. The following general conventions were used, some of which are explicated in the paragraph on phonology:

/i/ + /V/ (V≠i) > /yV/
/u/ + /V/ (V≠u) > /wV/[9]
/n/ + /i/ > /nyi/[10]
Low tone: ˋ
Falling tone: ˆ
Rising tone: ˇ
High tone: not marked
N always bears a diacritic when syllabic
A long vowel is represented as VV with the restrictions mentioned in 2.1 1°.

2. Phonology

2.1 There are five vowels (/i/, /u/, /e/, /o/, /a/) which can be combined with vowel quantity and tone to yield ten forms for each vowel. For instance, the different forms for the vowel /i/ are as follows: /í/, /ì/, /í:/, /ì:/, /ǐ:/, /î:/, /i̱:/, /ì̱:/, /ǐ̱:/, /î̱:/.[11] Complex tones and nasality are always associated with vowel quantity. Furthermore, both nasality and vowel quantity are only possible before a consonant inside a word, which means that they are excluded in word-final position. Exceptions are a few conjunctions, such as *àn* /à̱:/ or *èn* /è̱:/ (yes) and *tǒ̱* (no) and the word *mbû* (or *mbuwù* ocean). The most used vowels are the low and high[12] vowels. In word-final position, /e/ will alternate with /a/, /o/ with /u/ (examples (1)(a)), but not the opposite (examples (1)(b)). In certain cases /i/ will freely alternate with /e/ and with /o/ (examples (1)(c)). In the pronunciation of many speakers, /e/ never occurs in this position. All these cases stress the preference of the language for low and high vowels, especially in word-final position.

(1) (a) mupânde=mupânda *torn* (active past participle)
dilòòlò=dilòòlù *evening*
(b) kwebeja≠*kwebeje *to ask*
mupanda≠*mupande *torn* (passive past participle)
tulù≠*tulò *sleep*
(c) kumwambilayè=kumwambilayì *he told him*
mwoyo=mwoyi *heart*
byôbyo=byôbi *them*

The following rules are used for the representation of vowels and tones:[13]

1° (a) V > [V:] / —— NC
(b) V > [V:] / CG —— $
(c) V > [V:] / #G ——

Because of these rules, the vowels in bold in the examples below are written only once although they are bimoraic:

(2) kunanga *to love*, kukwàta *to catch*, webè *your(s)*, yà *of*

2° H's are not represented, being the most frequent.

3° The M, which is responsible for downdrift, is not distinctive. Being phonologically predictable, no special sign is used to represent it:

R > M: / H ——
L > M / H ——

(3) tatwěbè *your father*, patwǎyi *when we went* are respectively pronounced:
[‾ —— _] and [– —— –] (or [‾ / –]),
manàyi *games* is pronounced:
[‾ – –] (or [‾ _ –])

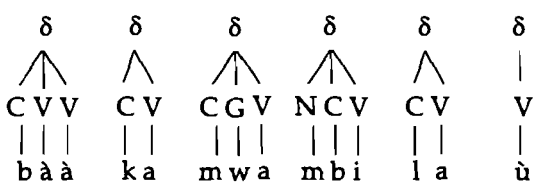
2.2 There are 17 consonants: /m/, /n/, /ɲ/, /ŋ/, /b/, /v/, /l/(/d/), /z/, /ʒ/, /p/, /f/, /t/, /s/, /ʃ/,/k/, /ɸ/, /tʃ/. Some of these are conventionally represented as follows:

/ɲ/	: ny	/tʃ/	: c (or tsh)
/ŋ/	: ng	/ʒ/	: j
/ʃ/	: sh	/ɸ/	: p

/d/ is in complementary distribution with /l/ after /n/ and before /i/.

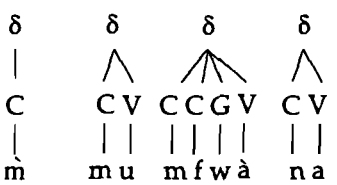
2.3 The syllable structures are CV, V, NCV and CGV.[14] There are variant forms as illustrated below. Example (4)(b) illustrates that the predicative morpheme ǹ *it is* (and its combinatory variants) is syllabic. The same is true for the morpheme ń-/ǹ- (sc first person sing) as shown in example (4)(c). A C-type syllable is often heard in sentence final position also, where the vowel is probably aspirated.

(4) (a)



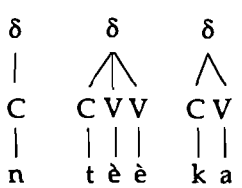
they told them (those things) to him

(b)



he resembles me

(c)



(I) putting

2.4 Excluded phoneme combinations are:

(a) CG: nyw, cw, zw, jw, shw.
(b) CV: si, zi, ti, ni;[15] ve, va, fe, fa;[16] vo, fo; li.[17]
(c) CCV: it is the moraic nature of N and G which accounts for the tolerance of NC and CG, as shown in note 15.
(d) NVl: only [ŋV] is permitted, otherwise l undergoes nasal assimilation and becomes n.

(5)	/lú-mòn-ílú/	> lumwènu	*mirror*
	/kú-tùm-íl-á/	> kutùmina	*to send to*
	/kú-sùm-íl-íl-á/	> kusùminyina	*to persist*
	but:		
	/kú-kàng-íl-á/	> kukàngila	*to enclose, to shut* + preposition

2.5 There are two types of loanwords:

2.5.1 Loanwords which are completely integrated into Lubà phonology, although they will display features which are rather rare, such as a low np or no prefix at all (as in some Lubà kinship terms). At this stage, the phonetic structure of the language is not disturbed by the introduction of new sounds, the following general principles being applied:[18]

1° Vowel epenthesis, which results in syllabification of clusters. Particularly, if the borrowed noun begins with a cluster with initial [b] or [k], [u] and [a] respectively are appended, which results in CV-type syllables corresponding to classes 12 and 14 nps. Very often, when the foreign word ends with a consonant, Cilubà appends an identical vowel to the preceding consonant, unless the phonetic features of this consonant exert an influence:

$C_1 \begin{Bmatrix} C_2 \\ \# \end{Bmatrix}$ > $C_1 V \begin{Bmatrix} C_2 \\ \# \end{Bmatrix}$

(6)	classe	> kàlaasà	*classroom, school*
	clerc	> kàleelèkà	*white-collar worker*
	cravate	> kàlàvwandà	*tie*
	bloc	> bùlokò	*prison*

As a rule, the quality of the appended vowel is determined by the adjacent phonemic features.[19] In most cases, however, a low vowel will be inserted, as its frequency in the language is the highest among the vowels.[20]

(7) (a)	bath (Eng)	> (m)baafù	*bath, basin*
	pas op (Du)	> kusopwesha[21]	*to warn*

	Gabriel	> Ngaabùdyèlà	*Gabriel*
	corned beef (Eng)	> kòlònèbefù/kòlènèbefù	*corned beef*
	diable	> dyabùlù	*devil*
	bifteck	> bìfùtekà	*beefsteak*
	soupe	> nsupù	*soup*
(b)	juge	> nzujì	*judge*
	chemise	> nsùmijì (-sà)	*shirt*
	glace	> dìkàlaashì	*glass*
	belge	> beelèjì	*Belgian*
	chave (Port)	> nsapì	*key*

Quite often, when a word ends in /e/, it will freely alternate with /a/, as happens in normal Lubà words:

(8)	fête	> fetè/fetà	*celebration, party*
	cassette	> kàsetè/kàsetà, kàset	*tape cassette*

Instead of /u/, a glide may be appended. In the second and third examples, /u/ is inserted after /v/ and /f/ because the sequences /va/ and /fa/ are not permitted:

(9)	franc	> mfùlangà or mfwàlanga	*money*
	cravate	> kàlàvwandà	*tie*
	tofali (Kswa)	> dìtàfwadì	*brick*

2° Epenthesis of an np (mostly class 1 nasal np or class 5 np):

(10)	boy (Eng)	> dìbooyì	*servant*
	carro (Port)	> dìkalù	*bicycle*
	bath (Eng)	> mbaafù	*bath, basin*
	glass (Eng)	> dìkàlaashì	*glass*
	mpira (Kswa)	> mùpilà	*pullover*
	baraza (Kswa)	> dìbàlaasà	*verandah*
	sapato (Port)	> cìsàbaatà	*shoe*
	sentry (Eng)	> nsentedi	*sentry*
	limão (Port)	> didìmà	*lemon*
	kopo (Port)	> dikopo	*cup*

3° Whenever there is a formal resemblance between the first syllable (or article plus first syllable) of a foreign word and a Lubà np, the former is adapted to match the shape of a Lubà np (cf. Chart 1); e.g. [ly, lo, lɔ] > /lu/; [me] > /mi/; [li] > /di/); [to, tɔ] > /tu/; [b] > /bu/; [by] > /bi/; [k] > /ka/ (examples (11)(a)). When this is not possible, a nasal prefix is used (examples (11)(b)). In some cases, a foreign initial syllable is felt to be a plural prefix and is subsequently made to alternate with a Lubà singular prefix (examples (11)(c))[22].

(11)	(a)	bloc	> bùlokò 14	*prison*
		classe	> kàlaasà 12	*classroom*
		courant d'eau	> kàlàndê 12	*trench*
		machine	> màshinyì 6	*car*
		cassette	> kàsetà 12	*tape cassette*
		bus	> bisà 8	*bus*
		coeur-de-boeuf	> kàlàbefù 12	*kind of fruit*
		l'hôpital	> lùpìtaadì 11	*hospital*
		l'histoire	> dìstwâr 5	*story*
		lunette	> lùneetà 11	*spectacles*
		caixete (Port)	> kashèètà 12	*box*[23]
	(b)	pato (Port)	> mpaatu 1	*duck*
		soupe	> nsupù 1	*soup*
		canezou	> nkanzu 1	*dress*
		sukari (Kswa)	> nsùkaadì 1	*sugar*
		juge	> nzujì 1	*judge*
		pão (Port)	> mpaù 1	*bread*
		pataco (Port)	> mpatà 1	*5-franc coin or note* (in colonial times)
	(c)	tomate	> tùmatà 13 (cf. kàmatà 12)	*tomatoes*
		mes habits	> mìzàbì 4 (cf. mùzàbì 3)	*cassocks*
		minute	> mìnutà 4 (cf. mùnutà 3)	*minutes*
		million	> mìlyô 4 (cf. mùlyô 3)	*millions*

4° Extrasyllabic truncation:[24]

(12)	épingle	> mpengèlà	*safety pin*
	appel	> mpeelù	*call*
	américain	> màlèkaanyì	*kind of cloth*
	essuie-mains	> sùmê	*towel*
	indépendance	> dìpàndà	*independence*

5° Final or penultimate nasal vowel > velar + vowel:

(13)	franc	> mfwàlangà	*money*
	sabão (Port)	> nsàbangà	*soap*
	botão (Port)	> mbòtangà	*button*

6° /g/ > /ng/ (sometimes /k/):

(14)	grec	> kàlekà	*praying-place* (for adepts of Bupoostòòlò, a syncretic religion)
	gâteau	> kààtô	*cake*
	gare	> ngâlà	*station*
	garfo (Port)	> ngàlafù	*fork*
	grâce	> ngaasà	*mercy*
	Gabriel	> Ngaabùdyèlà	*Gabriel*
	gold (Eng)	> ngôlù	*gold*

7° In a few cases, a voiceless stop will become voiced:

(15)	guitare	> cìdâlà	*guitar*
	kabati Kswa	> kabàdì	*cupboard*
	tampon	> cìtambì/-pì	*seal*

There is one known case in which a voiced stop alternates with a voiceless:

(16)	salade	> (màfutà àà) nsaalàtà	*(oil for) salad*

2.5.2 Loanwords which retain some of their original phonological features and are thus only partially nativized, as in the following examples, all from French. All these words are relatively recent, and it is unlikely that they will naturally undergo further nativization. Rather, many of the words which were fully nativized (*mfwàlànsa*, *ngâlà*, and so on), tend to be pronounced as in French. The older pronunciation, it seems, becomes associated with poor schooling. The different changes enumerated below are certainly the result of a greater familiarity with French if not through education, at least through the media. In these words the following phenomena are observed:

1° Nasal vowels and complex (namely falling) tones appear in word-final position:

(17)	l'histoire	> dìstwâr	*story*
	famille	> fàmî	*family*
	contrat	> kòntrâ	*contract*
	pardon	> pàrdõ	*sorry*

2° Absence of np. Such words belong to gender 1/4 (see 3.2 and 3.3 2°):

(18)	secret	> sèèkèlê	*secret*
	congé	> kònjê	*off day, holiday*
	parti	> pàrtî	*(political) party*

pick-up (Eng)	> pìkepà	*delivery van*
client	> kìdìyâ	*client, customer*

3° Words from French display the following general tonal pattern L ... H, L ... F or L ... FL where H or F corresponds to the accented syllable in French. Such patterns increase the number of L nps, as well as the number of stems with L's:

(19)			
	allumette	> àlàmeetà	*match*
	acide	> àsîdà	*acid*
	politique	> pòlìtikà	*politics*
	fenêtre	> fineetèlà	*window*
	sida	> sìdâ	*aids*

4° All sorts of clusters and CV sequences are tolerated, in violation of the restrictions mentioned in 2.4:

(20)				
	C+r:	mfrangà	*money*	< franc
		muprofetà	*prophet*	< prophète
	s+C:	mùpoostòòlò	*apostle*	< Gr apostolos
		eskè	*question phrase*	< est-ce que
		dìspànsêlà	*dispensary*[25]	< dispensaire
	y+w:	bùywâlà	*kettle*	< bouilloir
	C+s:	tààksî	*taxi*	< taxi
		tèleksè	*telex*	< télex
	l+i:	pòlìtikà	*politics*	< politique
	d+a:	àsîdà	*acid*	< acide
		dààkôr	*all right*	< d'accord
		dàyêr	*anyway*	< d'ailleurs
	Nvl:	kòntènêlà	*container*	< Eng container

5° French phonemes such as /œ/, /y/, /g/ and especially /r/ are tolerated. One notices even the phenomenon of hypercorrection, by which e.g. [r] is pronounced instead of [l].

(21) Philomène > Phiromène

3. Morphology

3.1 The Lubà noun has one of the prefixes listed in column 2 of Chart 1. This prefix has a H and is monomoraic. There are very few cases of L nps[26]. Grouping nouns by genders rather than by classes will best show us the difference between pure Lubà or fully nativized words and partially nativized words. A gender is defined as a morphosyntactic pair of classes whose members, different from ø, generally represent the singular and plural forms respectively[27]. The

involved affixes are the np, npq, `pp, cc, sc, oc, pe, po, ad, dd1 and dd2. The np, which can have variants (cf. e.g. gender 1/4 in Chart 1, in which class 1 np can be mu-, N or Ø) and can even be regarded as a word in classes 16, 17 and 18 and therefore written separately, is not taken into account for the definition of gender. In the pair Ø/6, the right member has a collective rather than a plural meaning; in the other pairs containing a Ø, the opposition singular/plural is irrelevant. According to the system generally used in Bantu languages, the number of genders for Cilubà appears to be 21:

(22)			
	1/2	mwâna, bâna	*child(ren)*
	1/4	ntambwa, ntambwa	*lion(s)*[28]
	3/4	mucì, micì	*tree(s)*
	5/6	dijiba, majiba	*lake(s)*
	5/Ø	dipìta	*passing*
		dyàkabì	*bad luck*
	Ø/6	mâyi	*water*
	7/8	cibelu, bibelu	*thigh(s)*
	7/0	cikongo	*Kikongo; like the Bakongo*
	8/8	bidyàa, bidyà	*porridge(s)*
	8/Ø	bikolè	*hard, very*
	11/4	lulengu, ndengu	*poison(s)*
	11/Ø	lùkàsà	*quickly*
	12/13	kantu, tuntu	*small thing(s)*
	12/Ø	kakesè	*a little*
	Ø/13	tuminu	*nasal mucus*
	14/6	bulaba, malaba	*soil(s)*
	14/Ø	buntu	*humanness, humanity*
	15/Ø	kwakula	*to speak*
	16/Ø	pa mèèsà	*on the table*
	17/Ø	ku mèèsà	*at the table*
	18/Ø	mu nzùbu	*in the house*

These genders are made up of the 16 class numbers contained in Chart 1, plus Ø to express the absence of a class. In this chart, independent nominals (nouns), take one or two of the nps listed in column 2, whereas dependent nominals (qualificatives) only take the canonical variant (labelled npq) of the corresponding np:

(23)					
	mwâna	(np: mu-)	mwîmpè	(npq: mu-)	*a nice child*
	mùkooko	(np: mù-)	mwîmpè	(npq: mu-)	*a nice sheep*[29]
	nzùbu	(np: N-)	mwîmpè	(npq: mu-)	*a nice house*
	sààkooshì	(np: Ø-)	mwîmpè	(npq: mu-)	*a nice bag*

Qualificatives are adjectives, past participles and ordinals from 1 to 6:

(24)

mwâna mwîmpè	*a nice child*
mwâna mulààle	*a sleeping child* (cf. kulààla *to sleep*)
dìkalù diitânu	*fifth bicycle*

The prefixes used in (22) are primary nps. In some classes (2, 6, 7, 8, 12, 13 and 14), there exists a second set of nps which are phonologically distinct from the primary nps. They precede a full noun, i.e. they are used before another np, which can be Ø in loanwords. A secondary np is always bimoraic (CVV) and bears a high H. It is indicated by a + sign after the conventional class number or morpheme. The locative prefix can be secondary, but it remains monomoraic (CV). In this case, it is written separately and can be regarded as a word rather than a morpheme:

(25)

kàkalù 12/kaadìkalù 12+	*a small bicycle*
tunkanzu (tuu+n+kanzu) 13+	*little dresses*
pa mucì mucyàmàkàne	*on the cross*
kù baabèndè	*abroad*
mu eu nzùbu	*in this house*

In column 4, which lists the pe's (used in subject relatives, possessives and connectives), the following rule is applied: H# > L/{F,L}$ —— #`p̀p:

(26)

bâna # `bănàyi	> bânà bănàyi	*the children who have played*[30]
bâna # `bèèbè	> bânà bèèbè	*your children*
matùnga # `àà luuyà	> matùngà àà luuyà	*warm countries*

The tone of the pe's in column 8 is in contrast with the adjacent tone:

(27)

kumufùndayè	*he accused him*
pààmufùndàye	*when he accuses him*

The examples below are translated literally in order to illustrate the use of chart 1. The class affixes, which are sometimes modified by some morphonological rule, are given in bold type:

(28) class 1

Mwâna	**mwîmpè**	**wa**	Ilunga	**wălu.**	**Ùmupèèshè**	cyàlòmbà**ye.**	
Child	*nice*	*of*	*Ilunga*	*has come.*	*Give him*	*what he'll ask.*	

class 3

Mucì	**mwîmpè**	**wà**	Ilunga	**wăcìbuku.**	**Nêtùwùôshèpùùmàu.**	
Tree	*nice*	*of*	*Ilunga*	*is broken.*	*We'll burn it*	*when it dries up.*

class 5
Dìkalù dîmpè dyà Ilunga **dyănyangùkù.** Nêndìlongòlòlè pààfikàdi.
Bicycle nice of Ilunga is broken. I'll repair it when it arrives.
class 7
Cisanjî cîmpè cyà Ilunga **cyănyangùkù.** Nêncìlongòlòlè pààfikàci.
Radio nice of Ilunga is out of order. I'll repair it when it arrives.

3.2 Firstly, the Lubà infinitive has class 15 np and ends in -a, exceptions being a few defective stems:

(29) -di *to be,* -tu *to be often* or *generally* (these verbs do not have class 15 nps)
kwanji *auxiliary verb meaning "x first"* (diachronically: kwanza)

Only very few foreign verbs have been fully adapted, such as:

(30)	bénir	> kubèènesha	*to bless*
	baptiser	> kubàtiiza	*to baptize*
	peindre	> kupenta	*to paint*
	pas op[31] (Du.)	> kusopwesha	*to warn*

Other verbs retain their original infinitive form in all tenses. Because they are kept phonologically intact (though they sometimes can be combined with ordinary verbal morphemes), they should perhaps rather be regarded as cases of code-switching, particularly as this principle is applied to any verb:

(31)	proposer	> netùbàpròpòzê	*we will propose them*
	concevoir	> kukònsèvwâr	*to conceive*
	définir	> kudèfînîr	*to define*
	se débrouiller	> kudìdèbrùyê	*to manage, to get on*
	remarquer	> ngâkarèmàrkê	*I noticed*
	investir	> ǹcinyì cyàkaènvèstîryi	*what did he invest?*
	comprendre	> kabààkukòmprandrè to	*they will not understand*

Secondly, new genders (or new combinations of classes) are created, as can be seen in Chart 2. The total number of genders is extended from 21 to 28, not counting the variant forms indicated by a and b. In this chart, the members of each gender have been illustrated with singular and plural examples, although, as has already been said, the opposition singular/plural is not relevant to all genders. It is obvious that an np (column 2) inside a gender can display various phonological shapes (shown with the letters a and b), whereas the class pair or gender remains constant (column 1) no matter the np variants[32]. The symbol Ø in column 1 means the noun is monoclass; in columns 2 or 3, it means that there is no np or that the apparent np is not relevant (cf. 5/4 or 14/4 in Chart 2). The following general observations can be made:

1° Gender 1/2 contains only human beings. Human beings belonging to classes 1, 7 and 12 are often found in subgender 1/2a, in which a noun is preceded by a secondary prefix:

(32)	mungàngà, baamingàngà	*doctor(s)*
	mfùmù, bamfùmù	*chief(s), king(s)*
	cilembi, baacilembi	*hunter(s)*
	kangìmbà, baatungìmbà	*singer(s)*
	mìnistrè, baamìnistrè	*minister(s)*
	pêrè, baapêrè	*Catholic clergyman(-men)*
	mùmpêlà, baamùmpêlà	*Catholic clergyman(-men)*

This gender does not only contain kinship terms as traditional grammars claim. It contains two loanwords[33] in which nps N- and baa- alternate:

(33)	virgo (Lat)	> mvirgò, baavirgò	*female virgin(s)*
	sacerdoce (Lat)	> nsàserdòsè, baasàserdòsè	*priest(s)*

2° Gender 1/4 normally contains only nouns with nps N- for both the singular and the plural. All foreign words, which do not naturally have a class prefix, or whose first syllable cannot be interpreted as such, are placed here.

(34)	tv	> tèèvê, tèèvê	*television set(s)*
	amende	> àmândà, àmândà	*fine(s)*

3° Genders 4/4, 6/6 and 8/8 are characteristic of nouns which use the same affixes for both the singular and the plural. All of these, except *bidyà* (porridge), are loanwords:

(35)	misa (Lat)	> misà 4/4	*holy mass(es)*
	mitraillette	>mìtràyetà 4/4	*riot gun(s)*
	machine	> màshinyì 6/6	*car(s)*
	budget	> bìdyê 8/8	*budget(s)*
	biberon	> bìbèrôn 8/8	*baby bottle(s)*

4° In gender x/6, a loanword from any gender except 1/2 may keep its np for the singular (often zero in loanwords) and append np maa+ for the plural. The singular nps belong most of the time to classes 1, 5, 6, 11, 12 or 14. One word was found belonging to class 3 np. Since the singular can be any class, it is indicated by x in the gender formula:

(36)	valise	> vààlîzà x(=1)/6	*suitcase(s)*
	radiateur	> rààdyàtêr x(=1)/6	*radiator(s)*
	camion	> kààmìnyô x(=12)/6	*lorry (lorries)*

moteur	> mwòtêr x(=3)/6	*motor*[34]
mission	> mìisyô x(=4)/6	*mission(s)*
disque	> di(i)skè x(=5)/6	*record(s)*
bombe	> bwômbà x(=14)/6	*bomb(s)*
loisir	> lwàzîr x(=11)/6	*recreation*

5° The prefixes in genders 12/4 and 12/13b are bimoraic and bear an L. They are the only genders where long and L primary nps are found:

(37)	camion	> kààmìnyô 12/13	*lorry (lorries)*
	quartier	> kààrcyê 13/4 or 12/13	*town area(s)*

6° In gender 14/4, class 14 np is associated with class 4 np in such a way that the first syllable is regarded as an np in the singular, but not in the plural:

(38)	bwômbà bwătaayìkì ku Tel Aviv	*a bomb exploded in Tel Aviv*
	bwômbà yătaayìkì	*bombs exploded*

7° Some nouns, most of which are loanwords, are found to belong to different genders:

(39)	mungàngà 1/2a or 1/4	*doctor(s)*
	kangìmbà 1/2a or 12/13a	*singer(s)*
	cilembi 1/2a or 7/8	*hunter(s)*
	kààrcê 12/4, 12/13b or x/6	*town area(s)*
	màshinyì 6/6 or x/6	*car(s)*
	teevê 1/4b or x/6 (cf. tv Fr)	*television set(s)*
	bwômbà 14/4 or x/6	*bomb(s)*

8° Since some genders (7/8, 11/4, 12/13 and 14/Ø) are possible with almost any noun by nominal derivation, only a selection of nouns (based on their frequency) belonging to them will be included in a basic dictionary. It goes without saying that among locatives, only locative nouns will be included (such as *pambèlu* (outside), and not *pa mèèsà* (on the table) in which the locative is used prepositionally). Of course, the three locative nps will represent three different entries, as they can have a prepositional function.

Because of the proliferation of genders due to loanwords, a chart like Chart 2 is indispensable in any modern Lubà dictionary. As a synopsis of all the concord possibilities, it allows the lexicographer to limit the metalinguistic information in the microstructure to a minimum. For example, the metalinguistic information provided by kangìmbà 1/2a or 12/13a is the following: the syntactic concords for this noun which designates a human being, occur in class 1 for the singular despite its np which belongs to class 12; its plural is in class 2

with the secondary np baa+, which is added either to the singular or to the plural noun:

(40) kangimbà mupyamùpyà uùvwù mumòna *the new singer you saw*
baakangìmbà (or baatungìmbà) bapyabàpyà baùvwà mumòna
the new singers you saw

Gender 12/13a, which is also possible for *kangàmbà* means that this word can also behave like any word of class 12, irrespective of its human content, which would require the use of class 1 npq, `pp, cc, sc, oc, pe, po, ad, dd1 and dd2 as in example (40). Thus:

(41) kangìmbà kapyakàpyà kàdi kìmba bîmpè *the singer sings well*
tungìmbà tupyatùpyà tùdi twìmba bîmpè *the new singers sing well*

The genders of the loanword *kààrcê* inform the reader, e.g. that one can say:

(42) kààrcê mipyamìpyà, maakààrcê mapyamâpyà or tùùrcê tupyatùpyà
new town areas

It is obvious that accurate gender indication provides a lot of useful information in a very condensed way. Frequency counts based on a much larger corpus will allow us to know which genders are used most when a noun belongs to more than one gender.

4. Conclusion

Words are borrowed not only because they come with new concepts, but also because they accompany new habits. In addition, shorter words are adopted more easily. Borrowing does not necessarily mean that the borrowing language lacks equivalent words or fails to coin them. Sociolinguistic reasons, such as prestige often intervene to favour foreign words. For instance, the French words for the numbers or for the months are preferred, although equivalents do exist in Cilubà. Words for technical objects or the metalanguage for specialized disciplines such as technology, linguistics, philosophy, economy, politics, etc. are most often borrowed from French. The case of Cilubà also illustrates that languages need not be in direct contact for words to circulate among them.

Phonologically, the pronunciation practices of the Balubà are undergoing changes due to prolonged exposure to French. As loanwords are being integrated into Cilubà, new phonemes ([R], [g], [œ]) and new combinations of phonemes are being incorporated.

The new Lubà morphology is characterized by the appearance of new genders. This change will influence the way metalinguistic information is presented in a dictionary. While with genuine Lubà nouns it was sufficient to men-

tion the singular form of a noun, the plural being automatically deduced, with loanwords it becomes necessary to mention the gender, i.e. the classes in which both the singular and plural forms concord, as this is no longer easily predictable. Furthermore, the following general tendencies are noticeable:

— fairly general use of classes 6 or 4 to mark the plural of inanimate objects, irrespective of the singular prefix;
— appearance of bimoraic primary nps sometimes with Ls;
— use of an np in the singular, but not in the plural; and
— extended use of the same np for both the singular and the plural.

There are often different forms of loanwords, assimilated and unassimilated, often depending on the speaker's attitude or background (e.g. *mfwàlangà/mfùlangà/mfrangà* money; *ngâlà/gârè, kârè* station).[35]

No attempt has been made in this article to explain the existence of a series of words related to food, for which one might expect a foreign origin. Most of them are words for New World crops which were introduced in Central Africa by the Portuguese since the 15th century, such as *cyômbe* (cassava), *mwenga* (sugar-cane), *dyamvwa* or *ditalà* (maize), *cìlùngà* (sweet potato), *kambelà* (peanut), *cikàkà* (pineapple), *ndùngù* or *kacìpì* (bird chilli). In earlier centuries some other crops reached Central Africa across the Sahara or the Indian Ocean from the Middle East or Southeast Asia, such as *cimenà* (yam), *lukùnda* (bean), *lunyimù* (pea), *ditàbàlà* (taro) and *cibòta* (banana).[36] Both phonologically and morphologically, these words are perfect Lubà words. One can hypothesize that over a few centuries the foreign words (whose sources remain unknown) were completely assimilated or that either new names were coined for the new products, or that some transfer of meaning took place from similar original crops to new ones. Proto-Bantu reconstructions have been proposed for *banana, sugarcane, peanut*[37] and *maize*, but except for *dikonde* (big banana), the Lubà forms are not related to any of the reconstructions.

A good understanding of the structure of loanwords will facilitate the task of coining neologisms through borrowing.

Chart 1: Affixes and Demonstratives

1	2	3	4	5	6	7	8	9	10	11	12
cl	np	npq	`pp	sc	cc	oc	pe	po	ad	dd1	dd2
1	mu-/N-, ø-	mu-	u-	ù-/à-	u-	-mu-	-ye/yè	-èndè	au	eu	wàwa
2	ba-/bà-, baa+	ba-	`bà-	bà-	=	=	-bu/bù	-àbù	abu	aba	bààba
3	mu-/mù-	mu-	`ù-	ù-	=	"=	-u/ù	-àù	au	eu	wàwa
4	mi-/mì-, n-, ø-	mi-	`ì-	ì-	=	"=	-yi/yì	-ài	ai	ei	yàya
5	di-/dì-	di-	`dì-	dì-	=	=	-di/dì	-àdì	adi	edi	dyàdya
6	ma-/mà-, maa+	ma-	`à-	à-	=	"=	-u/ù	-àù	au	aa	ààa
7	ci-/cì-, cii+	ci-	`cì-	cì-	=	=	-ci/cì	-àcì	aci	eci	cyàcya
8	bi-/bì-, bii+	bi-	`bì-	bì-	=	=	-bi/bì	-àbì	abi	ebi	byàbya
11	lu-/lù-	lu-	`lù-	lù-	=	=	-lu/lù	-àlù	alu	elu	lwàlwa

12	ka-/kà-, kaa+	ka-	`kà-	kà-	=	=	-ku/kù	-àkù	aku	aka	kààka
13	tu-/tù-, tuu+	tu-	`tù-	tù-	=	=	-tu/tù	-àtù	atu	etu	twàtwa
14	bu-/bù-, buu+	bu-	`bù-	bù	=	=	-bu/bù	-àbù	abu	ebu	bwàbwa
15	ku-	ku-	`kù-	kù-	=	=	-ku/kù	-àkù	aku	eku	kwàka
16	pa±/pà±	pa-	`pà-	pà-	=	=	-pu/pù	-àpù	apu	apa	pààpa
17	ku±/kù±	ku-	`kù-	kù-	=	=	-ku/kù	-àkù	aku	eku	kwàka
18	mu±/mù±	mu-	`mù-	mù-	=	=	-mu/mù	-àmù	amu	emu	mwàmwa

Chart 2: Genders

gen	np	np	sing	pl	translation	content
1/2	mu-	ba-	muntu	bantu	*man (men)*	Humans
1/2a	Ø	baa+	taatù	baataatù	*father(s)*	Humans: Kinship terms
1/2b	N-	baa-	nsàserdòsè	baasàserdòsè	*priest(s)*	Loanwords
1/4	mu-/`	mi-/`	mungàngà	mingàngà	*doctor(s)*	
1/4a	N-	N-	nnyuunyi	nnyuunyi	*bird(s)*	
1/4b	Ø	Ø	teevê	teevê	*television set(s)*	Loanwords
3/4	mu-/`	mi-/`	mucì	micì	*tree(s)*	
4/4	mi-/`	mi-/`	misà mìtràyetà	misà mìtràyetà	*holy mass(es)* *riot gun(s)*	Loanwords
5/6	di-/`	ma-/`	dibòku	mabòku	*arm(s)*	
5/Ø	di-	Ø	dimòna; dyàkabì		*seeing;* *misfortune*	Gerunds; connective words
5/4	di(i)-/`	Ø	diiskè	diiskè	*record(s)*	Loanwords
6/6	ma-/`	ma-/`	màshinyì	màshinyì	*car(s)*	Loanwords
Ø/6	Ø	ma-		mâyi	*water*	Collectives
x/6	x-	maa+	màshinyì	maamàshinyì	*car(s)*	Loanwords
7/8	ci-/`	bi-/`	cintu	bintu	*thing(s)*	Augmentatives
7/8a	cii+	bii+	ciidìkalù	biimàkalù	*big ugly bike(s)*	Augmentatives
7/Ø	ci-/`	Ø	cilubà citòòke		*the Lùba language* *like the Whites*	Languages; customs
8/8	bi-/`	bi-/`	bidyà byêla; bìdyê	bidyà byêla; bìdyê	*porridge(s)* *beer(s); budget(s)*	Loanwords
8/Ø	bi-	Ø	bikolè		*hard; very*	Adverbs
11/4	lu-	N-	lupènzu	mpènzu	*cockroach(es)*	

11/6	lù-	mà-	lùneetà	màneetà	*pair(s) of spectacles*	Loanwords
11/ Ø	lu-/ˋ	Ø	lubilu, lùkàsà		*quickly*	Adverbs
12/13	ka-/ˋ	tu-/ˋ	kantu	tuntu	*small thing(s)*	Diminutives
12/13a	kaa+	tuu+	kaacilamba	tuubilamba	*small bridge(s)*	Diminutives
12/13b	kàà-	tùù-	kààmìnyô	tùùmìnyô	*lorry (lorries)*	Loanwords
12/4	kàà-	Ø	kààrcê	kààrcê	*town area*	Loanwords
12/Ø	ka-/ˋ	Ø	kàbìdì		*again*	Adverbs
Ø/13	Ø	tu-		tuminu	*nasal muscus*	Collectives
14/4	bu-	Ø	bwômbà	bwômbà	*bomb(s)*	Loanwords
14/6	bu-	ma-	bulundà	malundà	*friendship(s)*	Abstract nouns
14/Ø	bu-	Ø	buntu		*humanity (humanness)*	Abstract nouns
14/Øa	buu+	Ø	buumungàngà		*medicine*	Abstract nouns
15/Ø	ku-	Ø	kumanya		*to know*	Infinitives
16/Ø	pa-/ˋ[38]	Ø	pambèlu, pànu		*outside; here*	Locatives
16/Øa	pa+		pa bulààlu		*on the bed*	
17/Ø	ku±/ˋ	Ø	ku mèèsà; kùnu		*at the table; here*	Locatives
			kù baabèndà		*abroad*	
18/Ø	mu±/ˋ	Ø	mu mbekeci		*in the bucket*	Locatives
			mù baamànsèba; mùnu		*at my uncles'; here*	

Notes

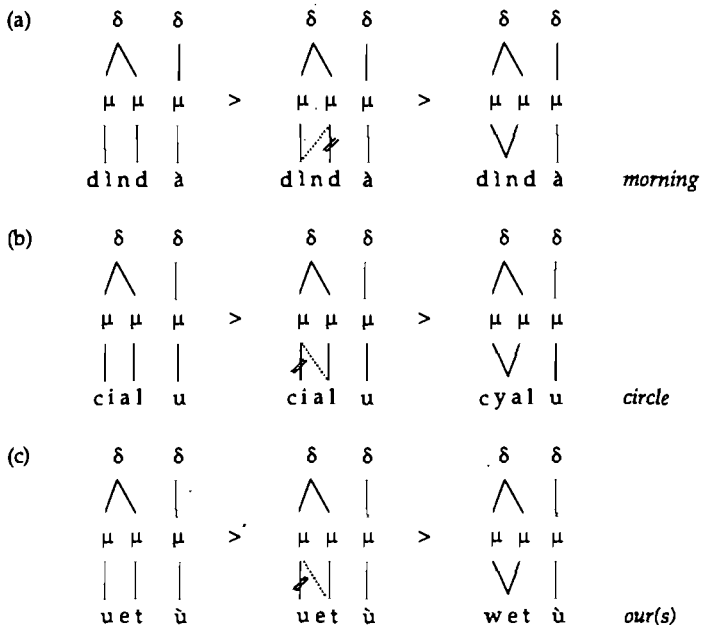
1. In cases where no abbreviated source language is given after the loanwords, the source language is French.
2. The term *Cilubà* refers to the language spoken by the Balubà and the Luluwà or Beena-Luluwà, while Lubà is the corresponding adjective. Cilubà which is classified by Guthrie (1971: 54) as L31, is related to Kisongye L23 (Congo), Kanyok L32 (Congo), Kilubà L33 (Congo) and Kaonde L41 (Zambia). Cilubà and Kiswahili are the main subjects in the Department of African Languages and Cultures of the University of Ghent (Belgium).
3. While French (spoken by barely 10% of the population) is the official language, there exists no legal text bestowing on the four African languages the status of national languages which they enjoyed before independence in 1960. The role of the African languages in the education system has even been restricted to the first two years of primary school, instead of six as in colonial times. Curiously enough, it is during the "authenticity" campaign in 1972 that the role of French has particularly been reinforced. At that time, all the magazines in African languages were suppressed (Ngalasso 1986: 16-20) and a lot of words adapted to French use

(e.g. *septante* became *soixante-dix*, etc.). This did not prevent the African languages from being used intensively for daily communication and, with the collapse of the education system, the expected improvement of competence in French does not seem to have been achieved. It is however true that the exposure to French has been stronger than ever before, which has had an obvious influence on borrowing strategies.

4. Bongo (1977: 360), who was a general secretary at the Ministry of National Education in 1977, gives a figure of 4 500 000, whereas Ngalasso (1986: 12) gives ±3 000 000 and indicates that this figure corresponds to the population supposed to live in the area where Cilubà is actively spoken. All these figures are just guesses, since they do not include the important Lubà communities in Shaba, Kinshasa and elsewhere. Moreover, no statistical surveys have been carried out for several decades.
5. These Cokwe enjoyed such prestige that it became tradition for prospective Luluwà chiefs to travel to Angola to receive recognition mainly in exchange for ivory. Subsequently, many Luluwà chiefs made trade expeditions to Angola. Trade contacts between the Beena-Luluwà, the Cokwe and the Pombeiros (adapted to Bimbàdi in Cilubà) developed particularly in chief Kalamba Mukenge's time (last quarter of the 19th century). Some Luluwà local markets became important trade centres as long-distance trade was developing. Kalamba Mukenge's village in particular, played a major role in the Luso-African trade in Kàsaayì. Angola's influence was so great that the most important post in West-Kàsaayì (which was later to become Luluabourg) was called Malandji (or Malandi), after a location with a similar name in Angola (Malange) (Petridis 1997: 42-45).
6. Kalanda (1963), Mpoyi (1987), Mukenji Mulenga (1981), *Tekemenayi* 1993-1996, unpublished letters in Cilubà from 1960 to 1995. A more comprehensive corpus is being built up in the Department of African Languages and Cultures of the University of Ghent, using modern computer techniques. This will no doubt be very useful for future lexicographical and other linguistic works.
7. Since no study has as yet been carried out to determine the basic vocabulary in Cilubà, I provisionally use this figure which is based on statistics for English (cf. e.g. West 1976 or Bertrand and Lévy 1972), just to show that one needs quite a small number of words to communicate.
8. Foreign verbs (from French) are found mostly in intrasentential code-switching.
9. wu or yi are only written when they are syllabic and in some special cases.
10. One exception is when *n* is in initial position. Cf. note 15.
11. Underlined vowels are nasal.
12. Counts carried out on a 90-minute ordinary conversation recorded on cassette revealed not only that /a/ is the most frequent vowel (followed by either /u/ or /i/ according to whether one considers the H or the L as shown in the charts below), but also that there are 62% of H vs. 38% L.

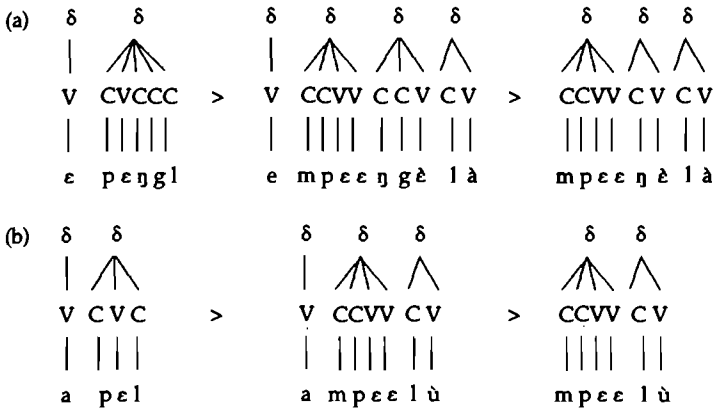
V	%	V	%
a	39,2	à	46,9
u	29,3	ì	23,0
i	18,8	ù	19,2
e	9,0	è	8,3
o	3,5	ò	2,3

13. These rules account for the compensatory lengthening triggered by prenasalization and glide formation. In these processes, the nasal and the high vowels are devocalized and transfer their morae to the vowel placed left and right respectively (Hubbard 1995).



14. In CV-, NCV- and V-type syllables, V can be monomoraic or bimoraic. In CGV types, V is always bimoraic, except in final position.
15. In these syllables, C will normally palatalize. s+i is only found in the emphatic word sì (sì wàyi *he's gone, you know*, sì m̀mêma *it is me, indeed*) and in the verb kusìnsa/kusìnsakaja *to encircle*; n+i is only found in nì (associative *with, and*, or conjunction *that*), ni (conjunction *whether*) and in ni- and nì- (future and concessive tense markers respectively). In other words, s+i and n+i occur almost exclusively in monosyllabic words or as initial syllables. The enclitic emphatic morpheme -s is always attached to the final vowel.
16. These will normally incorporate a G.
17. li>di (an exception to the latter is in the phrase bùcyâcyà bwìlílà *day out day in*). Otherwise, the consonant is palatalized.
18. Similar principles are used in other Bantu languages. Cf. e.g. Kunene (1963) for Southern Sotho or Batibo (1996) for Kiswahili.
19. This happens in other languages too, as in Tswana borrowings (Batibo 1996: 36).
20. We have noticed that even when a vowel is not inserted as in code-switching, the concord still happens in class 12 (clavier kâmvwà musùmba *the keyboard I bought*).
21. This originally Dutch word was either directly borrowed from Dutch (spoken by a great deal of Belgians in Congo) or would have reached Cilubà (and some other African languages too) via Fanagalo. In the Lubà infinitive, the initial syllable pa- was dropped and the remaining part, pronounced [soø] or [soøw] regarded as a root.

22. This phenomenon is known in other languages too: (Ar) mistar, a deverbative noun > (Kswa) mistari 4 *lines*, cf. mstari 3 (Knappert 1970: 81); (Fr) petit pois > (Kisanga) bìtìpwâ 8 *peas*, cf. kìtìpwâ 7 (Coupez 1974).
23. Although this word also exists in class 3 (mushèètà), the existence of kashèètà is opposed to Knappert's (1970: 79) generalization that "the Bantu speakers seem to have rejected this form of the word since the first syllable *ka-* has the shape of the prefix of class 12, which denotes small things; for big things and for things made of wood, the *mu-* prefix is used and has therefore been substituted".
24. This process can be explained as follows:



25. One even finds st in a word which is not originally a loanword: cìtàncist *a person doing business in diamonds* < kutanta *to prosper*.
26. In De Clercq and Willems (1960) there are approximately 210 Lubà nouns with an L np (loanwords were excluded from the count).
27. Some authors do not distinguish between gender and class. Cf. Hinnebusch: "Bantu languages also divide their noun universe into genders ... usually referred to as classes and numbered in singular/plural pairs ... These genders can normally be identified by the shape of the affixes, and if not, then by the grammatical concord they govern" (1989: 466). My definition is very close to Guthrie's: "Chaque fois que des groupes de classes d'un type régulier se rencontrent avec des nominaux indépendants de même radical, ces groupes sont appelés 'genres'. L'espèce de genre la plus commune est celle qui comporte deux classes correspondant à une distinction entre le singulier et le pluriel." Guthrie also defines "class" in the following terms: "Une classe est définie sur le plan morpho-syntaxique comme un schème d'accord bien défini, consistant en le préfixe d'accord d'un nominal indépendant, un ou plusieurs types de préfixes caractéristiques des nominaux dépendants (qualificatifs, démonstratifs, numéraux, etc.) et un préfixe utilisé dans les verbaux, tous les membres de la série des préfixes étant morphonologiquement identiques" (1967: 392). Cf. also Schadeberg: "Enkelvoud en meervoud van telbare naamwoorden horen bij verschillende klassen. Op basis daarvan kunnen de klassen 1 t/m 15 in paren (genera) worden gegroepeerd" (1986: 5).
28. This pair is traditionally represented as 9/10, which is right if one only takes np as classification criterion, with np N- for both the singular and the plural. But for classifying nouns in genders, syntactical concord prevails and therefore it is useless to maintain classes 9/10,

whose concords are exactly the same as those of 1/4. The consequence of this is that class 1 does not contain only human beings with np mu-, but also any noun with prefix N- or ø- (e.g. mwâna ùdi ùnàya *the child is playing*; nkwasa ùdi pambèlu *the chair is outside*; tèlèvìzyôn ùdi pa mèèsà *the TV set is on the table*). It is rather pair 1/2 which characterizes human beings. Except for mungàngà 1/4 *doctor*, there seems to be no human beings in 1/4.

29. As stated in note 26 (De Clercq and Willems 1960) gives about 210 nouns with an L np which are not loanwords. Locative nps bear an L in some words and phrases (e.g. kù baabèndà *in foreign countries*). In the following proverb the locative prefix of class 17 has an L: Bàtu bààya kù baamwandà / Kabàtu bààya kù baaweetù. *One should always be impartial* (literally: *One goes to the matter / One does not go to the brothers*).
30. This is different from: bâna # bănàyi > bâna bănàyi *the children have played.*
31. Imperative form of oppas *to be careful, to pay attention*. Cf. note 21.
32. Variants with an L are just shown by an L sign after the slash, in order to save column space. Thus, lu-/` means lu-/lù-.
33. These two words belong to the religious vocabulary which was almost entirely coined by Bishop A. de Clercq at the beginning of this century, using Latin or Greek as source languages. However, although these neologisms have been used in the Catholic Church for almost a century, the Lubà Bible translators (1994) decided to replace most of them by seemingly more adequate Lubà words or phrases which, beside being generally longer, are polysemous:

aanyìmà	>	mwoyo	*spirit*
bàtismò	>	dyowesha *baptism* < kwowesha	*wash*
bàtistà	>	mwoweshi	*baptist*
bible	>	mukàndà wà Mvìdi Mukùlù	*Bible* (i.e. *Book of God*)
ditùkù dyà nsabato	>	ditùkù dyà cijila	*sacred day*
dyabòlò	>	sàtânà	*devil*
èkèleezìyà	>	cisà cyà Maweeja	*church* (i.e. *people of God*)
èvànjelìyò	>	mukenji mulenga	*gospel* (i.e. *good news*)
kèrùbinè	>	cilòbò cyà mu dyulu	*cherubin* (i.e. *hero from heaven*)
mpàganò	>	mwena cisàmba cikwàbò	*pagan* (i.e. *belonging to another tribe*)
muskribè	>	mumanyi wa dîyi	*scribe* (i.e. *the one who knows the law*)
mwàpostòòlò	>	mutùmìlbwe kùdì Mfùmù	*apostle* (i.e. *the one sent by the Lord*)
mwena Kristò	>	Mwena Yezù	*Christian*
nsàserdòsè	>	mwakwidi	*priest* (in the Old Testament *mulàmbudi* or *mukùbi* are used instead)
paasàkà	>	dipàtuka dyà mu Èjipìtù	*Easter* (i.e. *going out of Egypt*)
-nsanto	>	-a cijila	*sacred*
pèntèkostè	>	cibìlù cyà dinowa	*Pentecost* (i.e. *harvest feast*)
ùkàrìstiyà	>	didyà dyà Mfùmù	*Eucharist* (i.e. *Lord's meal*)
(m)virgò	>	nsongààkàjì mujimè	*virgin*

The Catholic missionaries did not always care about Lubà phonology, which resulted in coining queer words such as *cyàltârè* (altar), *cìshiifèrì* (figure), *mòmpêrè* (father), *Kristò, nkùrusè* (i.e. cross), *Petrò, Markùsè, Màteùsè, Yòwanèsè, Ìzràel, àrâbè*, etc. Protestant missionaries, on the other hand, made a greater effort to adapt their neologisms, e.g. *Kilistò, mucì mucyàmàkàne* (i.e. cross), *Peetèlô, Maakà, Maatààyò, Yona, Iisàlèèlà, aalàbà*, etc. According to Father Paul Lissens (editor of the Catholic Bible, 1994), the Catholics and the Protestants finally agreed to use a unified vocabulary (oral communication, July 1996).

34. Cf. *Tekemenayi* 87:8, 1994.

35. However, it is difficult to say whether a speaker changes attitudes inside the same conversation when he uses different forms of the same loanword, as often happens. On one of our cassettes, the same speaker uses at very short intervals: *mùlâbà, mùrâbà* and *mwena àrâbà* (Arab).

36. Mpoyi (1987: 14) claims that *mpondà* (millet), *tumbumba* (sorghum), *matàbàlà* (taro), *bìlùngà byà nsenga* (sweet potatoes) and *bimenà* (yams) were introduced in the Congo by the Bantu around 2000 BC at the same time as agriculture and handicraft. Unfortunately no sources are mentioned.

37. Linguistic evidence shows that these crops seem to have been known to Proto-Bantu speakers. However, peanuts were either known by the same name as some other crop, or were introduced under various names after Bantu had become current, probably by transfer from terms for some local crops (Guthrie 1970: 30-31). Guthrie also shows that sugar-cane was not known to Proto-Bantu speakers, probably being introduced independently to the east at the end of the Bantu dispersion and in more recent times to the west (1970: 31). According to Gregersen (1968: 3-4, 1977: 149), though, no Proto-Bantu forms are possible for crops which are known to have been introduced no more than 500 years ago, such as *maize* or *peanuts*.

38. Available data suggest that a secondary locative np with L only occurs before a noun with a secondary np (which has a H), e.g.: mù baamànsèba *at my uncles'*, kù baaweetù *at my brothers'*, kù baabèndà *abroad*. Moreover, examples were found only for classes 17 and 18.

Bibliography

ACCT (Agence de Coopération Culturelle et Technique). 1983. *Lexiques thématiques de l'Afrique Central, Zaïre, Ciluba, Activités économiques et sociales.* Paris: CERDOTOLA.

Alexandre, P. 1967. Note sur la réduction du système des classes dans les langues véhiculaires à fonds bantu. *La classification nominale dans les langues négro-africaines* (Colloques Internationaux du Centre National de la Recherche Scientifique, Aix-en-Provence, 3-7 juillet 1967. Sciences Humaines): 277-290. Paris: CNRS.

Bader, Y. and R. Mahadin. 1996. Arabic Borrowings and Code-Switches in the Speech of English Native Speakers Living in Jordan. *Multilingua* 15(1): 35-53.

Batibo, H.M. 1994. Does Kiswahili have Diphthongs?: Interpreting Foreign Sounds in African Languages. *South African Journal of African Languages* 14(4): 180-186.

Batibo, H.M. 1996. Loanword Clusters Nativization Rules in Tswana and Swahili: A Comparative Study. *South African Journal of African Languages* 16(2): 33-41.

Bertrand, Cl.-J. and Cl. Lévy. 1972. *L'anglais de base.* Paris: Classiques Hachette.

Bold, J.D. 1968 (1951). *Fanagalo: Phrase Book, Grammar and Dictionary.* Johannesburg: Hugh Keartland.

Bongo, A. 1977. Zaïre. Sow, I. (Ed.). 1977. *Langues et politiques de langues en Afrique noire: L'expérience de l'Unesco*: 359-363. Paris: Nubia.

Bunduki, K. 1975. Essai de lexique linguistique français-ciluba. Coll. "Travaux et Recherches". Lubumbashi: CELTA (Centre de Linguistique Théorique et Appliquée).

Burssens, A. 1946a. *Tonologische schets van het Tshiluba.* Antwerp: De Sikkel.

Burssens, A. 1946b. *Manuel de Tshiluba.* Antwerp: De Sikkel.

Chuwa, A.R. 1988. Foreign Loan Words in Kiswahili. *Kiswahili* 55(1)-(2): 163-172.

Clark, J.D. 1970. African Prehistory: Opportunities for Collaboration between Archeologists, Ethnographers and Linguists. Dalby, D. (Ed.). 1970. *Language and History in Africa*: 1-19. New York: APC.

Coupez, A. 1954. *Études sur la langue luba.* Tervuren: Musée Royal de l'Afrique Centrale.

Coupez, A. 1974. *Notes du cours de sanga.* Unpublished course. University of Brussels.

Croegaert, L. 1985. *Premières Afriques: Histoire et découvertes d'un continent.* Paris: Didier Hatier.

De Clercq, A. 1914. *Dictionnaire Luba: Luba-Français, Français-Luba.* Brussels: A. Dewit.

De Clercq, A. 1937. *Dictionnaire Luba (Luba-Français).* Léopoldville: Procure des Missions de Scheut.

De Clercq, A. and Willems, E. 1960. *Dictionnaire Tshiluba-Français.* Léopoldville: Imprimerie de la Société Missionnaire de St. Paul.

Dzokanga, A. 1979. *Dictionnaire Lingala-Français.* Leipzig: VEB Verlag Enzyklopädie.

Dzokanga, A. 1995. *Nouveau dictionnaire illustré Lingala-Français.* Paris: INALCO.

Forson, B. 1981. Phonological Regularities in Akan-English Code-Switching. *Studies in African Linguistics*, Supplement (Precis of papers from the Twelfth Conference on African Linguistics, Stanford, April 10-12): 29-34.

Gabriel, Fr. 1921. *Étude du Tshiluba.* Brussels: Ministère des Colonies.

Gabriel, Fr. 1922. *Dictionnaire Tshiluba-Français.* Brussels: A. Dewit.

Gabriel, Fr. 1925. *Dictionnaire Français-Tshiluba.* Brussels: A. Dewit.

Gregersen, E.A. 1968. Words and Things in African Prehistory. *Anthropological Linguistics* 10(3): 1-4.

Gregersen, E.A. 1977. *Language in Africa.* New York: Gordon & Breach.

Guthrie, M. 1967. Variations in the Range of Classes in the Bantu Languages. *La classification nominale dans les langues négro-africaines* (Colloques Internationaux du Centre National de la Recherche Scientifique, Aix-en-Provence, 3-7 juillet 1967. Sciences Humaines): 341-353. Paris: CNRS.

Guthrie, M. 1970. Contributions from Comparative Bantu Studies to the Prehistory of Africa. Dalby, D. (Ed.). 1970. *Language and History in Africa*: 20-49. New York: APC.

Guthrie, M. 1971. *Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages. Part I, Vol. 2.* Farnborough, Hants: Gregg International.

Guthrie, M. et al. 1967. Classe et genre. *La classification nominale dans les langues négro-africaines* (Colloques Internationaux du Centre National de la Recherche Scientifique, Aix-en-Provence, 3-7 juillet 1967. Sciences Humaines): 391-397. Paris: CNRS.

Hinnebusch, Th.J. 1989. Bantu. Bendor-Samuel, J. (Ed.). 1989. *The Niger-Congo Languages*: 450-473. New York: University Press of America.

Hlongwane, J.B. 1995. The Growth of the Zulu Language and its Structural Changes. *South African Journal of African Languages* 15(2): 60-65.

Hubbard, K. 1995. Morification and Syllabification in Bantu Languages. *Journal of African Languages and Linguistics* 16(2): 137-155.

Hyman, L.M. 1975. *Phonology: Theory and Analysis.* New York: Holt, Rinehart and Winston.

Kabuta, J. 1980. *Le thème nominal polysyllabique dans les langues bantoues.* Unpublished M.A. dissertation. University of Brussels.

Kabuta, J. 1995. *De morfologie van het Lubà werkwoord.* Unpublished course. University of Ghent.

Kabuta, J. 1996. *Inleiding tot de structuur van het Cilubà.* Unpublished course. University of Ghent.

Kadima, B.A. et al. 1995. *Terminologie grammaticale et pédagogique: Lexique Français-Ciluba, Ciluba-Français.* Kinshasa: Éditions Universitaires Africaines.

Kalanda, M. 1963. *Tabalayi.* Léopoldville: Concordia.

Kalonji, Z. 1993. *La lexicographie bilingue en Afrique francophone, l'exemple français-cilubà.* Paris: L'Harmattan.

Knappert, J. 1970. Contribution from the Study of Loanwords to the Cultural History of Africa. Dalby, D. (Ed.). 1970. *Language and History in Africa*: 78-88. New York: APC.

Knappert, J. 1989. Les mots swahili empruntés au grec, aux langues romanes et américaines. Le swahili et ses limites. Ambiguïté des notions reçues. Romb, M.-Fr. (Ed.). 1989. *Table ronde internationale du CNRS, Sèvres, 20-22 avril 1983*: 41-57. Paris: Éditions Recherche sur les Civilisations.

Kubela, M.K. 1986. Le schème tonal dans les mots ciluba d'origine étrangère. *Annales Aequatoria* 7: 221-225.

Kunene, D.P. 1963. Southern Sotho Words of English and Afrikaans Origin. *Word* 19(3): 347-375.

Laman, K.E. 1936. *Dictionnaire Kikongo-Français.* Brussels: Mémoires. Institut Royal Colonial Belge. Section des Sciences Morales et Politiques.

Langacker, R.W. 1973. *Language and its Structure: Some Fundamental Linguistic Concepts.* New York: Harcourt Brace Jovanovich.

Maalu-Bungi, L.-L. 1979. La recréation en littérature orale: l'exemple des mankònkù et des màstwâr luluwa. *La civilisation ancienne des peuples des Grands Lacs* (Colloque de Bujumbura, 4-10 septembre 1979, organisé par le Centre de Civilisation Burundaise). Publié avec le concours de l'UNESCO: 16-29. Paris: Karthala-C.C.B.

Meeussen, E. 1944-1959. Syntaxis van het Tshiluba. *Kongo-Overzee* 9 (1-3): 81-105, (4-5): 113-159, 1944-45; 10-11 (1-3): 89-114, (4-5): 218-249, 1946-1947; 12-13 (3): 143-184, 1957; 23 (5): 303-315, 1958; 24 (4-5): 256-264, 1959.

Meeussen, E. 1951. Tooncontractie in het Ciluba (Kasayi). *Kongo-Overzee* 17 (4-5): 289-291.

Meeussen, E. 1962. *Notes de grammaire Luba-Kasayi.* Tervuren: MRAC.

Milroy, J. 1992. Social Network and Prestige Arguments in Sociolinguistics. Bolton, K. and H. Kwok (Eds.). 1992. *Sociolinguistics Today: International Perspectives*: 146-162. London/New York: Routledge.

Morrison, W.M. 1906. *Grammar and Dictionary of the Buluba-Lulua Languages (As Spoken in Upper Kassai and Congo Basin).* New York: American Tract Society.

Morrison, W.M. 1939. *Dictionary of the Tshiluba Language.* Revised and enlarged by a Committee of the American Presbyterian Congo Mission. Luebo: John Leighton Wilson Press.

Mpoyi, M. 1987. *Lwendu lwa Baluba.* Mbujimayi: Diyoseze dya Mbujimayi.

Mukendi, K. 1989. *Ciluba cikendame.* Unpublished lecture organized by the association Lugha za Afrika and presented at the University of Brussels.

Mukenji Mulenga (New Testament). 1981. Mbujimayi: Diyoseze dya Mbujimayi.

Mutombo, H. 1977. *Les variations linguistiques en Lubà-Kàsaayì.* Unpublished Ph.D. dissertation. Lubumbashi: Université Nationale du Zaïre.

Ngalasso, N.-M. 1986. État des langues et langues de l'État au Zaïre. *Politique Africaine* 23: 6-27 *(Des langues et des États).* Paris: Karthala.

Petridis, C. 1997. *Context en morfologie van de plastische kunst bij de Luluwa (Zuid-Centraal Zaïre), Vol. I.* Unpublished Ph.D. dissertation. University of Ghent.

Richardson, I. 1967. Linguistic Evolution and Bantu Noun Class Systems. *La classification nominale dans les langues négro-africaines* (Colloques Internationaux du Centre National de la Recherche Scientifique, Aix-en-Provence, 3-7 juillet 1967. Sciences Humaines): 373-390. Paris: CNRS.

Robins, R.H. 1975. *General Linguistics: An Introductory Survey.* London: Longman.

Samain, A. s.a. *La langue kisonge.* Bibliothèque-Congo XIV. Brussels: Goemaere.

Samarin, W. 1966. Self-annulling Prestige Factors among Speakers of a Creole Language. Bright, W. (Ed.). 1966. *Sociolinguistics: Proceedings of the UCLA Sociolinguistics Conference, 1964*: 188-213. The Hague/Paris: Mouton.

Schadeberg, Th. 1986. *Kleine structuurcursus Umbundu.* Leyden: Rijksuniversiteit.

Stappers, L. 1943-1945. De middelhooge toon in het Tshiluba. *Kongo-Overzee* 9-10 (4-5): 261-264.

Stappers, L. 1949. *Tonologische bijdrage tot de studie van het werkwoord in het Tshiluba.* Brussels: K.B.K.I.

Tekemenayi. 1993-1996. 77: 5, 7, 9, 17; 87: 3, 8-10; 88: 6, 7-11; 89: 3, 5, 6, 9, 11; 90: 3, 11; 91: 6, 7; 99: 6, 7, 9.

Van Avermaet, E. and Mbuya, B. 1954. *Dictionnaire Kilubà-Français.* Tervuren: MRAC.

West, M. 1976. *A General Service List of English Words.* London: Longman.

Whiteley, W.H. 1967. Swahili Nominal Classes and English Loan-Words: A Preliminary Survey. *La classification nominale dans les langues négro-africaines* (Colloques Internationaux du Centre National de la Recherche Scientifique, Aix-en-Provence, 3-7 juillet 1967. Sciences Humaines): 157-174.

Willems, E. 1986. *Dictionnaire Français-Tshiluba.* Kananga: Éditions de l'Archidiocèse.

Yukawa, Y. 1992. A *Classified Vocabulary of the Luba Language.* Tokyo: ILCAA (Institute for the Study of Languages and Cultures of Asia and Africa).

Analysis of the Word-Initial Segment with Reference to Lemmatising Zulu Nasal Nouns

M.H. Mpungose, *Zulu Dictionary Project, University of Zululand, Durban-Umlazi Campus, Isipingo, South Africa*

Abstract: The process of lemmatising nasal nouns in the Zulu lexicon is problematic. The traditional method is to lemmatise a Zulu lexical noun by etymological noun-stem. This practice creates difficulties in harmonising lexical nouns with their syntactic application. Most authors and dictionary-makers are inconsistent in identifying the word-initial segment which determines the letter of the alphabet under which the lexical noun should be included. Consequently, dictionary users do not find Zulu dictionaries user-friendly. This article therefore proposes the principle of "a noun without initial vowel" as a method for lemmatising Zulu nasal nouns. It concludes that it is not necessary to delve into the derivational history of a lexical noun, but rather to focus on the product of the operation of morphophonological rules. The article also suggests the need to identify the distinctiveness of the segments of a syllable and to acknowledge that identical forms of a segment do occur at different segmental positions (initial, medial and final). Finally it is argued that the Zulu nasal noun class prefix is constructed according to an open syllable pattern defined by a general CV-formula based on a VCV noun prefix open syllable pattern.

Keywords: ADJOINED LETTER, COMPOUND, COMPOSITE, CONSONANT, ELEMENT, ETYMOLOGICAL, EVOLUTIONARY, HOMORGANIC, INITIAL, INTRAVOWEL, LEMMA, LEMMATISE, LEXICAL, MORPHOPHONOLOGICAL, NASAL, NOUN CLASS PREFIX, SEGMENT, SYLLABLE, VOWEL

Opsomming: Ontleding van die woordinisiële segment met verwysing na die lemmatisering van nasale naamwoorde in Zoeloe. Die proses van lemmatisering van nasale naamwoorde in die Zoeloeleksikon is problematies. Die tradisionele metode is om leksikale selfstandige naamwoorde in Zoeloe volgens die etimologiese naamwoordstam te lemmatiseer. Hierdie gebruik veroorsaak moeilikhede by die harmonisering van leksikale selfstandige naamwoorde met hul sintaktiese toepassing. Die meeste outeurs en leksikograwe is inkonsekwent in die identifisering van die woordinisiële segment wat die letter van die alfabet bepaal waaronder die leksikale selfstandige naamwoord geplaas moet word. Gevolglik vind woordeboekgebruikers Zoeloewoordeboeke nie gebruikersvriendelik nie. In hierdie artikel word die beginsel van "'n selfstandige naamwoord sonder inisiële klinker" dus voorgestel as 'n metode om nasale naamwoorde in Zoeloe te lemmatiseer. Daar word tot die gevolgtrekking gekom dat dit nie nodig is om op die afleidingsgeskiedenis van 'n leksikale naamwoord in te gaan nie, maar dat daar eerder gefokus moet word op die produk van die werking van die morfofonologiese reëls. Die artikel gee dit ook ter oorweging dat dit nodig is om die onderskeidende kenmerke van die segmente van 'n sillabe te identifiseer en om te erken dat identiese vorme van 'n segment in verskillende segmentele posisies (inisieel, mediaal en finaal) voorkom. Ten slotte word voorgestel dat die prefiks van die nasale

naamwoord in Zoeloe saamgestel word aan die hand van 'n oop sillabepatroon gedefinieer deur 'n algemene KV-formule gebaseer op 'n VKV-naamwoordprefiks-oopsillabepatroon.

Sleutelwoorde: NAASLIGGENDE LETTER, SAMESTELLING, SAAMGESTEL, KONSONANT, ELEMENT, ETIMOLOGIES, EVOLUSIONÊR, HOMORGANIES, INISIEEL, INTRAKLINKER, LEMMA, LEMMATISEER, LEKSIKAAL, MORFOFONOLOGIES, NASAAL, NAAMWOORDKLAS, SEGMENT, SILLABE, KLINKER

1. Introduction

Both traditional methods for lemmatising Zulu lexical items, namely the word-stem tradition and the full-word tradition, come in for criticism. These methods lead to controversies among dictionary users, dictionary makers, linguists and metalinguists. Marggraff (1997) agrees with Van Wyk's statement (1995: 82-83) that:

> A dictionary is a compilation of lexical items, not a grammar ... All dictionaries therefore assume grammatical knowledge on the part of the user ... In the case of most languages a knowledge of important facts concerning the morphology and even some morphophonological processes are also regarded as a necessary prerequisite to the use of a dictionary.

The application of these two traditional methods in the Zulu dictionary-making process results in dictionaries which:

— are not user-friendly,
— assume grammatical knowledge,
— are linguistically inconsistent, and
— are uneconomical.

The basic function of a dictionary is to provide the dictionary user with comprehensible internal lexicographical information that ought to satisfy the user's needs at that moment. The lemma (head-word), as the most important dictionary entry, commands access into the required internal lexicographical information and indicates "each respective lexical unit in its canonical form" (Zgusta 1971: 249-250). The meaningful rapport between the lemma in Zulu dictionaries and Zulu dictionary users is therefore the primary concern in this analysis of the Zulu nasal noun initial segment:

> Most lexicographers derive at least some satisfaction from the knowledge that the product of their labours can help ordinary language users in situations of communicative conflict or deficit (Hartmann 1983: 6).

1.1 Historical background

As early as 1857 J.L. Döhne and later also J.W. Colenso (1861, 1905), A.T. Bryant (1905, 1917), C. Roberts (1905) and R.C.A. Samuelson (1923) were already indi-

vidually confronted with the problem of lemmatising Zulu lexical items. Subsequent recent contributions, including those by Doke et al. (1948, 1958, 1964, 1990), A.C. Nkabinde (1982, 1985) and S.L. Nyembezi (1992) among others, do not offer a method different from the two traditional methods for lemmatising Zulu lexical items. The difficulty is however unfairly borne by dictionary makers who were and are expected:

(i) to identify the Zulu lexical word as distinct from the grammatical word;
(ii) to be conversant with the word-initial segment which determines the letter of the alphabet under which Zulu lexical items are recorded;
(iii) to harmonise the Zulu lexicographical rules with linguistic factors;
(iv) to coalesce lexical rules with rules that govern Zulu orthography; and
(v) to choose a method of lemmatising words that seems best for Zulu lexical items (i.e. listing words, as Taylor (1991: 179) says, "in their appropriate phonological form").

1.2 Textual examples

In support of the hypothesis that the initial nasal consonant segment of a noun-stem is "homorganically pronounced" (Malmkjaer 1991: 26), and therefore needs not be depleted from its composite nasal consonant segment when a nasal noun is lemmatised, this article uses data taken from two Zulu monolingual dictionaries and one bilingual dictionary where more examples can be found:

(i) Monolingual dictionaries

A.C. Nkabinde: *Isichazamazwi* 2 (1985) (henceforth ACNK)
S.L. Nyembezi: *aZ Isichazimazwi Sanamuhla Nangomuso* (1992) (henceforth SLNY)

(ii) Bilingual dictionary

C.M. Doke et al.: *English – Zulu / Zulu – English Dictionary* (1990) (henceforth DMSV)

1.3 Lemmatising tendency

The three dominant tendencies adopted by most dictionary makers are:

(i) The Zulu lemmatising practice is based on well-defined lexical theories formulated to suit languages with distinctive writing systems in which the first letter of a noun or a noun-stem is distinct. For example, Table 1 demonstrates that the initial segments in the words ***pig*** and ***kolobe*** are distinct forms. There can be no arbitrariness in lemmatising or looking up these nouns under the letter of the first consonant segment, viz. **P** and **K** respectively. In Zulu the noun ***ingulube*** is recorded under the letter **N**

by ACNK, DMSV and SLNY while Döhne (1857) lists it under the letter G. However, ACNK, DMSV and SLNY inconsistently lemmatise most nasal nouns which are classified by form with the noun ***ingulube*** either under one letter or under more than one letter. This tendency is evident even with the same author (see examples in Tables 9(a) and 10(a)).

Table 1: Conjunctive and disjunctive writing system

English:	**The pig** squealed the whole night.
Sotho:	**Kolobe** e tsetsetse bosiu bohle.
Zulu:	**Ingulube** itswininize kwaze kwasa.

(ii) The same lexical item in Zulu is included under or excluded from the letter **N** or **M** or lemmatised under more than one letter in the case where the phonemic segment of the syllable-initial consonant of the stem, homorganically pronounced, forms the combination in the nasal compound consonant segment or the nasalised consonant segment.

Table 2: Multiletter inclusion or exclusion

The lemma for ***imbuthuma*** is:	>	**-mbuthuma** or
	>	**-buthuma** or
	>	**-bhuthuma.**
The lemma for ***intshakaza*** is:	>	**-ntshakaza** or
	>	**-tshakaza** or
	>	**-shakaza.**

(iii) The etymological structure of the lexical item at premorphophonological level is used to determine the base-form for the canonical form of the lemma.

Table 3: The etymological lexical structure

PREMORPHOPHONOLOGICAL LEVEL	PHONOLOGICAL PROCESS		LEXICAL ITEM
-buthuma	N + b = **mb**	>	***imbuthuma***
-bhuthuma	N + bh = **mb**	>	***imbuthuma***
-tshakaza	N + tsh = **ntsh**	>	***intshakaza***
-shakaza	N + sh = **ntsh**	>	***intshakaza***
-hambiso	N + h = **nk**	>	***inkambiso***
-linganiso	N + l = **nd**	>	***indinganiso***

1.4 Lack of user-friendliness

These methods of lemmatising Zulu nasal nouns are inconsistent and therefore inconvenient and not user-friendly to dictionary users. The lexemes listed in the speaker's lexicon are those that are regular in his/her current language (Bauer 1988: 195). Thus the words lemmatised in a dictionary should be those that are actually used. Posthumus (1994: 35) discourages the application of etymological forms and says: "It is accepted that language does not operate with non-existing or meaningless forms" (such as those illustrated in Tables 2 and 3, e.g. **-buthuma, -bhuthuma, -tshakaza, -shakaza, -hambiso, -linganiso** — my examples). To use them as lemmas is therefore "unacceptable". The article postulates that lemmas need to be perceived and treated at a level of "the output of the morphological and phonological rules of the different strata put together" (Malmkjaer 1991: 323, see also Katamba 1989: 257-258). Hence the dictionary user can look up the morphophonological product but not its evolutionary process under the lemma in the dictionary.

2. A nasal consonant segment

The protoforms of "all class prefixes are bimorphic in Zulu" except class prefix 1(a) which is "truly monomorphic" (Hlongwane 1995: 62). The evolutionary noun class prefix reduction processes (Mini 1992, 1995) produce noun classes prefixed by "abstract segments" (Katamba 1989: 181, Zgusta 1971: 120) which have perceptible structures of noun class prefixes in some lexical items, for example, **-li-** in noun class (5), **-ni-** in noun classes (9) and (10), and **-lu-** in noun class (11):

noun class (5) **-li-**:				
indiki	*ili*ndiki	lemma	>	**-ndiki**
intshontsho	*ili*ntshontsho	lemma	>	**-ntshontsho**
inkankane	*ili*nkankane	lemma	>	**-nkankane**
inono	*ili*nono	lemma	>	**-nono**
inuku	*ili*nuku	lemma	>	**-nuku**
noun class (9) **-ni-**:				
imbaba	*ini*mbaba	lemma	>	?
inhlwa	*ini*nhlwa	lemma	>	?
imvula	*ini*mvula	lemma	>	?
intshakaza	*ini*ntshakaza	lemma	>	?
ingulube	*ini*ngulube	lemma	>	?
noun class (10) **-ni-**:				
izinhlwa	*izi-*n*i*nhlwa	lemma	>	?
izimvula	*izi-*n*i*mvula	lemma	>	?
izintshakaza	*izi-*n*i*ntshakaza	lemma	>	?
izingulube	*izi-*n*i*ngulube	lemma	>	?

noun class (11) **-lu-:**

umbimbi	*u-lumbimbi*	lemma	>	**-mbimbi**
undanda	*u-lundanda*	lemma	>	**-ndanda**
unyawo	*u-lunyawo*	lemma	>	**-nyawo**
unjongwe	*u-lunjongwe*	lemma	>	**-njongwe**
umonya	*u-lumonya*	lemma	>	**-monya**

The nasal noun classes (9) and (10) are unique and distinct. They consist of:

— the protoforms of the nasal noun class prefix which have an intravowel nasal consonant element;
— the perceptible protoforms of the nasal noun-class prefix which become homorganic with the initial consonant of the noun-stem; and
— the productive rule that governs incompatibility of adjoined consonant sounds.

The combination of protoforms of the nasal noun-class prefix with the initial consonant of noun-stems under noun classes (9) and (10) creates problems regarding the lemmatising of Zulu nasal nouns. Meinhof (1932: 95-96) states:

> When a nasal is preceded by ***n*** < B. ***ni*** the ***n*** disappears, e.g.
> ***n*** + ***ny*** > ***ny***, e.g. ***u-nyawo*** (11) "foot" pl. ***izinyawo*** for ****izin-nyawo***.
> ***n*** + ***n*** > ***n***, e.g. ***izi-ne*** (10) "four" > ****izin-ne***.
> ***n*** + ***m*** > ***m***, e.g. ***u-monya*** (11) "python" pl. ***izi-monya*** for ****izin-monya***, cf. ***i-mini*** (9) "middle of the day" pl. ***izi-mini*** for ****izin-mini***.
> The nasals remain unchanged before the vowels.
> (B. denotes the protoforms of Ur-Bantu, and * denotes protoforms.)

On the basis of a proposition like ***i-mini : izi-mini : *izin-mini***, this article attempts to test analogically by a reverse formula that lemmatising practice needs to ignore the morphophonological process in order to pick up initial segments of the noun-stems, that are the morphophonological products. For example, the noun ***intshakaza***, grouped in the same nasal noun class (9) as ***imini***, is selected here to form the problematic part of the equation, that is: If ***imini*** : ***-mini***, then ***intshakaza*** : ?

If						
imini (class 9)	>	***i-mini***	:	***izi-mini***	:	****izin-mini*** (given) and
		i-mini	:			****in-mini*** (formula),
then						
intshakaza (class 9)	>	***i-ntshakaza***	:	***izi-ntshakaza***	:	****izin-ntshakaza*** and
		i-ntshakaza	:			****in-ntshakaza***;
therefore, the lemma for		***imini***	>	***-mini*** and		
	for	***intshakaza***	>	***-ntshakaza.***		

It is logical to accept that the same theory holds in noun classes (9) and (10) with all nasal nouns whose initial nasal consonant segments are either simple

or homorganic. For example, the proposed solution to find lemmata for lexical items under noun class (9) listed as examples above is:

imbaba	>	**-mbaba**
inhlwa	>	**-nhlwa**
imvula	>	**-mvula**
intshakaza	>	**-ntshakaza**
ingulube	>	**-ngulube**

Note that the evolutionary forms of the noun ***intshakaza***, illustrated briefly in Table 14(a) showing developments from Ur-Bantu **-Ni-**, are too remote to be discussed in detail under lemmatisation. Hence the form ****in-ntshakaza*** is conveniently treated as example to indicate the "protoform" at premorphophonological level.

2.1 Types of nasal consonant segments

The two types of nasal consonants that occur in Zulu as consonant segments are:

— the primary nasal consonant segments, and
— the secondary nasal consonant segments.

They can be summarised schematically as in Table 4 where their relations are shown.

2.1.1 The primary nasal consonant segments

The primary or radical nasal consonant segment is found in Zulu as:

— a simple alveolar nasal consonant segment **n** [n];
— a simple bilabial nasal consonant segment **m** [m];
— a syllabic bilabial nasal consonant segment **m̩** [m̩] which occurs only if it precedes an adjoined consonant, whereas a syllabic alveolar nasal consonant is not permitted in Zulu; and
— an intravowel nasal consonant element, that is, a nasal consonant segment occurring between vowels with no adjoined consonants since a nasal consonant is never preceded by an adjoined consonant in Zulu, unless the preceding adjoined consonant is syllabic.

The features illustrated in Table 5 show no divergence in the system of lemmatisation by the Zulu dictionary makers ACNK, DMSV and SLNY. Under the primary nasal consonant segment type only the syllabic bilabial nasal consonant cluster reflects some problems. The combination of syllabic **m̩** with the initial consonant of the first syllable of the noun-stem creates indistinct forms of initial segments. This is discussed under 2.2 and shown in Tables 6, 7 and 8.

Table 4 **TYPES OF NASAL CONSONANTS**

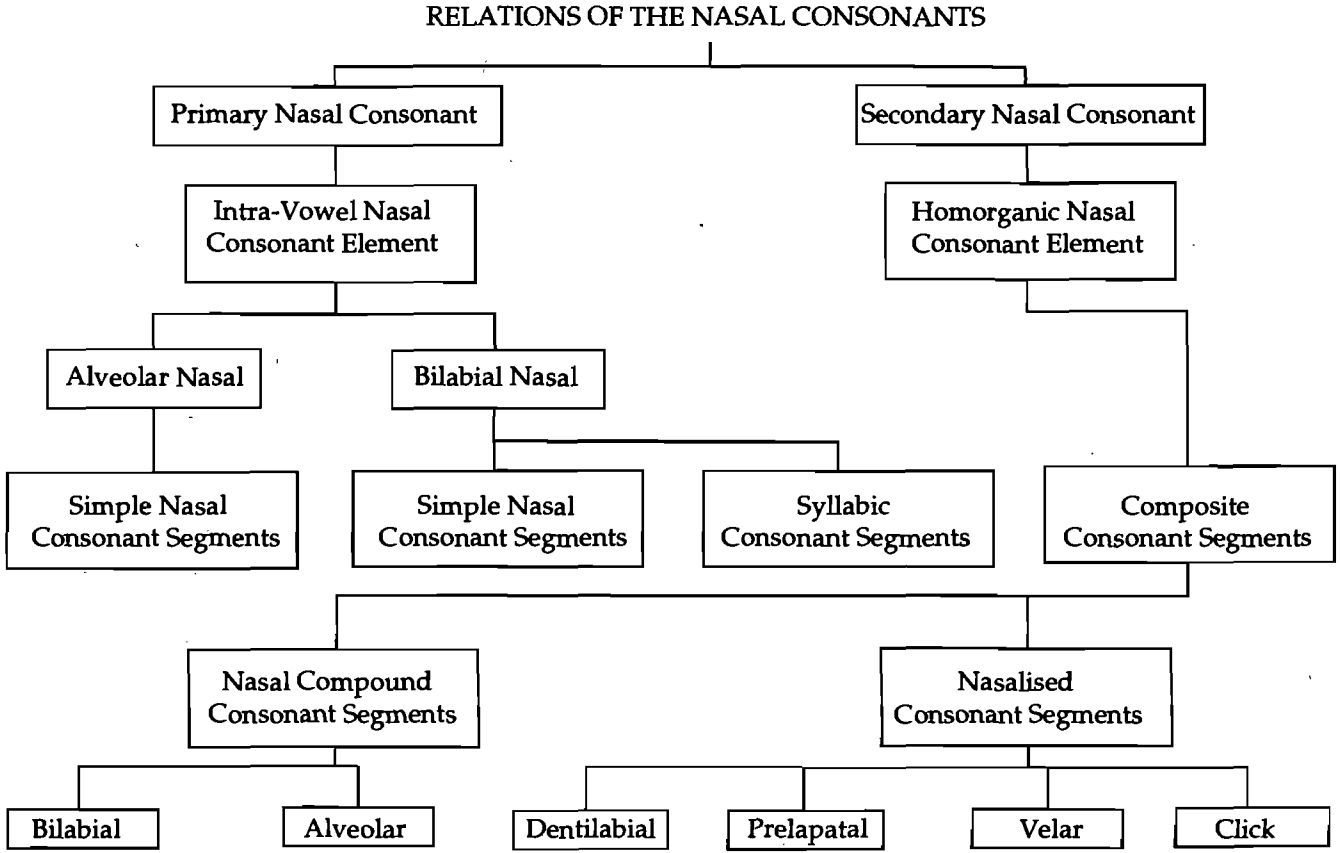


Table 5: The primary nasal consonant segments

Lexicon with Intravowel Nasal Consonant	Postulated Initial Consonant Segment	Letters under which Lemmatised by		
		ACNK	DMSV	SLNY
inamfunamfu	**n**	—	N	N
inembe	**n**	N	N	N
inomfi	**n**	—	N	N
inunu	**n**	—	N	N
imamba	**m**	—	M	M
imeli	**m**	M	M	M
imenenja	**m**	—	M	M
imini	**m**	M	M	M

2.2 Syllabic bilabial nasal consonant vs. homorganic nasal consonant segments

The ability to distinguish homographic consonant clusters between the syllabic bilabial nasal and the homorganic nasal consonant segments of written lexical items in general Zulu orthography presupposes a knowledge of the Zulu word structure, otherwise lemmatisation becomes problematic. This is attributable to current orthography. Therefore, problems in lemmatisation cannot be solved either with phonetic scripts or with specific diacritic markings for the syllabicity of segments. Hence it becomes difficult to determine or to explain the lexical rule to be applied for the initial consonant segment in both the syllabic bilabial nasal consonant and the homorganic nasal consonant segments occurring at the same segmental environment in a Zulu lexical item.

Doke et al. (1990) acknowledge the existence of this problem. For example, they divide the bilabial consonant into two dictionary alphabets, viz. **B** (or **b**) and **ɓ**, whereby **B** is an equivalent of **Bh** (devoiced bilabial explosive) and **ɓ** an equivalent of **B** (voiced bilabial implosive), that is, **B** = **Bh** and **ɓ** = **B**, *but* **ɓ** ≠ **Bh**, hence, **n** + **B** > **mb**, **n** + **Bh** > **mb**, and **n** + **ɓ** > **mb** (but either **mbh** < **n** + **Bh** or **mɓ** < **n** + **ɓ** as a single segment is incompatible). They (1990: 15, 57) assert that the two consonants **B** (or **b**) and **ɓ** are "phonetically distinct" in Zulu:

> Stems of nouns commencing in ***imb*** are sometimes recorded under ***ɓ***; sometimes under ***mb***. When, however, it is ascertainable that the initial of the root is ***b***, and in cases where the real initial is to-day unascertainable, these words are recorded under ***b***.

The problem is comparatively illustrated in Table 6 by using the "minimal pair test" propounded by Katamba (1989: 22-23).

Table 6: The minimal pair test

Syllabic Nasal at Initial Consonant Clusters	Lexical Item	Homorganic Nasal Initial Consonant Segment	Lexical Item
mb	*imbala* (class 4)	mb	*imbala* (class 9)
mb	*umbala* (class 3)	mb	*umbalane* (class 3a)
mb	*umbethe* (class 3)	mb	*umbekle* (class 3a)
mf	*imfula* (class 4)	mf	*imfumba* (class 9)
mv	*imvithi* (class 4)	mv	*imvithi* (class 9)

Table 7: The syllabic nasal consonant segments

Lexicon with Syllabic Nasal at Initial Consonant Clusters	Postulated Initial *Consonant* Segment	Letters under which Lemmatised by		
		ACNK	DMSV	SLNY
imbala	B	—	6	B
umbala	B	B	6	B
umbethe	B	—	6	B
imfula	F	—	F	F
imvithi	V	—	V	V

Table 8: The homorganic nasal consonant segments

Lexicon with InitialHomorganic Nasal Consonant Segment	Postulated Initial Consonant Segment	Letters under which Lemmatised by		
		ACNK	DMSV	SLNY
imbala	mb	Mb	Mb/B	Mb/Bh
umbalane	mb	—	Mb	Mb
umbekle	mb	—	B	Bh
imfumba	mf	—	F	Mf/F
imvithi	mv	—	V	V

2.3 The secondary nasal consonant segments

Two secondary nasal types of consonant segments that cause a major problem in lemmatising a Zulu lexical item are:

— the nasal compound consonant segments, and
— the nasalised consonant segments.

Each type of nasal consonant segment is generated by phonological factors. Each consists of composite consonant clusters with a nasal consonant which is homorganically pronounced. Furthermore, each type of segment constitutes a single (phonetic) sound. The first actual consonant of a noun-stem therefore determines the letter under which the lexical item is recorded (see Tables 9(a) and (10)(a)). It also distinguishes each lexical item by its form and word meaning (Tables 6 and 12). The fact that the actual nasal consonant is a catalyst for nasal compound consonant segments, makes it distinct from the nasalised type of consonant segments. This feature is only realised by use of phonetic script at postmorphophonological level. The distinction is illustrated in the column "Postulated Initial Homorganic Nasal Consonant" in Tables 9 and 10.

Table 9: The nasal compound consonant segments

PREMORPHOPHONOLOGICAL LEVEL	POSTMORPHOPHONOLOGICAL LEVEL	
	Postulated Initial Homorganic Nasal Consonant Segment	**Phonetic Sounds**
(i) Bilabial Consonant Segments		
N + b	> **mb**	> [mb]
N + bh	> **mb**	> [mb]
N + ɓ	> **mb**	> [mb]
N + p	> **mp**	> [mp']
N + ph	> **mp**	> [*mp'*]
(ii) Alveolar Consonant Segments		
N + d	> **nd**	> [nd]
N + l	> **nd**	> [nd]
N + t	> **nt**	> [nt']
N + th	> **nt**	> [nt']
N + s	> **ns**	> [nts']
N + z	> **nz**	> [ndz]
N + dl	> **ndl**	> [ndɬ]
N + hl	> **nhl**	> [ntɬ']

Table 9(a): The nasal compound consonant segments

Lexicon with Initial Homorganic Nasal Consonant Segments	Postulated Initial Homorganic Nasal Consonant Segments	Letters under which Lemmatised by ACNK	DMSV	SLNY
imbaba	mb	Mb/–/Bh	Mb/B/Bh	Mb/B/Bh
imbiza	mb	Mb/–/Bh	–/–/Bh	Mb/–/Bh
imbuthuma	mb	–/–/–	–/–/Bh	Mb/B/Bh
impandla	mp	–/–/Ph	–/–/Ph	Mp/–/Ph
impoko	mp	–/P/–	–/P/–	Mp/–/–
impontshi	mp	–/–/Ph	Mp/–/–	Mp/P/Ph
indaxandaxa	nd	–/D/–	Nd/D/–	Nd/D/–
indikimba	nd	–/D/–	–/D/–	Nd/D/–
indinganiso	nd	–/–/–	Nd/–/L	Nd/–/–
indodakazi	nd	–/D/–	–/D/–	–/D/–
intezazane	nt	–/–/Th	Nt/–/–	Nt/–/Th
intombi	nt	–/–/–	–/–/Th	Nt/–/Th
insada	ns	–/S/–	Ns/–/–	Ns/S/–
insephe	ns	Ns/–/–	Ns/S/–	Ns/S/–
insumpa	ns	–/S/–	–/S/–	Ns/S/–
inhliziyo	nhl	–/–/–	–/Hl/–	Nhl/–/–
inhlwa	nhlw	–/Hl/–	Nhlw/Hl/–	–/Hl/–

Table 10: The nasalised consonant segments

PREMORPHOPHONOLOGICAL LEVEL	POSTMORPHOPHONOLOGICAL LEVEL	
	Postulated Initial Homorganic Nasal Consonant Segment	Phonetic Sounds
(i) Dentilabial Consonant Segments		
N + f	> mf	> [ɱf]
N + v	> mv	> [ɱv]
(ii) Prepalatal Consonant Segments		
N + y	> ny	> [ɲ]
N + j	> nj	> [ɲdʒ]
N + sh	> ntsh	> [ɲtʃ]
N + tsh	> ntsh	> [ɲtʃ]

(iii) **Velar Consonant Segments**

N + g	> **ng**	> [ŋg]
N + k	> **nk**	> [ŋk']
N + kh	> **nk**	> [ŋk']
N + h	> **nk**	> [ŋk']
N + kl	> **nkl**	> [ŋkɬ']

(iv) **Clicks Consonant Segments**

N + ch	> **nc**	> [ŋ/]
N + c	> **ngc**	> [ŋ/g]
N + gc	> **ngc**	> [ŋ/g]
N + qh	> **nq**	> [ŋ!]
N + q	> **ngq**	> [ŋ!g]
N + gq	> **ngq**	> [ŋ!g]
N + xh	> **nx**	> [ŋ//]
N + x	> **ngx**	> [ŋ//g]
N + gx	> **ngx**	> [ŋ//g]

Table 10(a): The nasalised consonant segments

Lexicon with Initial Homorganic Nasal Consonant Segment	Postulated Initial Homorganic Nasal Consonant Segment	Letters under which Lemmatised by		
		ACNK	DMSV	SLNY
imfe	**mf**	–/F/–	–/F/–	Mf/F/–
imfene	**mf**	–/F/–	–/F/–	Mf/F/–
imvakazi	**mv**	–/V/–	–/V/–	Mv/V/–
imvula	**mv**	–/–/–	Mv/V/–	Mv/V/–
inyama	**ny**	–/–/–	Ny/–/–	Ny/–/–
inyanga	**ny**	Ny/–/–	Ny/*Vowel* A	Ny/–/–
inyoka	**ny**	Ny/–/–	Ny/*Vowel* O	Ny/*Vowel* O
inja	**nj**	Nj/–/–	Nj/J/–	Nj/–/–
injobo	**nj**	–/–	–/J/–	–/J/–
intsha	**ntsh**	–/–/Sh	–/–/Sh	–/–/Sh
intshakaza	**ntsh**	–/–/Sh	Ntsh/Tsh/–	Ntsh/Tsh/Sh
ingaco	**ng**	–/G/–	Ng/–/–	Ng/G/–
ingobo	**ng**	Ng/–/–	Ng/G/–	Ng/G/–
inkambiso	**nk**	–/–/H	Nk/–/H	Nk/–/H
inkantolo	**n**	–/–/–	Nx/–/Xh	Nx/–/–

3. Analysis

The dictionary entry consists of two parts (Zgusta 1971: 249-252). The first part is called the lemma (head-word), which indicates the lexical item itself. This part is most important since it acts as a lexicographic information cursor by which the dictionary entry is identified. The second part contains all information that refers to the first part. Access to the second part is never direct. It is always reached through the first part.

Bauer (1988: 9) comments on how words differ and states that:

— the *grammatical word* is discussed in terms of *its description*, and comprises the second part of the dictionary entry, e.g. verb, noun, past participle, etc., while
— the *dictionary word* is discussed in terms of *its form* (orthography or spelling) and comprises the first part of the dictionary entry, i.e. the lemma (head-word).

This article therefore analyses the first part of the dictionary entry.

3.1 The syllable and the segment

The syllable is a unit intermediate between the segment and the word. The Zulu lexical item is constructed on an open syllable pattern. For the facilitation of the lexicographic principle of lemmatisation, each syllable structure is perceived in this article to consist of a consonant segment C plus a vowel segment V, based on a CV-formula. According to Abercrombie (1980: 39, 42), a vowel is a segment of a syllable which can stand alone as a syllable to form a lone-vowel syllable, while a consonant is a segment of a syllable which, when placed alone as a syllable, becomes subjected to laws that regulate syllabic consonants and sounds collocation. This article examines the form and function of a nasal noun initial segment by which the lemma is identified in a Zulu dictionary. Abercrombie (1980: 39) asserts that "the segments of the syllable are *identified* by their sound".

3.1.1 Form and function

The algorithm is conveniently derived from Clements and Keyser (1983). It is applied with the purpose to define the form (Tables 11(a) and 11(b)) and function (Table 12) of a Zulu syllable. It forms the basis for arguments and examples throughout this article. It maintains that:

(i) fully formed core syllables must satisfy the language-particular syllable structure conditions;
(ii) fully formed core syllables are constructed in a pattern so that:
    — V-elements are prelinked to core syllables and C-elements to the left are adjoined one by one as long as the configuration resulting at each step satisfies all relevant syllable structure conditions;

— the syllable-initial consonants are maximised to the extent consistent with the syllable structure conditions of the language in question; and
— the construction of a syllable is in onion-like fashion, that is, built up from the centre outward (Clements and Keyser 1983: 38), hence the rule that applies to the identification of a segment can be likened to "the onion metaphor" (Katamba 1989: 258-259).

Similarly, the segment in a lexical item from the Zulu lexicon is structured in the same fashion as characterised above. The difference lies in the autonomy of the phonetic value of each segment. Each segment, either the vowel segment or the consonant segment, constitutes its own sound by which it is identifiable. The form of a syllable pattern in a lexical nasal noun can be featured as follows:

Table 11(a) ***intshakaza*** = V + CV + CV + CV
Table 11(b) ***intsha*** = V + CV

The first V-element represents a lone-vowel syllable and/or a vowel segment. It functions as a word-category marker (see Tables 11(a), 11(b), 14(a) and 14(b)). Hence, it influences change in lexical form and contributes to change in a word meaning. The noun in Table 11(a) is constructed of four syllables or of seven segments (4V + 3C) and in Table 11(b) of two syllables or three segments (2V + C). While proposals in this article are to:

— lemmatise a nasal noun without its initial vowel, and
— recognise the first homorganic nasal consonant in a nasal composite consonant segment of a noun-stem for lemmatisation,

the structure of a lemma can therefore for example be constructed as follows:

Table 11(a), the lemma for ***intshakaza*** > **-ntshakaza** (CV + CV + CV), and
Table 11(b), the lemma for ***intsha*** > **-ntsha** (CV).

Table 11(a): Form of a syllable and a segment

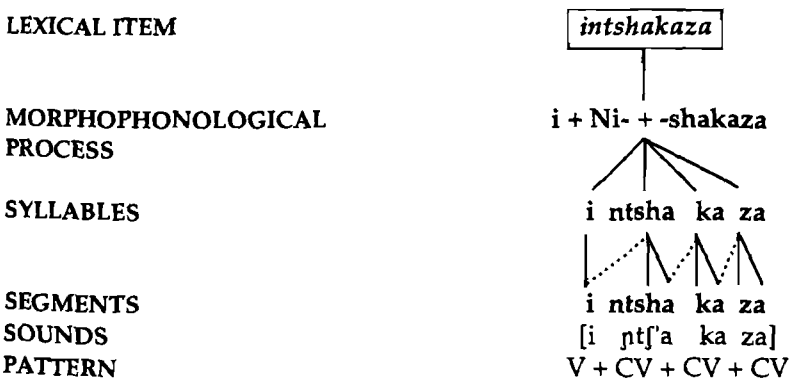
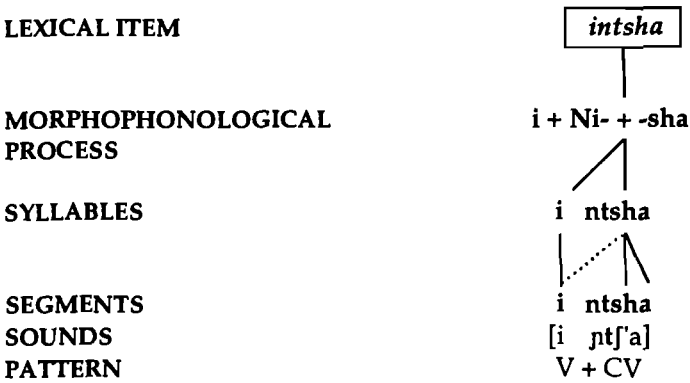


Table 11(b): Form of a syllable and a segment



DMSV records the word ***intsha*** with lemma **-sha** under the letter **S**. It lists the word ***intshakaza*** under the letter **N** with lemma **-ntshakaza**. ACNK and SLNY lemmatise both words ***intsha*** and ***intshakaza*** under the letter **S** with lemmas **-sha** and **-shakaza** respectively. SLNY lemmatises the word ***intshakaza*** under three different letters: **N** (**-ntshakaza**), **T** (**-tshakaza**) and **S** (**-shakaza**). This inconsistency is well illustrated in Table 10 (a). It is paradoxical that DMSV lists the segment **ntsh** separately as a dictionary entry under the letter **N** whereas it records the word ***intsha*** and many more with the same form in the initial position of the nasal noun-stem, under the letters **S** or **T**. However, Doke et al. (1996: 606) justify their inconsistency by saying:

> **ntsh** [ɲtʃ'] Prepalatal nasal preceding the ejective prepalatal affricate, generally the result of homorganic nasal influence upon ***sh***, but sometimes upon ***tsh***. (For words commencing in ***intsh-*** or ***izintsh-*** not listed under ***-ntsh*** see under ***-sh*** or ***-tsh***.

3.2 Internal cohesion of a segment

The initial nasal consonant segment in various Zulu word categories is also identifiable by form and function. It is observed in the following examples that the morphophonological product in each word category overrides the morphophonological process.

3.2.1 Syntactic application

The sentences in Table 12 illustrate a comment by Nkabinde (1992: 85) that:

> The structure and function of words are closely interrelated. This makes it difficult to treat the morphology of words apart from their syntactic application.

It can be observed that the segment **ntsh** of the core syllable-initial consonant maintained its form as well as its function in each word category as illustrated in Tables 11(a), 11(b) and 12.

Table 12: Syntactic application

(i) **Intshakaza** isiqalile ukushazwa. (Noun):
The flower-tuft has begun to wither.

(ii) **Le ntshakaza** imhlophe qwa. (Demonstrative):
This flower-tuft is snow-white.

(iii) Yiphunga **lentshakaza** leli olichazayo. (Possessive):
It is *the flower-tuft's* smell that you are describing.

(iv) Lo mmbila **awunantshakaza**. (Negative):
This maize stalk *has no flower-tuft.*

(v) Lo mmbila **usunentshakaza**. (Affirmative):
This maize plant *has now grown a flower-tuft.*

(vi) **Yintshakaza** lena obuza ngayo. (Copulative):
It is the flower-tuft that you are inquiring about.

(vii) Izilokazane ziqaqele **entshakazweni**. (Locative):
Insects adhere in a thick mass *around the flower-tuft.*

(viii) Unezinwele ezimayikayika **njengentshakaza**. (Adverb):
You have hair hanging and flapping about *like a flower-tuft.*

(ix) Sengibasathe **ntshakaza** uyamila, lutho. (Vocative):
I have waited for a long time for *(you) the flower-tuft* to grow, but there is no progress.

3.2.2 Positional mobility of a composite segment

The same nasal consonant composite segment can occupy different positions in a lexical item from the Zulu lexicon without losing its form and phonetic value. It can be placed in the initial syllable, medial syllable or final syllable positions in lexical items. The common practice of depleting a lemma's actual initial nasal consonant from its combination with the nasal consonant composite segment in its initial position in the first syllable of the noun-stem is discouraged in this article. This is demonstrated in Table 13.

Table 13: Positional mobility of a composite segment

Initial syllable:	**ntsh**	>	[ɲtʃ']
e.g. *i***ntsh***akaza; i***ntsh***a; i***ntsh***ibongo*			
Medial syllable:	**ntsh**	>	[ɲtʃ']
e.g. *ipe***ntsh***isi; isanku***ntsh***ane*			

Final syllable: **ntsh** > [ntʃ']
e.g. *iwoli***ntshi**; *inkwa***ntshu**; *impo***ntshi**

3.2.2.1 Reduplication

In the ordinary writing system, the first and second parts of reduplicated stems in a lexical item are joined to form a single lexical item which is written according to the current Zulu morphophonological form. But, when some dictionary makers lemmatise, they still change the first part of such a lexical item to reflect its form at premorphophonological level. This can be illustrated by comparing ACNK, DMSV and SLNY in Table 13(a). It can be observed that this does not greatly affect the lexical items which are categorised under primary nasal consonant segments. DMSV and SLNY demonstrate a uniform approach under this category e.g. ***inamfunamfu***. Dictionary makers tend to differ when lemmatising the lexical items with duplicated stems under the secondary nasal composite consonant segments, e.g. ***inkelenkele***. The postulated initial homorganic nasal composite segment for the noun ***inkelenkele*** is **nk**. For instance:

— ACNK records this noun under the letter K as **-KHELENKELE** and under H as **-HELE(N)HELE**;
— DMSV records it under the letter H as **-hele(n)kele** and under N as **-nkelenkele**; and
— SLNY lists it under the letter H as **-hele(n)hele** and under N as **-nkelenkele**.

In Table 13(a), the initial segments which contrast with the postulated initial segment in each lexical item, are highlighted under each author.

Table 13(a): Reduplicated noun-stems

Lexicon with Homorganic Nasal Initial Consonant Element	Postulated Homorganic Nasal Consonant Segment	Letters under which Lemmatised by		
		ACNK	DMSV	SLNY
inamfunamfu	–	–	N	N
imbidlimbidli	**mb**	–/–/–	–/–/**Bh**	Mb/–/**Bh**
impithimpithi	**mp**	–/–/**Ph**	–/–/**Ph**	–/–/–
indaxandaxa	**nd**	–/**D**/–	Nd/**D**	Nd/**D**/–
inzuthunzuthu	**nz**	–/–/–	Nz/**Z**	Nz/**Z**/–
intshiphantshipha	**ntsh**	–/–/–	Ntsh/–/**Sh**	Ntsh/–/**Sh**
inkavunkavu	**nk**	–/–/**Kh**	Nk/–/–	Nk/–/**Kh**
inkelenkele	**nk**	–/**Kh**/**H**	Nk/–/**H**	Nk/–/**H**
inkinyankinya	**nk**	–/–/**Kh**	Nk/–/–	Nk/–/**Kh**
inkumunkumu	**nk**	–/–/**Kh**	Nk/–/**Kh**	Nk/–/**Kh**
inxakanxaka	**nx**	–/–/–	–/**X**/–	Nx/**X**/–

3.2.3 Internal stability of a segment

The ordering of consonant items within the same nasal consonant composite segment of a syllable is fixed and noncontrastive. Hence it is not feasible to rearrange the same items and still arrive at the same or at the canonical nasal consonant composite segment. The constraint is ascribed to the rules of generative linguistics and incompatible sound laws which determine the acceptability of the morphophonological product, e.g.

(i) N (tsh, sh) = > **ntsh**, but
(ii) N (t, ts, th, ths, st, sth, ht, hs, hts) ≠ > **ntsh**.

Taylor (1991: 130-131), writing about the possibility of combining words into phrases, says that it is "a question of the compatibility of the feature specifications of the component forms, compatibility being in terms of selection restrictions". The acceptability of word combinations is clear: "either the feature specifications are compatible, or they are not". The same applies to the possibility of combining letters into segments, and the acceptability of these segments, in this case consonant combinations. The combination in example (ii) above does not satisfy the conditions of Zulu segment structures. The pattern of consonant configuration that results after operation in each step obeys neither the distributive law nor satisfies laws regulating the compatibility of sounds in Zulu. Therefore the operation in example (ii) cannot be acceptable in Zulu.

3.2.4 Evolutionary morphophonological product

The development of language orthography usually responds to current language demands. Logically, the current orthography of a lexical item normally indicates the first letter in the initial segment of the first syllable of the noun-stem under which the lexical item is listed in Zulu. Most dictionary-makers indicate this by a capital letter and precede it by a hyphen. When Theunissen (1943: 83) illustrates the use of capitals for the initial letter of the stem of words, he says:

> It should be noted that some Class : 5 (Doke) nouns with Class : 3 (Doke) plurals, have a *different stem* in the singular than in the plural, e.g. ***iNkosi*** but ***amaKhosi, iNdodakazi*** but ***amaDodakazi.***

The nouns ***iNkosi, iNdodakazi*** and ***iNtshakaza*** belong to the same noun class. This article asserts that they are, in principle, recorded under the same letter, i.e. N. In practice, however, this is not so. ACNK, DMSV and SLNY lemmatise these nouns differently. They seem to have considered the evolutionary morphophonological process more important than the lexical product itself. To prove that a lexicographical consistency exists instead of this practised inconsistency, this article provides a column for postulated initial consonant segments

in Tables 9(a), 10(a) and 13(a) that shows uniformity in the homorganic nasal composite consonant segments, viz. **nk**, **nd** and **ntsh** respectively, for these nouns.

Table 14: Evolutionary morphophonological product

(a)

	NOUN	LOCATIVE
EVOLUTIONARY	**i + (Ni- + -shakaza)**	**e + (Ni- + -shakaza) + -ini**
CURRENT FORM	**i + -Ntshakaza**	**e + -Ntshakazeni**
CONTEXTUAL FORM	***intshakaza***	***entshakazeni***
INITIAL SEGMENT	**ntsh**	**ntsh**
SOUND	[nʧʻ]	[nʧʻ]

(b)

	NOUN	LOCATIVE
EVOLUTIONARY	**i + (Ni- + -dodakazi)**	**e + (Ni- + -dodakazi) + -ini**
CURRENT FORM	**i + -Ndodakazi**	**e + -Ndodakazini**
CONTEXTUAL FORM	***indodakazi***	***endodakazini***
INITIAL SEGMENT	**nd**	**nd**
SOUND	[nd]	[nd]

4. Conclusion

It is difficult to engage in lemmatising in Zulu without adopting the word-stem tradition. The method for selecting the letter under which the lexical item, the nasal noun in particular, is to be recorded or looked up in the dictionary, remains controversial. Hence inconsistency, uncertainty, lack of user-friendliness, linguistic assumptions and uneconomical methods of compiling are all unavoidable. This article therefore proposes that the nasal consonant at the initial segment of the nasal noun-stem not be depleted from its homorganic composite nasal consonant segment when a nasal noun is lemmatised. All lexical items used as examples are listed in existing dictionaries, with some showing segments that are contrary to the segments shown in the column for the postulated initial homorganic nasal consonants.

Glossary of Zulu words in the text

amadodakazi	daughters
amakhosi	chiefs
imamba	mamba (*Dendroaspis* spp.)
imbaba	the palm of the hand
imbala (class 4)	colours
imbala (class 9)	actuality, reality; fire spot on the leg
imbidlimbidli	heavy, stout person
imbiza	earthenware pot
imbuthuma	fire of glowing embers; large log fire
imeli	mare
imenenja	manager
imfe	sweet reed
imfene	baboon
imfula	rivers
imfumba	pile of goods
imini	middle of the day
impandla	baldness; bald-headed person
impithimpithi	confusion, commotion; confused affair
impoko	grass flower
impontshi	pouch; small skin bag
imvakazi	veil; hair-fringe hanging over the forehead (of a young woman)
imvithi (class 4)	species of large, shady tree
imvithi (class 9)	wreckage; heap of ruins
imvula	rain
inamfunamfu	sticky substance
indaxandaxa	person or thing dripping wet; lazy person; coward
indiki	person suffering from an hysterical disease
indikimba	subject; main facts, essential points
indinganiso	measure, standard for guidance
indodakazi	daughter
inembe	the last matter passed at confinement; medicine used to aid parturition
ingaco	cultivated field
ingobo	small stomach of a beast
ingulube	pig
inhliziyo	heart
inhlwa	flying termite
inja	dog
injobo	a strip of wild-cat's skin forming the loin-covering of a man
inkambiso	custom
inkankane	hadeda ibis (*Bostrychia hagedash*)
inkantolo	magistrate's court; charge office
inkavunkavu	course, husky kind of food

inkelenkele	dizziness; calamity
inkinyankinya	difficulty; difficult problem; difficult job; quandary
inkosi	chief
inkumunkumu	anything gritty, sandy
inkwantshu	cramp; contraction of muscles
inomfi	birdlime
inono	neat person
insada	abundance; large quantity; large number
insephe	springbuck
insumpa	wart
intezazane	maiden
intombi	girl of marriageable age
intsha	youth (young people)
intshakaza	flower-tuft
intshibongo	person with a long face, the lower part and frontal bone curved
intshiphantshipha	shy, retiring, evasive person
intshontsho	small piece of meat taken as a perquisite
inuku	untidy person
inunu	monster
inyama	meat
inyanga	moon
inyoka	snake
inxakanxaka	confusion, disorder; complicated structure or mechanism
inzuthunzuthu	supple, pliant object; calm, pleasant weather; limp, weak person or animal
ipentshisi	peach
isankuntshane	the adder-tongue fern (*Ophioglossum reticulatum*)
iwolintshi	an orange
izimini	some other times
izimonya	pythons
izimvula	rainfalls
izine	fours
izingulube	pigs
izinhlwa	flying termites
izintshakaza	flower-tufts
izinyawo	feet
umbala	colour
umbalane	yelloweyed canary (*Serinus mozambicus*)
umbekle	Cape robin (*Cossypha caffra*)
umbethe	dew
umbimbi	conspiracy
umonya	python
undanda	a tall person
unjongwe	acidity
unyawo	foot

Bibliography

Abercrombie, D. 1980. *Elements of General Phonetics.* Edinburgh: University Press.

Bauer, L. 1988. *Introducing Linguistic Morphology.* Edinburgh: University Press.

Clements, G.N. and S.J. Keyser. 1983. *CV Phonology: A Generative Theory of the Syllable.* Cambridge: The MIT Press.

Döhne, J.L. 1857. *Zulu–Kafir Dictionary.* Cape Town.

Doke, C.M., D.M. Malcolm, J.M.A. Sikakana and B.W. Vilakazi. 1990. *English–Zulu/Zulu–English Dictionary.* Johannesburg: Witwatersrand University Press.

Hartmann, R.R.K. 1983. *Lexicography: Principles and Practice.* London: Academic Press.

Hlongwane, J.B. 1995. Growth of the Zulu Language and its Structural Changes. *South African Journal of African Languages* 15(2): 60-65.

Katamba, F. 1989. *An Introduction to Phonology.* London: Longman.

Lyons, J. 1971. *Introduction to Theoretical Linguistics.* Cambridge: University Press.

Malmkjaer, K. 1991. *The Linguistics Encyclopedia.* London: Routledge.

Marggraff, M. 1997. *Aspects of Lemmatization in Nguni.* Unpublished paper presented at the Second International Conference of the African Association for Lexicography, held at the University of Natal, Durban, 14-16 July 1997.

Meinhof, C. 1932. *Introduction to the Phonology of the Bantu Languages.* Berlin: Dietrich Reimer.

Mini, B.M. 1992. Problems in Lexicographical Work in the Xhosa Dictionary Project. *General and Technical Lexicography in Practice.* Pretoria: National Terminology Services.

Mini, B.M. 1995. Lexicographical Problems in isiXhosa. *Lexikos* 5: 40-56.

Mzolo, D. 1968. The Zulu Noun without the Initial Vowel. *African Studies* 27(4): 195-210.

Nkabinde, A.C. 1975. *A Revision of the Word Categories in Zulu.* Unpublished D.Litt et Phil. dissertation. Pretoria: University of South Africa.

Nkabinde, A.C. 1985. *Isichazamazwi* 2. Cape Town: Oxford University Press.

Nkabinde, A.C. 1992. Lexicography in Zulu. *General and Technical Lexicography in Practice.* Pretoria: National Terminology Services.

Nyembezi, S.L. 1992. *aZ Isichazimazwi Sanamuhla Nangomuso.* Pietermaritzburg: Reach Out.

Posthumus, L.C. 1994: Word-based versus Root-based Morphology. *South African Journal of African Languages* 14(1): 28-36.

Taylor, J.R. 1991. *Linguistic Categorization: Prototypes in Linguistic Theory.* Oxford: Clarendon Press.

Theunissen, S.B. 1943. Zulu Orthography. *Native Teachers' Journal* 23: 81-83.

Van Wyk, E.B. 1995. Linguistic Assumptions and Lexicographical Traditions in the African Languages. *Lexikos* 5: 82-96.

Zgusta, L. 1971. *Manual of Lexicography.* Prague: Academia/The Hague: Mouton.

Ziervogel, D. 1986. *Speech Sounds and Sound Changes of the Bantu Languages of South Africa.* Pretoria: Unisa.

A Multilingual, Multicultural and Explanatory Music Education Dictionary for South Africa — Using Wiegand's Metalexicography to Establish its Purposes, Functions and Nature*

Maria Smit, *Department of Music, University of Stellenbosch, South Africa*

Abstract: Wiegand's metalexicography is used to establish the purposes, functions and nature of a multilingual, multicultural, and explanatory music education dictionary for South Africa. Specific types of dictionaries have specific purposes. Special-field dictionaries should fulfil the purpose of conveying information on knowledge in special fields. They should also solve communication conflicts. The genuine purposes of special-field dictionaries, according to Wiegand, are to convey either linguistic information on terms, or encyclopedic information, or both. The needs of users should be taken into account when determining the functions of a dictionary. When the functions of a dictionary containing music terms from South Africa is considered, social factors in South African music education also have to be taken into account. The planned dictionary will have a linguistic and a communicative function. It will also have a cognitive and scientific function, fulfilling an educational need. With regard to the nature of the planned dictionary, it will have to contain elements of different types of dictionaries, such as explanatory dictionaries, translation dictionaries, and learner's dictionaries. A thematic arrangement will be followed, supplemented by an alphabetical index. Two versions of the dictionary will have to be published, namely, a more scholarly version for specialists, with more types of information, as well as a more popular version for nonspecialists.

Keywords: MULTICULTURAL DICTIONARY, MULTILINGUAL DICTIONARY, EXPLANATORY DICTIONARY, SPECIAL-FIELD DICTIONARY, WIEGAND, PURPOSES OF DICTIONARIES, FUNCTIONS OF DICTIONARIES, ENCYCLOPEDIC INFORMATION, LINGUISTIC INFORMATION, DICTIONARY USE, PROTOCOLS, LEARNER'S LEXICOGRAPHY

Opsomming: 'n Veeltalige, multikulturele en verklarende opvoedkundige musiekwoordeboek vir Suid-Afrika. Wiegand se metaleksikografie word gebruik om die doelstellings, funksies en aard van 'n veeltalige, multikulturele en verklarende opvoedkundige

* This paper was presented at the Second International Conference of the African Association for Lexicography, held at the University of Natal, Durban, 14-16 July 1997.

musiekwoordeboek vir Suid-Afrika te bepaal. Spesifieke tipes woordeboeke het spesifieke doelstellings. Vakwoordeboeke behoort die doelstelling te hê om inligting oor kennis in vakgebiede oor te dra. Hulle behoort ook kommunikasiekonflikte op te los. Die ware doelstellings van vakwoordeboeke is, volgens Wiegand, om óf linguistiese inligting oor terme oor te dra, óf ensiklopediese inligting, of albei. Die behoeftes van die gebruikers moet in ag geneem word wanneer die funksies van 'n woordeboek bepaal word. Wanneer nagedink word oor die funksies van 'n woordeboek wat musiekterme uit Suid-Afrika bevat, moet ook sosiale faktore in die Suid-Afrikaanse musiekopvoeding in ag geneem word. Die beplande woordeboek sal 'n linguistiese en 'n kommunikatiewe funksie hê. Dit sal ook 'n kognitiewe en wetenskaplike funksie hê om in 'n opvoedkundige behoefte te voorsien. Wat die aard van die woordeboek betref, sal dit elemente van verskillende tipes woordeboeke moet bevat, soos verklarende woordeboeke, vertalende woordeboeke en aanleerderswoordeboeke. 'n Tematiese rangskikking sal gevolg word, aangevul deur 'n alfabetiese indeks. Twee weergawes van die woordeboek sal gepubliseer moet word, naamlik 'n meer akademiese weergawe vir spesialiste, met meer tipes inligting, en 'n meer populêre weergawe vir niespesialiste.

Sleutelwoorde: MULTIKULTURELE WOORDEBOEK, VEELTALIGE WOORDEBOEK, VERKLARENDE WOORDEBOEK, VAKWOORDEBOEK, WIEGAND, DOELSTELLINGS VAN WOORDEBOEKE, FUNKSIES VAN WOORDEBOEKE, ENSIKLOPEDIESE INLIGTING, LINGUISTIESE INLIGTING, WOORDEBOEKGEBRUIK, PROTOKOLLE, AANLEERDERSLEKSIKOGRAFIE

In this paper, a brief motivation for the use of H.E. Wiegand's theory of metalexicography in conceptualising an explanatory music dictionary containing terms from all music cultures in South Africa, will be offered.

Wiegand (1984: 559) distinguishes four main research fields within metalexicography. These are: (i) research on the use of dictionaries ("Wörterbuchbenutzungsforschung"), (ii) research into criticism of dictionaries ("kritische Wörterbuchforschung"), (iii) the study of the history of dictionaries ("historische Wörterbuchforschung"), and (iv) the general theory of dictionaries ("systematische Wörterbuchforschung"). Within the general theory of dictionaries the following constituent theories are distinguished by Wiegand (1983: 44): (i) a general section, (ii) an organisational section, (iii) a theory of lexicographical language research, and (iv) a theory of lexicographical language description. Although Wiegand discusses them separately, these components form part of the whole theory, and should all be taken into account in the planning of any dictionary.

1. The purposes of the planned dictionary

Those aspects of Wiegand's metalexicography which can help lexicographers establish the most important purposes and functions of dictionaries are mainly drawn from the constituent theories which deal with research into dictionary use, and the general section of the general theory of dictionaries. Wiegand

(1983a: 314-315) has several practical suggestions for compilers who have to establish the purposes of their dictionaries. For example, one should incorporate the planned dictionary into the "dictionary scene" (Wiegand 1983a: 314) as a whole. The purposes should be formulated to fulfil those needs which can be identified by studying existing dictionaries of the same type. Even such dictionaries which are in their planning or compilation phase should be taken into account. Not only should one look at the collection of lemmata, but also at the contents of the articles of these dictionaries. From this type of research, one can learn a great deal about the advantages and disadvantages of existing and future dictionaries, and try to improve on them. This practice links with the second and third research areas of Wiegand's theory, namely, research into dictionary criticism and into the history of dictionaries.

Specific types of dictionaries have specific purposes. Wiegand (1976: 118) illustrates this assumption by characterising the explanation of word meanings as the "historically constant purpose" of monolingual dictionaries. Bilingual dictionaries, on the other hand, have other purposes. To illustrate this, Zaiping and Wiegand (1987: 229) discuss the purposes of the proposed comprehensive German-Chinese dictionary. They claim that there are scientific, cultural and political reasons why such a dictionary is necessary.[1] Special-field dictionaries ("Fachwörterbücher") have still other purposes. According to Wiegand (1993: 2), one could assume that this type of dictionary should convey information on knowledge in special fields.

Wiegand (1979: 49) also explains that special-field dictionaries are meant to solve communication conflicts which arise particularly when people read texts. Terminologies, he continues (1979: 50), have as main purpose the prevention of communication conflicts with regard to special-field languages.

Wiegand (1988: 743) even suggests that one should further classify the different types of special-field dictionaries according to their genuine purposes ("genuine Zwecke"). This is because some special-field dictionaries tend to give information which is of a more linguistic nature, while others tend to give encyclopedic information, i.e. of an extralinguistic nature. There are even those which equally give both types of information (Wiegand 1988: 747). In accordance with the genuine purposes of a particular special-field dictionary, compilers should decide whether the intended dictionary should convey linguistic information, encyclopedic information, or both. This decision will influence the methods the compilers will follow during the compilation process (Wiegand 1988: 751).

The needs of users and the way in which they will use the planned dictionary should also be taken into account. If one wants to plan the dictionary from a user's perspective, one should incorporate Wiegand's suggestions on typical user needs. In his initial phase of research on dictionary use, Wiegand (1977, 1977a: 57ff) suggests that empirical research should be done in order to determine more exactly what users need from dictionaries. Wiegand (1987: 179) explains that the whole reason for empirical research on dictionary use is to

ensure that new dictionaries, or revisions of older ones, will be more effective. One can measure the effectiveness of dictionaries according to the occurrence of successful consultations by users. Questionnaires which can reveal information about the different situations of dictionary use, should be drawn up. The questions may vary from those about which people actually own dictionaries and which dictionaries they own, to those about the type of information users need — whether they need to know something about the language (i.e. semantic knowledge) or something about the thing to which the word refers (i.e. encyclopedic knowledge).

This process could lead to the formation of a typology of situations of dictionary use. Wiegand (1977a: 70) distinguishes between (i) situations of passive language use (i.e. when users experience problems when they read texts), (ii) situations of active language use (i.e. when users experience problems when they have to write texts), and (iii) other situations of dictionary use (which are not relevant here).

In a later phase of his research, Wiegand collaborated with Ripfel (Ripfel and Wiegand 1986) to obtain preliminary information for the formulation of a theory of learner's lexicography. They argued (1986: 492) that a draft for a medium-sized, one-volume German learner's dictionary could be designed from such a theory. They started out their investigation by assessing different methods used by researchers up to that stage. After having tried to classify and critically discuss the methods and results[2], they came to the conclusion that a suitable way of conducting empirical research with regard to dictionary use is the utilisation of protocols. Protocols are utilised when subjects perform certain tasks in which dictionaries are used, at the same time commenting on these tasks by giving information on the type of search questions they had, how they went about to solve their search problems and in which ways the dictionaries used were adequate to solve these problems.

Lexicographers should firstly formulate the purposes of dictionaries in general terms. Secondly they should classify them into groups in such a way that they can derive specific and concrete lexicographical purposes for each dictionary type differentiated by Wiegand's theory of dictionary typology. Such purposes are then set out in the general section of the dictionary plan.[3] This plan is dealt with by Wiegand in his theory about the activities involved when establishing a dictionary basis and processing this basis into a lexicographical file.

As regards the music dictionary planned, at this stage the following can be said about its general purposes: (i) it will fulfil an educational need as there are no other dictionaries of this type yet; (ii) it will serve a scientific purpose, because it will provide subject-field information for research and education in ethnomusicology; and (iii) it will further a cultural aim, because it will enhance communication in the field of music.

Music terms are used in several situations. Firstly, for the practical performance of music, students have to understand the meaning of music terms in

order to be able to perform the music correctly. This can be characterised as a situation of passive language use. Problem situations which might occur are ones such as word gaps with respect to simplex, compound or derivative gaps, word meaning gaps, word usage gaps, word discrimination gaps, etc.

Secondly, in writing academic dissertations or papers on musical aspects, activities and instruments, students do not only have to understand the meaning of terms (as in passive language use), but they also have to be able to use the terms in their writing (as in active language use). In the process of writing dissertations or papers, students will use literature in which unknown terms occur. They then need the "lexical-semantic knowledge" which is needed in active language use (Wiegand 1977a: 78-80) to "place" a particular term within the relevant semantic "field" or "network." Terms have to be used in a functional and communicative way in this type of language use.

Sometimes problematic situations within African music terminology may also occur because students do not know the morphological structure of the language from which the term comes. The noun denoting a song type such as *umtshotsho* is derived from *uku-thsotsha* which is a verb. Users will have to know where to search for these terms, whether at the normal alphabetical place, or at the place where the stem of the word is inserted alphabetically. This problem links with a general problem nonnative speakers of African languages experience when they use dictionaries in African languages.

Research into dictionary use done by Wiegand, as well as that done by others such as Tono, Bensoussan et al. and Mitchell, mainly concern the use of general dictionaries. For special-field lexicography, however, similar tasks to write protocols to describe their use of dictionaries may be given to subjects. From the tasks set by Wiegand (1985) to the subjects in his research project, it can be concluded that there are no dictionaries available yet which give an extensive treatment of the different kinds of African music terminology.[4] This also implies that it is impossible to compare different types of dictionaries with each other. Lists containing tentative dictionary articles will have to be compiled especially for the purpose of writing protocols, using information in textbooks, journal articles and dissertations. Only then can students be given excerpts from texts in which they may find unknown words which they may want to look up. The texts with which music students will work, will also differ from the one suggested by Wiegand. This is because the texts will not necessarily be translated into another language. It will entail a situation of passive as well as active language use, which means that the students will have to understand what they read and will have to be able to use the information obtained in a paper on the particular topic at hand. Music terminology will not necessarily be translated, because loanwords will be used in, for example, an English text. The meanings and especially the cultural contexts in which these music terms are used, however, are important.

Subjects will have to write down the problems they experienced, and the strategies they used to solve these problems. They could indicate which parts

of the given "dictionaries" or word lists helped them most in solving problems, in other words, which types of information were most useful.

Only then will researchers be in a position to analyse the answers to these protocols and determine which types of information will be needed in the articles of a music dictionary of the type planned. One could divide the different problem situations into different categories, for example the problems with regard to word gaps, usage gaps, word discrimination gaps, derivation gaps, etc. The possible relationship between the questions of the users and the actual dictionary use can also be determined. Hypotheses could then be formed, which could be tested in later research. Next one could draw up a theoretical framework (or frame, to use Konerding's (1993) approach), as Wiegand suggests, on the basis of which an appropriate dictionary article can be planned. Purposes for this type of dictionary which can finally be formulated, might be to solve search problems with regard to word gaps, usage gaps, word discrimination gaps, derivation gaps, etc.

2. The functions of the planned dictionary

When one looks at the purposes of the planned dictionary from the perspective of a social, cultural and political point of view, as Zaiping and Wiegand (1987) do, there are still some other factors to keep in mind. These are, for example, communicative needs, cognitive needs, cultural needs and scientific needs of the potential users. If these needs are considered, one will also be able to determine the functions which such a dictionary will have in the society. For example, some of the social factors in South African music education up to the present need to be taken into consideration. Traditionally, only Western music has been promoted in schools. With the new curricula, this situation has changed. A music dictionary is needed which could fulfil some of the historical, social and cultural needs of the South African society as a whole. African music terms should be recorded in an appropriate way in order to prevent the indigenous cultures from disappearing in a society where all aspects of life have become increasingly westernised.

The main function of the proposed dictionary will consequently have to be to adequately explain the meanings and use of African music terms. In this sense, it will have a linguistic and communicative function. A dictionary which could solve communication conflicts in South African music education is urgently needed. No dictionaries are currently available to students who need to study different kinds of African music. The dictionary should take into account the etic/emic debate in ethnomusicology. Meaning explanations and types of information should be presented in such a way that students from other cultures may understand the semantic and cultural contexts in which expressions are used.

The planned dictionary will have to fulfil a cognitive and scientific function by providing the possibility for research within the field of music, ethno-

musicology and anthropology.

In addition, a dictionary such as the proposed one should also fulfil an educational need by empowering students to use dictionaries and to interpret the subject matter. One should therefore pay attention to the reference skills of the potential users. These skills, or the absence of such skills, will also have an influence on the eventual access structure of the dictionary. Less educated users will need more guidance than specialists to find the information they want. Wiegand emphasises the usefulness of different access structures. His suggestions in this regard are important for making dictionaries accessible to different users in different ways.[5] This aspect links with the function of empowerment that a dictionary of the proposed type will have. Various innovative access structures will therefore have to be devised for users with different levels of reference skills and different search priorities. Especially in a computerised version of such a dictionary, one should explore the different possibilities of access structures when multi-media or hypertext is used.

3. The nature of the planned dictionary

The purposes of a dictionary such as the planned music dictionary have a definite influence on the types of information which will have to be included. For various reasons, it is foreseen that the planned dictionary will have to be a typological hybrid. It will have to possess characteristics from different dictionary types in order to fulfil the needs of the potential users. It should not only be a special-field dictionary in the traditional sense. For example, it will have to incorporate elements of explanatory dictionaries, because encyclopedic as well as meaning explanations will have to be included. It will in addition have to be translatory, because terms from African musical cultures might have translation equivalents in other languages. Elements of learner's lexicography will also have to be incorporated, because students have to "acquire" a "special-field language", namely musical terms from different indigenous cultures. Furthermore, it is believed that such a dictionary would do well to follow a thematic arrangement instead of an alphabetical one, because concepts could then be studied within their natural contexts.

In terms of Wiegand's metalexicography, it is necessary to distinguish between different types of dictionary use and between different types of users. Firstly, students in secondary education, undergraduates as well as educated nonspecialists who want to acquire knowledge about the music cultures of South Africa, will be one of the target groups. A music dictionary should be able to provide suitable information which these students could use in their musicological and anthropological studies. Secondly, the dictionary will also have to be published in a more scholarly version for the sake of specialists. This version of the dictionary will have to supply enough information to enhance and encourage scientific research within ethnomusicology. References should be made to relevant literature to make it possible for researchers to obtain additional information when needed. It should, therefore, be possible to distinguish

between a more popular version of the dictionary for the purposes of nonspecialists and a more academic version for the purposes of specialists. In a computerised version of the planned dictionary, the use of video clippings, sound recordings and "lexicographical narration", as Wiegand (1977: 107) calls it, have interesting possibilities when they can be used within the frames of the dictionary articles at the specific places where they are needed.

4. Summary

Wiegand's theoretical framework is the only one which deals extensively with all aspects of dictionary planning and making. It is clear that, for the purposes of the planned music dictionary, Wiegand's framework is certainly useful. This does not mean that one can use it without any adaptations. The theory should be used as a basis for determining what is needed to compile a customised dictionary which will fulfil the needs of the potential users. Wiegand has for example not dealt with all aspects to the same extent. Certain aspects, such as communication in special fields, user situations in the case of special-field dictionaries, the latest developments in computerisation, aspects of corpus lexicography, etc., will have to be worked out in more detail before a music dictionary of the kind proposed here can be compiled.

It is, however, of crucial importance that aspiring lexicographers pass through all the phases of Wiegand's theory in order to lay a sound theoretical basis for the planning and compilation of their dictionaries. In the case of the proposed music dictionary which will include terms from African and possibly also Indian and Western music, this is also valid. No final decisions or conclusions can be made or drawn with regard to any of the mentioned aspects before one has not completed all the steps in Wiegand's theoretical framework.

Notes

1. For example, such a dictionary would enhance communication between Germans and Chinese who share many scientific and technical projects. There are a few existing dictionaries, but they do not contain recent lexical items. In determining the purposes of this German-Chinese dictionary, the compilers took into consideration the needs of the users. In the first place the dictionary is meant for Chinese users who know German well and need the dictionary in university research and lectures. In the second place, German users may also find it a useful dictionary. The dictionary also has several functions: Firstly, it fulfils the needs of Chinese speakers who have to translate from German to Chinese ("Herübersetzen.") Secondly, a Chinese speaker who wants to produce German texts ("Hinproduktion") may also find the dictionary useful. For German users, it may be helpful in situations of producing Chinese texts in the first place, and in the second place it helps them with translating from Chinese to German.

   Cf. also Kromann et al. (1991: 2712-2713) who list a few purposes of bilingual dictionaries. For example, they claim that bilingual dictionaries can serve as "important tools in language learning". Furthermore, they are "useful aids to travel abroad and communication in foreign

languages, necessary tools in the commercial world and public administration, and indispensable for secretaries dealing with foreign-language correspondence, translators and interpreters." They (1991: 2712) also refer to "specialized translation dictionaries", which they consider important in international specialised communication between companies, public authorities and international organisations. Then there are also "translation dictionaries of the scholarly historical-philological type, which serve research in the humanities and the interpretation of older texts and cultures", for example, translation dictionaries with Biblical Hebrew, Classical Greek and Latin as the source languages.

2. The research by Ard (1982), Hatherall (1984), Bensoussan et al. (1984), Mitchell (1983), Tono (1984) and Wiegand (1985) was taken into account here.
3. Wiegand (1983: 58, note 35) states that one can list the purposes of a dictionary in the dictionary introduction by characterising in an organised way the situations of use for the dictionary type at hand. One could regard situations of dictionary use as a triangle of (i) dictionary user, (ii) the question which the user directs towards the dictionary, and (iii) and the dictionary itself.
4. One exception is within the field of musical instruments, where the *New Grove Dictionary of Musical Instruments* (Sadie:1980) includes the terms for African musical instruments. Other publications which also deal with musical instruments are e.g. Sachs (1962), Brincard (1989), Wegner (1984) and Norborg (1987). For other types of terms such as those designating musical activities and musical forms, there are no lexicographical publications in which these terms can be looked up.
5. Cf. also McArthur (1986: 178-179), who stresses the importance of working from a menu (as on the computer), via "an alphabetic or thematic-and-numerical indexing system", and operating "according to our own system of priority". He (1986: 183) believes that this helps in "democratizing" the handling of the information.

References

Ard, J. 1982. The Use of Bilingual Dictionaries by ESL Students while Writing. *ITL Review of Applied Linguistics* 55: 1-27.

Bensoussan, M. et al. 1984. The Effect of Dictionary Use on EFL Text Performance Compared with Student and Teacher Attitudes and Expectations. *Reading in a Foreign Language* 2: 262-276.

Brincard, M-T. 1989. *Sounding Forms: African Musical Instruments.* New York: The American Federation of Arts.

Hatherall, G. 1984. Studying Dictionary Use: Some Findings and Proposals. Hartmann, R.R.K. (Ed.). 1984. *LEXeter '83 Proceedings. Papers from the International Conference on Lexicography at Exeter, 9-12 September 1983*: 183-189. Tübingen: Max Niemeyer.

Konerding, H-P. 1993. *Frames und lexikalisches Bedeutungswissen. Untersuchungen zur linguistischen Grundlegung einer Frametheorie und zu ihrer Anwendung in der Lexikographie.* Reihe Germanistische Linguistik 142. Tübingen: Max Niemeyer.

Kromann, H-P. et al. 1991. Principles of Bilingual Lexicography. Hausmann, F.J. et al. (Ed.). 1991. *Wörterbücher. Ein internationales Handbuch zur Lexikographie / Dictionaries. An International Encyclopedia of Lexicography / Dictionnaires. Encyclopédie internationale de lexicographie*: 2711-2728. Handbücher zur Sprach- und Kommunikationswissenschaft 5.2. Berlin: De Gruyter.

McArthur, T. 1986. *Worlds of Reference.* Cambridge: Cambridge University Press.

Mitchell, E. 1983. *Search-do Reading: Difficulties in Using a Dictionary.* Formative Assessment of Reading, Working Paper 21. Aberdeen: College of Education.

Norborg, A. 1987. *A Handbook of Musical and Other Sound-Producing Instruments from Namibia and Botswana.* Musikmuseetsskrifter 13. Stockholm: [A. Norborg].

Ripfel, M. and H.E. Wiegand. 1986. Wörterbuchbenutzungsforschung. Ein kritischer Bericht. *Germanistische Linguistik: Studien zur neuhochdeutschen Lexikographie* 6: 490-520.

Sachs, C. 1962. *Real-Lexikon der Musikinstrumente.* Hildesheim: Olms.

Sadie, S. (Ed.). 1980. *The New Grove Dictionary of Music and Musicians.* London: MacMillan.

Tono, Y. 1984. *On the Dictionary User's Reference Skills.* Unpublished B.Ed. thesis. Tokyo Gakugei University.

Wegner, U. 1984. *Afrikanische Saiteninstrumente.* Berlin: Museum für Völkerkunde.

Wiegand, H.E. 1976. Synonymie und ihre Bedeutung in der einsprachigen Lexikographie. Moser, H. (Ed.). 1976. *Probleme der Lexikologie und Lexikographie: Jahrbuch 1975 des Instituts für deutsche Sprache*: 118-180. Düsseldorf: Pädagogischer Verlag Schwann.

Wiegand, H.E. 1977. Einige grundlegende semantisch-pragmatische Aspekte von Wörterbucheinträgen. Hyldgaard-Jensen, K. (Ed.). 1977. *Kopenhagener Beiträge zur germanistischen Linguistik*: 59-149. Copenhagen: Universitetsforlaget.

Wiegand, H.E. 1977a. Nachdenken über Wörterbücher: Aktuelle Probleme. Drosdowki, G., H. Henne and H.E. Wiegand. 1977. *Nachdenken über Wörterbücher*: 51-102. Mannheim: Bibliographisches Institut.

Wiegand, H.E. 1979. Kommunikationskonflikte und Fachsprachengebrauch. Mentrup, W. (Ed.). 1979. *Fachsprachen und Gemeinsprache: Jahrbuch 1978 des Instituts für deutsche Sprache*: 25-58. Düsseldorf: Pädagogischer Verlag Schwann.

Wiegand, H.E. 1983. Überlegungen zu einer Theorie der lexikographischen Sprachbeschreibung. *Germanistische Linguistik* 5-6: 35-72.

Wiegand, H.E. 1983a. Zur Geschichte des deutschen Wörterbuchs von Hermann Paul. *Zeitschrift für germanistischen Linguistik* 11: 301-320.

Wiegand, H.E. 1984. Prinzipien und Methoden historischer Lexikographie. Besch, W. et al. (Ed.). 1984. *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*: 557-620. Handbücher zur Sprach- und Kommunikationswissenschaft 2.1. Berlin: De Gruyter.

Wiegand, H.E. 1985. Fragen zur Grammatik in Wörterbuchbenutzungsprotokollen. Ein Beitrag zur empirischen Erforschung der Benutzung einsprachiger Wörterbücher. Bergenholtz, H. and J. Mugdan. 1985. *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch, 28-30.6.1984*: 20-98. Lexicographica Series Maior 3. Tübingen: Max Niemeyer.

Wiegand, H.E. 1987. Zur handlungstheoretischen Grundlegung der Wörterbuchbenutzungsforschung. *Lexicographica* 3: 178-227.

Wiegand, H.E. 1988. Was eigentlich ist Fachlexikographie? Munske, H.H. et al. (Ed.). 1988. *Deutsche Wortschatz: Lexikologische Studien*: 729-790. Berlin: De Gruyter.

Wiegand, H.E. 1993. Zur Unterscheidung von semantischen und enzyklopädischen Daten in Fachwörterbüchern. Unpublished paper.

Zaiping, P. and H.E. Wiegand. 1987. Konzeption für das grosse deutsch-chinesische Wörterbuch (zweiter Entwurf). *Lexicographica* 3: 228-241.

"Oumense het blotvoet gebruik": Nederlandse taalresten in de variëteiten van het Afrikaans*

Karin van Lierop, *VDO Opleidings- en Adviescentrum, Hogeschool van Arnhem en Nijmegen, Nederland*

Abstract: "Oumense het blotvoet gebruik": Dutch relics in the regional varieties of Afrikaans. The Bureau of the WAT has developed a survey of all varieties of Afrikaans. One part thereof is the survey of Dutch language relics in the Afrikaans regional varieties which is discussed here. From earlier surveys it has been shown that many Dutch relics are present; the Dutch language and in particular the dialect from Holland from the 16th and 17th centuries have had a strong influence on (the origin of) Afrikaans. Questionnaires were used to check with 183 informants from the Northern and the Western Cape whether Dutch relics are still present in the varieties they use, on the fonetic as well as the lexical level.

The survey has shown that Dutch relics are found on both levels. In the pronunciation especially there is still [ø] where standard Afrikaans has [e], and less often [y] is found instead of [œy]. It is noticeable that the [əi]/[i] variation is present less frequently, while in Dutch a parallel development in the diphtongisation of [y] and [i] has occurred. Afrikaans has seemingly undergone an independent development in this regard. Most phonetic relics were found in Namaqualand, particularly among informants with little schooling, and somewhat more among women than among men.

On a lexical level the results were quite different. There are still many Dutch relics present in word meaning as well as in knowledge of idioms. The Muslim community of Cape Town scores particularly high in this regard, while phonetic relics were scarcely found. Informants with higher education levels score more highly in this regard than those with lower schooling or less, and among men more lexical relics are recorded than among women.

Lastly, both on a lexical level and on a phonetic level, a number of indirect results were gathered from the survey which also point to the presence of Dutch language relics.

Keywords: WOORDEBOEK VAN DIE AFRIKAANSE TAAL, INVENTORYING OF RELICS, USE OF LANGUAGE DESCRIPTION, QUESTIONNAIRE, INFLUENCE OF DUTCH ON AFRIKAANS, PHONETIC LANGUAGE RELICS, SOUND VARIATION, LEXICAL LANGUAGE RELICS, WORD MEANING, IDIOMS, CORRELATIONS WITH SEX, SCHOOLING AND AGE, INDIRECT RESULTS

Samenvatting: Het Buro van die WAT heeft een onderzoek ontwikkeld naar alle variëteiten van het Afrikaans. Een deelproject hiervan is het hier te bespreken onderzoek naar Nederlandse taalresten in de variëteiten van het Afrikaans. Uit eerdere studies is gebleken dat er nog veel

* Een onderzoeksproject van het Bureau van die WAT in samenwerking met de Nederlandse Taalunie en het Kaaps Forum voor Neerlandistiek.

Nederlandse taalresten aanwezig zijn; de Nederlandse taal en in het bijzonder de Holland-variant uit de 16e en 17e eeuw heeft immers een sterke invloed gehad op (het ontstaan van) het Afrikaans. Via vragenlijsten is er bij 183 informanten uit de Noord-Kaap en de West-Kaap nagegaan of er nog Nederlandse taalresten aanwezig zijn in de variëteit die zij spreken, zowel op fonetisch als op lexicaal gebied.

Gebleken is uit onderzoek dat er op beide gebieden nog Nederlandse resten te vinden zijn. In de spraakklanken blijkt er vooral nog sprake te zijn van de [ø] waar standaard Afrikaans de [e] heeft en, in mindere mate, komt de [y] voor in plaats van de [œy]. Opvallend is dat de [əi]/[i]-wisseling veel minder voorkomt, terwijl er in het Nederlands een parallelle ontwikkeling is geweest in de diftongering van [y] en [i]. Kennelijk heeft het Afrikaans hier een eigen ontwikkeling ondergaan. De meeste fonetische taalresten zijn gevonden in Namaqualand, met name bij informanten met weinig opleiding en iets meer bij vrouwen dan bij mannen.

Op lexicaal gebied ziet het beeld er heel anders uit. Zowel in woordbetekenis als wat betreft de kennis van idiomen blijken er nog veel Nederlandse taalresten aanwezig te zijn. Hier scoort vooral de Moslimgemeenschap in Kaapstad erg hoog, terwijl daar nauwelijks fonetische resten zijn aangetroffen. Informanten met meer schoolopleiding scoren hier hoger dan die met lagere schoolopleiding of minder en bij mannen noteren we meer lexicale resten dan bij vrouwen.

Tot slot zijn er op lexicaal en op fonetisch gebied een aantal indirecte resultaten uit het onderzoek voortgekomen die eveneens wijzen op de aanwezigheid van Nederlandse taalresten.

Trefwoorden: WOORDEBOEK VAN DIE AFRIKAANSE TAAL, INVENTARISATIE VAN RELICTEN, NUT VAN TAALBESCHRIJVING, VRAGENLIJST, INVLOED VAN HET NEDERLANDS OP HET AFRIKAANS, FONETISCHE TAALRESTEN, KLANKWISSELING, LEXICALE TAALRESTEN, WOORDBETEKENIS, IDIOMEN, CORRELATIES MET GESLACHT, OPLEIDING EN LEEFTIJD, INDIRECTE RESULTATEN

1. Inleiding

Als het Afrikaans, zoals de opvatting heerst, geen dialecten kent, dan kent het toch zeker variëteiten, die in veel gevallen spreek- en/of streektaal gebonden zijn. Het *Woordeboek van die Afrikaanse Taal (WAT)* is in het verleden vaak verweten zich in zijn lexicografie uitsluitend te richten op het standaard Afrikaans en te weinig recht te doen aan regionale en sociale variëteiten. In het opnamebeleid echter van het *WAT* spreekt de redactie zich expliciet uit dat de lemmakeuze van het woordenboek een weerspiegeling moet zijn van de lexicale items waaruit de woordenschat van het Afrikaans bestaat en daarmee sluit het *WAT* materiaal in uit alle variëteiten van het Afrikaans, want "die *WAT* is 'n woordeboek vir Afrikaans, nie vir Standaardafrikaans nie" (Inleiding, *WAT* IX, 1994: i). Het *WAT* beschikt dan ook over een uitgebreide verzameling van streektaalmateriaal, regionaal gekleurde woorden en uitdrukkingen, en zoekt al langer naar uitbreiding van haar eigen onderzoek naar de variëteiten van het Afrikaans.

Onder auspiciën van het Kaaps Forum voor Neerlandistiek heeft het WAT een breed onderzoek ontwikkeld naar alle variëteiten van het Afrikaans. Een deelproject daarvan is het hier te bespreken onderzoek naar Nederlandse taal-

resten in de variëteiten van het Afrikaans, een onderzoek naar overblijfselen van archaïsch taalgebruik waarvan de herkomst is terug te voeren op het Nederlands of op Nederlandse dialecten die niet (meer) in het standaard Afrikaans voorkomen.

Uit materiaal van het WAT en van andere studies op dit gebied, is gebleken dat er veel Nederlandse taalresten voorkomen in de variëteiten van het Afrikaans. Vooral vanwege de geografische geïsoleerdheid van de sprekers van die variëteiten zijn zekere Nederlandse resten hierin bewaard gebleven, waar ze niet (meer) in het standaard Afrikaans voorkomen. Met het hier te bespreken project hebben we ons geconcentreerd op die resten, in vijf, geografische tamelijk geïsoleerde, gebieden: Namaqualand, Genadendal, Mamre, Saron en de Moslimgemeenschap in Kaapstad.

Om tot een systematische inventarisatie van deze relicten te komen, heeft het Buro van die WAT contact gezocht met de Nederlandse Taalunie in Den Haag. Onder de koepel van de Taalunie functioneert al enige jaren een overlegstructuur voor dialectlexicografie in Nederland en België, het overleg Regionale Woordenboeken, het ReWo. Twee medewerkers van het ReWo, Joep Kruijsen van de Nijmeegse Centrale voor Dialect- en Naamkunde in Nijmegen en Jacques van Keymeulen van het Seminar voor Nederlandse Taalkunde en Vlaamse Dialectologie in Gent, hebben samen met de hoofdredacteur van het WAT, Dirk van Schalkwyk, het vooronderzoek voor het project "Nederlandse taalresten in de variëteiten van het Afrikaans" in Namaqualand voorbereid en uitgevoerd. Inmiddels is het project afgerond en dit artikel biedt een inzicht in de resultaten die uit dit project zijn voortgekomen.

Nadat we hieronder kort ingaan op het nut van taalbeschrijving, geven wij een beschrijving aan van het project zelf. Daarin wordt ingegaan op het doel van het project, de fasering en de werkwijze ervan. Tevens gaan we in op de vragenlijst die we hebben gebruikt voor de enquêtes. In een volgende paragraaf bespreken we summier de invloed van het Nederlands op het Afrikaans in historisch perspectief, waarbij we ons beperken tot die aspecten die relevant zijn voor het onderhavige onderzoeksproject. Daarna bieden wij een overzicht van de resultaten en conclusies, zowel op fonetisch als op lexicaal gebied, zoals die uit het onderzoek zijn voortgekomen. We besteden hierbij tevens aandacht aan de correlatie tussen de aanwezigheid van Nederlandse taalresten enerzijds en de leeftijd, het opleidingsniveau en het geslacht van de informanten anderzijds. Bovendien vermelden we hier een aantal indirecte resultaten uit het onderzoek die wij voor ons doel noemenswaardig achten. We beëindigen dit artikel door de aandacht te vestigen op het belang van dit onderzoek voor de lexicografie.

2. Matjieshuizen of het nut van taalbeschrijving

Het nut van een taalbeschrijving als die van de variëteiten van het Afrikaans

ligt niet alleen op taalkundig vlak, maar heeft ook cultuurhistorische waarde. De traditionele woordenschat verdwijnt snel en is vaak nog enkel opgeslagen in het geheugen van de oudste generatie. Het is belangrijk dit materiaal op te vragen en te registreren en daarmee een historische woordenschat die hoort bij een samenlevingsvorm die aan het verdwijnen is of al is verdwenen, vast te leggen.

Registratie van het hiergenoemde materiaal en opname ervan in het woordenboek heeft natuurlijk in de eerste plaats een taalkundig nut, omdat het een deel van de taalwerkelijkheid ontsluit. Vooral via mondelinge enquêtes is het vaak mogelijk zeer fijn onderscheiden klankvariaties te geven, zoals ook is gebleken uit het hier te bespreken project. Belangrijker nog is de lexicaal-semantische en etymologische waarde van dit soort onderzoek. Woordenboeken openen hiermee namelijk de mogelijkheid semasiologische en onomasiologische woord- en betekenisvelden te bestuderen en de verspreidingsgeschiedenis van woordvormen en van betekenissen na te gaan. Kennis van variëteiten van een taal, toegankelijk gemaakt door een systematisch geordende verzameling van het materiaal, is een belangrijk bestanddeel van taalkundige kennis in ruimere zin.[1]

Daarnaast wordt met het beschrijven van variaties van het Afrikaans een cultuurlandschap ontsloten. Doordat een woord meestal langer blijft bestaan dan de referent, is lexicologisch onderzoek ook dienstig aan de cultuurgeschiedenis: de taal, de variatie geeft een afspiegeling van de alledaagse werkelijkheid van een bepaald deel van de bevolking in een bepaalde tijd. Men vergelijke bijvoorbeeld het woord *matjieshuis* waarvan de referent nagenoeg verdwenen is. Tijdens ons onderzoek hebben we in Steinkopf nog een aantal van deze matjieshuizen aangetroffen.

Het *WAT* zegt over *matjieshuis* onder andere het volgende: een matjieshuis is een "informele of tydelike woning wat bestaan uit 'n sirkelvormige raamwerk van pale wat met dwarslatte aan mekaar verbind en met matjies gedek word". Hier (in Namakwaland) *het die boere die matjieshuis ontwikkel en agter die veetroppe aan "verhuis"* (A. Coetzee in G.J. Labuscagne: Feesbundel, 1959, 47). Matjieshuizen kunnen makkelijk worden vervoerd, hetgeen een ideale behuizing is voor het nomadenvolk in Namaqualand, ideaal ook in het klimaat in deze streek: de matten zwellen op als het regent, waardoor ze nauwelijks water doorlaten en ze krimpen als het warm wordt, zodat de wind erdoorheen kan waaien. Nu de mensen andere vormen van bestaan hebben gevonden en niet meer trekken, zijn de matjieshuizen nagenoeg verdwenen.

Taal is niet alleen een afspiegeling van de culturele geschiedenis van een (deel van de) bevolking, maar in taal zien we ook sociale variaties; een taalvariëteit kan sociaal en situationeel gebonden zijn. In ons onderzoek blijkt dat bijvoorbeeld uit de verschillende resultaten bij informanten met een middelbare schoolopleiding of zij die lagere of geen schoolopleiding hebben gehad: de eerste groep neigt er veel meer naar de standaard Afrikaanse uitspraak, de uitspraak die sociaal aanvaard is, te hanteren dan de tweede groep.

3. Beschrijving van het project

3.1 Doel

In het onderzoek naar de taalvariaties binnen het Afrikaans spelen de talen die het Afrikaans hebben beïnvloed een bijzondere rol. Aangezien het Afrikaans voor een belangrijk deel uit het Nederlands is ontstaan — de woordenschat van beide talen loopt voor een groot gedeelte parallel — is dit deelproject van eminent belang.

De bedoeling van het onderzoek is na te gaan of verschijnselen die in de dialectische, sociolectische of etnolectische variëteiten van het Afrikaans voorkomen, maar niet in het standaard Afrikaans, verklaard kunnen worden als import uit het Nederlands (en de historische dialecten daarvan) als "toeleverancier" van taalvormen, ofwel als ontwikkelingen binnen het Afrikaans zelf.

3.2 Fasering

Het project is opgedeeld in vier fasen.

In een eerste fase is er een verkennend vooronderzoek geweest in Namaqualand in Springbok en omgeving. Namaqualands, als een van de opvallende variëteiten van het Afrikaans, zou een goede aanduiding kunnen bieden van de mogelijke resultaten.

In fase twee is er in samenwerking met de medewerkers uit Gent en Nijmegen een vragenlijst opgesteld voor het inwinnen van materiaal.

De derde fase bestond uit het werven en voorlichten van medewerkers en uit het toetsen en invullen van de vragenlijsten. Voor dit doel zijn in eerste instantie de vijf genoemde gemeenschappen benaderd: Springbok en omgeving, Genadendal, Mamre, Saron en de Moslimgemeenschap in Kaapstad.

In de vierde fase is ten slotte het verzamelde materiaal ontleed en verwerkt. Dit materiaal zal worden ingevoerd in het normale systeem van het Buro van die WAT.

3.3 Werkwijze

De gegevens voor het onderzoek zijn verzameld door middel van vragenlijsten. Daartoe zijn medewerkers aangezocht die de plaatselijke variëteit kennen en tegelijk taalkundig zodanig zijn onderlegd dat ze ook bij de opvraging en de eerste verwerking behulpzaam kunnen zijn. De medewerkers hebben plaatselijke zegslieden gezocht en hen de vragenlijst voorgelegd, waarna de vragenlijsten terug zijn gestuurd naar het Buro van die WAT waar ze door schrijfster dezes zijn verwerkt. De vragenlijsten zijn steeds afgenomen door betrokken medewerkers of door Van Schalkwyk zelf.

De gebruikte vragenlijst is tot stand gekomen op basis van resultaten uit het vooronderzoek in de eerste fase en bestaat uit twee delen: een fonetisch/

fonologisch deel en een lexicaal deel. In het eerste deel gaat het met name om de volgende klankverschijnselen: het behoud van [y] voor [œy] (als in *huus* voor standaard Afrikaans *huis*), het behoud van [i] voor [əi] (als in Zeeuws *tied* voor standaard Afrikaans *tyd*) en het behoud van [ø] voor [e:] (als in *veul* voor standaard Afrikaans *veel*).

Het tweede gedeelte betreft vragen naar zowel de woordenschat als idiomen. Het gaat hier om variëteiten die nog wel in het Nederlands maar niet meer, of niet meer algemeen in het standaard Afrikaans voorkomen, zoals *geit* voor *bok* of *nagel* voor *spyker*.

In de vragenlijst voor de eerste enquêteronde, in Namaqualand, is een greep gedaan uit de woordenlijst die is opgesteld uit de fase van het vooronderzoek (fase 1). De vragenlijst is niet uitputtend en in die zin kan deze eerste vragenronde als experimenteel worden beschouwd. Resultaten uit dit project kunnen waardevol zijn voor eventueel volgend onderzoek en kunnen in volgende vragenlijsten worden verwerkt.

In Namaqualand zijn 54 enquêtes afgenomen bij een dwarsdoorsnee van de bevolking van Springbok en omgeving. Men kan dus ook gezien het beperkte aantal ondervraagden in deze fase spreken van een monsteronderzoek. Voor de andere gebieden betreft het in totaal 129 vragenlijsten.

Naast de vragenlijsten is er een narratief deel waarin mondelinge overlevering is geregistreerd: vrije opvraging van volksverhalen. Resultaten van dit onderdeel worden door de schrijfster dezes in eerste instantie onderzocht op narratieve en historische kwaliteiten, maar kunnen tevens een belangrijke bijdrage leveren tot de kennis van de uitdrukkingsrijkdom van variaties van het Afrikaans. De verhalen zijn niet opgenomen in het onderzoeksrapport, maar zijn beschikbaar bij het Buro van die WAT.[2]

3.4 De vragenlijst

In elk onderzoeksgebied hebben we mensen benaderd die gaan werken met de vragenlijsten. Tijdens de voorlichting aan de medewerkers hebben we hen gewezen op een aantal belangrijke zaken:

- de vragenlijst moet altijd zo worden gehanteerd dat de vragensteller de informant niet een uitspraak of antwoord in de mond legt;
- het is te verkiezen dat de informant ouder is dan 65 jaar, tenzij er jongere kandidaten zijn die de variant van de streek goed spreken;[3]
- het is belangrijk dat de vragensteller informanten uit de hele opnamearea vraagt;
- de vragensteller moet de antwoorden van de informanten zo nauwkeurig mogelijk beoordelen en op de vragenlijst aanduiden.

De vragenlijst heeft een A-gedeelte en een B-gedeelte. In het A-gedeelte wordt gevraagd naar bijzonderheden over de informant. Het B-gedeelte bevat de eigenlijke vragenlijst.

Bij de vragen van het B-gedeelte zijn telkens antwoordmogelijkheden aangegeven in de lijst, daarnaast is er ruimte voor alternatieve antwoorden van de informant. Bovendien is aan de medewerker gevraagd aantekeningen te maken van opmerkelijke woorden of uitspraken tijdens de enquête.

In het A-gedeelte van de vragenlijst zijn nauwgezet de biografische gegevens van de informanten genoteerd om achteraf de taalvormen te kunnen duiden. Tevens hebben we deze gegevens gebruikt om te kunnen onderzoeken of er correlaties bestaan tussen de aanwezigheid van taalresten enerzijds en het geslacht, het opleidingsniveau en de leeftijd van de informanten anderzijds.

Opgetekend zijn:

— datum en plaats van de enquête en de naam van de medewerker;
— naam van de informant;
— geboortejaar en geboorteplaats, geslacht, plaats waar de informant is opgegroeid en de huidige verblijfplaats;
— de huistaal van de informant en de moedertaal van elk van zijn ouders;
— de taal/talen waarin de informant de schoolopleiding ontving;
— de opleiding/hoogste kwalificatie van genoten onderwijs van de informant.

Het B-gedeelte, de eigenlijke vragenlijst, is tot stand gekomen op basis van resultaten uit het vooronderzoek in de eerste fase en bestaat uit een fonetisch en een lexicaal deel.

In het eerste deel gaat het met name om klankverschijnselen; hierin wordt in de vorm van vraag- en invuloefeningen nagegaan of er in de variëteit van de informant sprake is van de volgende klankwisselingen:

— het behoud van [y] voor [œy], zoals in *huus* in plaats van *huis*, of dat de informant de [œy]-woorden misschien nog heel anders uitspreekt, b.v. een [u], zoals in *hoes*;
— op soortgelijke wijze onderzoeken we de [əi]/[i]-wisseling, zoals in *wief* in plaats van *wijf*;
— in een derde invuloefening onderzoeken we de [e:]/[ø]-wisseling van woorden als *besem* en *beusem*;
— met een vierde invuloefening gaan we na of er nog sprake is van een intervocalische [x] in de taalvariant van de informant, zoals bij woorden als *vogel* tegenover *voël*;
— in een vijfde onderdeel onderzoeken we "varia" als de [ø]/[o:]-wisseling, de [a]/[ɛ]-wisseling, en de [ɛ]/[a:]-wisseling.

Na dit fonetisch-fonologisch gedeelte van de vragenlijst, volgen twee lexicale onderdelen:

— onderzoek naar de woordenschat/betekenis van woorden (zoals *krank, rund, vaak*);

— in het laatste deel gaan we na welke idiomen bekend zijn bij de informant (uitdrukkingen als *die kans is verkyk* of *om met iets op die proppe te kom* of *hy is 'n hele Piet*).

In deze laatste onderdelen betreft het woorden en idiomen die in het standaard Afrikaans niet meer worden gebruikt, of waarvan de gebruiksfrequentie in standaard Afrikaans aan het afnemen is.

In het totaal hebben we 183 informanten bevraagd: 54 in Namaqualand, 32 in Genadendal, 30 in Mamre, 36 in Saron en 31 in de Moslimgemeenschap in Kaapstad.

4. Invloed van het Nederlands op het Afrikaans

In het fonetisch/fonologische onderdeel van het project hebben we het behoud van een vroeger taalstadium van het Nederlands (17e–19e eeuw) in het Afrikaans onderzocht. Dit verschijnsel is niet nieuw: Rademeyer (1938) heeft de kwestie beschreven en ook Kloeke (1950) heeft de rol van de historische dialecten in de opbouw van het Afrikaans ter sprake gebracht. Recenter onderzoek hiernaar treffen we aan bij Van Schalkwyk (1983) die zich speciaal richtte op de taal van de Rehoboth Basters en Links (1989) die de taal van de Kharkams onderzocht. Ook Ponelis (1993) besteedt in zijn studie over de ontwikkeling van het Afrikaans aandacht aan de invloed van het Nederlands op het Afrikaans.

Zowel Ponelis als Kloeke wijzen erop dat het niet de standaardtaal van Nederland is die van invloed was op de ontwikkeling van het Afrikaans. Ponelis ziet het Afrikaans als een koloniale en Hollandse variatie van het Nederlands: de Hollandse variëteit is dominant in de Nederlandse basis van het Afrikaans. Standaard Nederlands, in geschreven vorm, staat los van de spreektaalvariëteiten. Er zijn veel onderlinge verschillen tussen spreektaalvariëteiten van het Nederlands, maar ze hebben ook een aantal overeenkomsten die hen collectief onderscheiden van het Algemeen Beschaafd Nederlands en sommige komen eveneens voor in het Afrikaans. Als we die overeenkomsten in ogenschouw nemen, moeten we — aldus Ponelis — beseffen dat zowel het Afrikaans als de Nederlandse dialecten zijn veranderd sinds de vroeg moderne tijd en dat sommige overeenkomsten sindsdien zijn verdwenen uit het Afrikaans of uit sommige Nederlandse dialecten. Ponelis wijst er hier op dat de spreektaalkenmerken niet zijn overgenomen uit willekeurige, verschillende dialecten, maar dat ze waarschijnlijk allemaal een deel vormden van de 17e-eeuwse Hollandse of Amsterdamse spreektaal.

In ons onderzoek hebben we ons in eerste instantie gericht op de volgende drie klankverschijnselen:

(a) de [i]/[əi]-wisseling, als in het Zeeuwse *tied* voor het Nederlandse *tijd* en het Afrikaanse *tyd*;
(b) de [y]/[œy]-wisseling, als in *huus* voor standaard Afrikaans *huis*;
(c) de ronding van [e] tot [ø] als in dialectisch *zeuven* voor *zeven*.

Vanuit het standpunt van de historische dialectologie van het Nederlands, is het verbazingwekkend dat we in het vooronderzoek wel herhaaldelijk dialectisch on-Hollands [y] in plaats van [œy] aantroffen (in woorden zoals *kruus, bruun, duuwel*) en zelden of nooit [i] in plaats van [əi]. In de ontwikkeling van het Nederlands lopen de diftongeringen van de oude Westgermaanse [u] (via [y]) tot de tweeklank [œy] en van de lange [i:] tot de tweeklank [əi] immers parallel en gelijktijdig.

Het voorkomen van deze tweeklanken [œy] en [əi] in het Afrikaans was juist een hoofdargument voor Kloeke om als stamland van de Afrikaanse taal het zuidelijk deel van Holland aan te wijzen. Immers, zo redeneert Kloeke, op grond van aantallen immigranten zou men evengoed een Duits of een Engels grondpatroon van de Afrikaanse klinkers hebben kunnen vinden. De parallellie van beide tweeklanken echter in de Nederlandse (Hollandse) dialecten en Afrikaans is doorslaggevend. Vergelijk:

> Engels *tide*, Duits *Zeit* tegenover Nederlands *tijd* en Afrikaans *tyd*;
> Engels *bite*, Duits *beissen* tegenover Nederlands *bijten* en Afrikaans *byt*;
> Engels *house*, Duits *Haus* tegenover Nederlands *huis* en Afrikaans *huis*;
> Engels *brown*, Duits *braun* tegenover Nederlands *bruin* en Afrikaans *bruin*.

De Nederlandse tweeklanken [œy] en [əi] komen, volgens de kaarten van Kloeke, in zo goed als hetzelfde dialectgebied voor: in de driehoek met de basis aan de taalgrens en de top in de buurt van Den Helder. De ontwikkeling is begonnen in het zuiden en is uit de "zuidtaal" door de Hollandse dialecten overgenomen in de "noordtaal" en vandaaruit tot standaard geworden.

4.1 De [əi]/[i]-wisseling

Met zijn *ijs*-kaart geeft Kloeke een eerste houvast aan de taalgeograaf die op zoek is naar de bakermat van het Afrikaans. Het lijdt volgens Kloeke geen twijfel dat de Nederlandse dialecten, die worden gesproken in het gebied tussen de Waalse taalgrens in het zuiden en de punt van Noord-Holland in het noorden, door deze habitus verraden dat zij van alle Europese dialecten het naast verwant zijn aan het Afrikaans. Dat er sprake is van een oorspronkelijke lange [i:] blijkt uit de tegenwoordige Nederlandse schrijfwijze met het teken [əi] (*bijten, rijden*). De voorraad woorden met vroegere [i:] die nog in het Afrikaans leeft, is tamelijk groot en komt overeen met de voorraad gediftongeerde woorden die wij in het gebied op de kaart vinden: *afgryselik, bly, blyk, ry*, enzovoorts.

In het grootste deel van het vasteland van Holland moet de diftongering in de 17e eeuw definitief haar beslag hebben gekregen, misschien eerder. Kloeke concludeert dat alle Afrikaanse woorden met gespelde *y* (uit [i]) Hollands erfgoed zijn en dat ook de relicten uit Holland afkomstig zijn.

4.2 De [œy]/[y]-wisseling

Speciaal over de [œy] zijn interessante gevallen te melden in ons onderzoeksgebied. Evenals de oude [i] is ook de oude [u] zowel in het Nederlands als in het Hoogduits en Engels gediftongeerd. Maar de Nederlandse ontwikkeling wijkt van beide andere af, doordat aan de Nederlandse diftongering een tussenstadium is voorafgegaan: de [y] met umlaut, ontstaan in de tijd van de Frankische nederzetting uit en in de nabijheid van de Romaanse ontwikkeling.

Daarna komt de diftongering van de [y] tot [œy], ook eerst in het zuiden. De eerste sporen ervan kan men al vroeg in het Middelnederlands aanwijzen. In elk geval kreeg de diftongering zijn beslag nog voor het eind van de middeleeuwen in Brabant, met Antwerpen als centrum. Omstreeks 1600 begint de [œy] als voorname uitspraak in Amsterdam veld te winnen, wat des te makkelijker kon, omdat een aanloop tot diftongering ook in Holland reeds aanwezig was. In de 17e eeuw behoort de diftong [œy] tot de beschaafde uitspraak in Holland en komt zo in de standaardtaal terecht; Goeree en Overflakkee en de Zeeuwse eilanden blijven als relictgebied met [y] over. Terwijl de [y] enerzijds terrein verloor aan het Algemeen Beschaafd Nederlands, werd hij anderzijds tezelfdertijd expansief. Vanuit de lagere maritieme milieus, die in Amsterdam tot in de 18e eeuw de [y]-uitspraak bleven gebruiken, ging de [y] naar de Waddeneilanden, het gebied om de Zuiderzee; wellicht via Utrecht naar de Veluwe.

Kloeke heeft dit proces laten zien op zijn *huis*-kaart, waarop de drie achtereenvolgende gestalten van de Oudgermaanse lange [u] worden onderscheiden; het oudste [u]-stadium en het daarop volgende [y]-stadium zijn goed geconserveerd. Kloeke laat hier geografisch naast elkaar zien wat er chronologisch na elkaar is gebeurd:

(a) De oostelijke rand van het Nederlandse taalgebied (en ook de aangrenzende Nederduitse en Rijnlandse dialecten) en het Fries zijn met de uitspraak *hoes* op Oudgermaans standpunt blijven staan.
(b) Een reeks kleinere en twee grote *huus*-gebieden zijn op Middelnederlands standpunt blijven staan: West-Vlaanderen, Zeeland, Gelderland behalve de Betuwe, deels Overijssel, Drenthe, Groningen, Friesland en de Waddeneilanden.
(c) Het meest vooruitstrevende gebied met diftongering is Oost-Vlaanderen, de Brabantse provincies, half Utrecht en Zuid- en Noord-Holland.

Kloeke concludeert dat de *huis*-kaart, evenals de *ijs*-kaart, aantoont dat de naaste taalverwanten van het Afrikaans vooral te vinden zijn in het westelijke gebied. Binnen de grenzen daarvan, "en nergens anders", is volgens hem het stamland van de Afrikaanse [œy]-uitspraak te zoeken.

Ook bij de ontwikkeling van de oude [u] zijn enkele woorden in de oude toestand gefixeerd tot relicten; hier hebben we dus twee soorten: de oude [u] zelf en de geümlaute [u].

In het Nederlands tellen we een veertigtal [u]-relicten, waarvan ongeveer de helft mee naar Afrika is gekomen: bijvoorbeeld aspoestertjie (*poesten* naast mnl. *puysten* "blazan"), boer (etymologisch naast *buur*), enzovoorts. Als [y]-relict is het woord *ruzie* (vergelijk *roes*) naar Afrika gekomen. Zoals bij de [i] is de aanwezigheid van relicten een sterk bewijs voor overname.

4.3 De parallellie doorbroken: wel [y] en geen [i]?

Rademeyer (1938: 49) had al opgemerkt, en Kloeke valt hem in deze bij, dat bij bruin Afrikaanssprekenden de [y]-variant vaak wordt aangetroffen. Afgezien van het bekende *Duusman* voor *Diesman* "wit Afrikaanssprekende" < Duits- of Dietsman, hoort men in de taal van de Griqua's en de Rehoboth Basters wel *huus, tuus, muus*, enzovoorts. Kloeke (1950: 214) wijst op andere aanwijzingen waaruit blijkt dat de [y]-realisatie ook elders en niet alleen bij bruin Afrikaanssprekenden voorkomt en hij wijst de gedachte dat we hier "een oude echo van de taal der 17e-eeuwse blanken" te maken kunnen hebben, niet van de hand: "Men vergete niet, dat de taal van zuidelijk Zuid-Holland het meest heeft bijgedragen tot de vorming van het Afrikaans. Zuidelijk Zuid-Holland nu ligt niet alleen vlak tegen het ongediftongeerde gebied aan (het Zuidhollandse Goeree en Overflakkee behoort zelfs nog tot het monoftongische gebied) maar heeft de diftongering blijkbaar ook pas laat aangenomen."

Van de andere kant wijst Rademeyer, en volgens Kloeke terecht, op het bevreemdende feit dat de [i] bij de Basters niet "bewaard" is gebleven. Rademeyer (1938: 50) vraagt: "Sou die rede miskien wees dat die diftongering van [i] plaasgevind het voor die van [y], sodat die dialeksprekende immigrante van die 17e eeu onbekend was met die monoftongiese [i]?", een vraag die volgens Kloeke nog moet worden beantwoord, aangezien we overal in het Nederlands een parallellisme van [i] > [əi]- en [y] > [œy]-ontwikkeling (geografisch en historisch) kunnen constateren. De afwezigheid van [i] zou volgens Kloeke aan de taaltoestanden van de Basters kunnen worden toegeschreven, die blijkbaar "mengingen" ten gevolge hebben gehad die in Europa onbekend zijn — een intern Afrikaanse ontwikkeling dus.

Enerzijds wijst Kloeke dus voor het behoud van [y] op de nabijheid van de Zeeuwse dialecten, anderzijds voor de afwezigheid van [i] op intern-Afrikaanse ontwikkelingen.

4.4 Ronding van [e] tot [ø] als in *zeuven* voor *zeven*

Weijnen (1966[2]: 225) en Schönfeld (1970[8]: 49-50), en Kloeke (1950: 166) met hem, beschrijven deze zeer frequente ronding in woorden als *speulen, veul, beuzem* voor *spelen, veel, bezem* in de dialecten van Holland, Utrecht en Noord-Brabant, maar ook wel in het oosten van Nederland. We hebben deze wisseling opgenomen omdat er in Zuid-Afrikaanse literatuur over variaties (vgl. Van Schalk-

wyk 1983 en Links 1989) wordt gewezen op het bestaan ervan. Kloeke, die de vorm *seuwe* niet geattesteerd heeft gezien in het Afrikaans, merkt op dat deze, indien het een relictvorm in Zuid-Holland is — en dat bepleit hij op dialectologische grond — er te verwachten zou zijn. We hebben deze vorm inderdaad aangetroffen.

Verder worden in de vragenlijst nog enkele niet-standaard klankverschijnselen die in de literatuur waren gesignaleerd onder de loep genomen, zoals het voorkomen van [ɛ] voor [a:], en het behoud van de intervocalische [x] in *oge, dage, ogenblik* die Links (1989: 24) in Kharkam signaleert.

4.5 Lexicale resten

In het laatste onderdeel van de vragenlijst gaan we na welke lexicale resten van het Nederlands nóg bestaan in archaïsche variëteiten van het Afrikaans. Als bronnen hebben we voornamelijk gebruik gemaakt van Links (1989) en Botha et al. (1994).

Links gaat in zijn boek in op het metaforisch woord- en idioomgebruik van de Kharkams in Namaqualand. De meeste van de woorden, woordgroepen en uitdrukkingen die hij in een lijst weergeeft, zijn in het standaard Afrikaans onbekend en hij geeft er dan ook alle opgegeven betekenissen en connotaties bij met het oog op het achterhalen van de mogelijke etymologie. De bedoeling van deze paragraaf van onze vragenlijst is aanvullingen en geografische uitbreiding te verkrijgen op deze door Links ingeslagen weg.

Vaak zijn woorden uit de algemene taal verdwenen als de zaken waarvoor zij staan verdwijnen; als voorbeeld noemt Links *snuiter* en *konfoor*. Omdat echter vele mensen in zijn onderzoeksgebied (Namaqualand) een tamelijk geïsoleerd bestaan hebben, zijn veel "ouderwetse" zaken zoals de oude lengtematen *duim* (uit het Nederlands) en *jaart* (uit het Engels), en daarmee hun betekenaars, de woorden, hier bewaard gebleven (zie Links 1989: 69).

Enkele voorbeelden met Nederlandse lexicale restanten uit de vragenlijst zijn:

> *blootvoet* (Nederlands *blootvoets*), voor Afrikaans *kaalvoet*;
> *ontskiet* en partikel *ontskoot* (Nederlands *ontschieten*) voor Afrikaans *nie onthou nie*; enzovoorts.

In het onderzoek zijn ook uit Botha et al. (1994) uitdrukkingen overgenomen, die niet standaard Afrikaans zijn en die hun oorsprong in het Nederlands vinden. Voorbeelden hiervan zijn:

> *die kans is verkyk* (de kans is voorbij);
> *hy is 'n hele Piet* (hij is een belangrijke persoon);
> *om met iets op die proppe te kom* (om met iets voor de dag te komen).

5. Resultaten en conclusies[4]

Uitgaande van de theorie betreffende de invloed van het Hollands op het Afrikaans en gezien de aard van de taalresten die we hebben aangetroffen in de variëteiten van ons onderzoeksgebied, mogen we ervan uitgaan dat deze taalverschijnselen een overblijfsel zijn uit het Hollands. Zowel de spraakklanken als de lexicale taalverschijnselen zijn naar alle waarschijnlijkheid Nederlandse taalresten, aangezien het standaard Afrikaans een eigen ontwikkeling heeft doorgemaakt waarin deze verschijnselen niet (meer) of slechts sporadisch voorkomen.

Met name door het behoud van dialectische kenmerken van het Hollands, zoals de frequente ronding van de [e:] tot [ø] en ook van lexicale en idiomatische niet-standaard Afrikaanse maar wel Nederlandse elementen, lijkt het behoud van een oorspronkelijke variatie aannemelijk.

Naast de gevraagde resultaten heeft het onderzoek een aantal gegevens opgeleverd waarnaar niet direct is gevraagd, maar die vanuit taalkundig oogpunt interessant zijn om hier op te nemen. Van een aantal van deze taalverschijnselen kan men sterk vermoeden dat hier tevens sprake is van invloed vanuit het Nederlands.

5.1 Fonetische taalresten

Wat betreft de fonetische taalresten hebben we ons vooral gericht op de [œy]/[y]-, de [əi]/[i]- en de [e:]/[ø]-wisseling en de aanwezigheid van de intervocalische [x]. Daarnaast hebben we een aantal andere wisselingen onderzocht, namelijk [e:]/[o:], [ɛ]/[a:], [ø]/[ɔ] en [a]/[ɛ].

Een overzicht van de soorten fonetische resten:

[œy]/[y]-wisseling		10%	(206)
[e:]/[ø]-wisseling		18%	(166)
[əi]/[i]-wisseling	[i]	3%	(57)
	[i:]	1%	(11)
intervocalische [x]		1%	(6)
varia		3%	(26)

Een vergelijking van de verschillende onderzoeksgebieden op aanwezigheid van Nederlandse taalresten:

Namaqualand	16%	(334)
Genadendal	3%	(31)
Mamre	8%	(89)
Saron	1%	(7)
Moslimgemeenschap	1%	(10)

De [e]/[ø]-wisseling is met 18% de meest voorkomende fonetische taalrest gebleken. We hebben dit klankverschijnsel onderzocht in de woorden *sewe, speel, veertig, veel* en *besem*. De klankwisseling komt het meest voor in Namaqualand (46%). Opmerkelijk is dat Links (1989: 9) in zijn onderzoek naar de Kharkamstaal deze wisseling slechts in twee gevallen heeft aangetroffen, terwijl nu blijkt dat in de rest van Namaqualand dit klankverschijnsel (de [ø] voor de [e:]) veel frequenter voorkomt.

De [œy]/[y]-wisseling (in de woorden *bruin, duim, kruis* (rug), *kuiken, muis, suiker, volstruis, kruis* (kerk), *vuis, buite, duiwel, huis*) is de tweede meest frequent voorkomende fonetische taalrest: in 10% van de gevallen horen we een [y] in plaats van een [œy]. Ook wat deze spraakklank betreft, springt Namaqualand er uit: in dit gebied heeft deze wisseling met 20% de hoogste frequentie.

De [əi]/[i]-wisseling (in de woorden *by, grys, konyn, vyf, kyk, gebyt, ryk, skryf, slyp, wyfie*) is — in tegenstelling tot de [œy]/[y] — veel minder vaak aangetroffen: gemiddeld slechts voor 3%; de lange [i:] komt maar voor 1% voor. Kloeke (1950: 215), Rademeyer (1938: 50) en Links (1989: 21) hebben gewezen op het feit dat de [i] veel minder bewaard is gebleven dan de [y], terwijl de [əi]/[i]-wisseling meestal samengaat met de [œy]/[y]-wisseling. Kloeke, refererend aan de taal van de Basters, schrijft dit "afwijkende gedrag der kleurlingentaal" toe aan specifiek-Afrikaanse taaltoestanden bij de Basters, maar een werkelijke verklaring voor de afwezigheid van de parallelle ontwikkeling moet volgens hem nog gevonden worden.

Het verschijnsel blijkt zich nu voor te doen in een groter gebied dan alleen bij de Basters, waarbij we overigens in gedachten moeten houden dat de Basters oorspronkelijk afkomstig zijn van Namaqualand. Bovendien, in het zuiden van de West-Kaap, bij de Moslimgemeenschap in Kaapstad, doet zich het omgekeerde voor: hier treffen we de [y]-klank slechts 1 keer aan, terwijl er 9 maal de [i]-klank is geconstateerd.

Deze resultaten geven nog steeds geen antwoord op de vraag waarom de Zuidafrikaanse ontwikkeling die bijna absolute parallellie van de oude [i] en de oude [u] in de Nederlandse dialecten tegenspreekt. Volgens Joep Kruijsen, die dit verschijnsel heeft besproken in een lezing in april 1998 in Nijmegen, is de combinatie niet onmogelijk. Hij haalt hierbij Weijnen (1991: 31) aan, die er in zijn bespreking van de *huis*-kaart van Kloeke op wijst dat het uiterste zuid-westen van Zuid-Holland, Voorne-Putten, dat tegen het overwegend [y]-realiserend Zeeuws aanligt, zich kenmerkt door een enigszins open korte [y]-klank. Kruijsen vraagt zich af of hier misschien de schakel ligt die de historische dialectkunde kan leggen tussen de *bruune suuker* in Steinkopf en de zuidelijkste Zuidhollandse dialecten.

Voorlopig lijken we ons echter te moeten houden aan Kloekes opvatting dat de doorbreking van de parallellie te beschouwen is als een intern Afrikaanse ontwikkeling, met dien verstande dat de [i], zij het veel minder, overigens wel voorkomt.

Het voorkomen van de intervocalische [x] is door Links (1989: 24) al geconstateerd in de taal van de Kharkams. Uit ons onderzoek (naar de woor-

den *reën, leuen, voël, spieël, seël, oomblik*) blijkt dat dit verschijnsel nog voorkomt in een breder gebied, zij het niet zeer frequent. In totaal komt de intervocalische [x] 6 maal (1%) voor: 4 maal in Genadendal en 2 maal in Namaqualand.

Voor de overige spraakklanken die we hebben onderzocht, komen de [ø]/[ɔ]-wisseling in *seun* (9 maal: 6%) en de [œ]/[ɛ]-wisseling in *murg* (8 maal: 5%) het meest voor; de eerste spraakklank treffen we het meest aan in Namaqualand (11%) en de tweede het meest in Mamre (13%). De [ø:]/[o:]-wisseling in *kneukels* komt in totaal 4 maal (3%) voor, de [ɛ]/[a:]-wisseling in *kerse* 3 maal (2%) en de [a]/[ɛ]-wisseling in *vars* 2 maal (1%).

Vergelijken we de verschillende onderzoeksgebieden met elkaar, dan blijkt dat in Namaqualand de meeste fonetische taalresten (334 totaal: 16%) zijn aangetroffen. In Mamre vinden we in totaal 89 resten (8%) en in Genadendal 31 (3%). In Saron (7: 1%) en in de Moslimgemeenschap (10: 1%) treffen we nog maar sporadisch spraakklanken aan die als taalrest uit het Nederlands kunnen gelden.

5.2 Lexicale taalresten

Om na te gaan of er nog lexicale taalresten van het Nederlands bestaan in de variëteiten van ons onderzoeksgebied, hebben we gevraagd naar zowel de betekenis van afzonderlijke woorden als naar de betekenis van idiomen. Door de aard van de vragen hebben we vooral kunnen nagaan of de woorden en idiomen nog aanwezig zijn in de passieve woordenschat van de respondenten. Alleen bij de invuloefening waar wordt gevraagd naar de woorden *blootvoet* en *ontskiet/ontskoot* hebben we informatie gekregen over de actieve woordenschat van de respondenten.

Een totaaloverzicht van de resultaten per regio:

Namaqualand	23%	(252)
Genadendal	31%	(197)
Mamre	25%	(147)
Saron	15%	(106)
Moslimgemeenschap	53%	(326)

Een totaaloverzicht van de resultaten uit de *Woordeskat:*

Wat beteken die volgende woorde?

Krank	63%	(116)
'n Rund	13%	(24)
'n Stier	38%	(69)
'n Geit	12%	(21)
'n Keuken	3%	(6)
'n Nagel	8%	(14)
Suutjies	72%	(132)
Vaak	1%	(1)

Vul die ontbrekende woord in

Iemand wat nie sokkies en skoene aan het nie, loop ..... *blootvoet*	20%	(36)
Wanneer ek iets vergeet het, het dit my ..... *ontskiet/ontskoot*	5%	(9)

Een totaaloverzicht van de resultaten uit de *Idiome:*

Die kans is verkyk/verkeke	63%	(116)
Om 'n gat in die hand te hê	33%	(60)
Om baie pyle op jou boog te hê	25%	(46)
Om met iets op die proppe te kom	29%	(53)
Hy is 'n hele Piet	19%	(34)
Om iemand die mette te lees	35%	(64)
In die dae van Olim	34%	(62)
Om nie met iemand te kan opskiet nie	44%	(81)
Om veel met iemand op te hê	26%	(47)
Om van die reën in die drup te kom	18%	(32)

De resultaten uit het lexicografische gedeelte van het onderzoek tonen een totaal ander beeld als dat van het fonetische deel, met name wat betreft de spreiding van de aanwezige taalresten over het onderzoeksgebied.

Van de acht items uit de *Woordeskat* zijn de woorden *suutjies* (72%) en *krank* (63%) het meest bekend in alle onderzoeksgebieden. In de Moslimgemeenschap in Kaapstad zijn de informanten het meest bekend (38%) met de woorden uit dit onderdeel, gevolgd door Genadendal (31%). Maar ook in de andere gebieden kennen veel respondenten nog de betekenis van de woorden: in Namaqualand 26%, in Mamre 23% en in Saron 17% van het aantal gevallen.

Opmerkelijk is dat in de Moslimgemeenschap, die zo hoog scoort bij de betekenis van de woorden, de actieve woordenkennis wat betreft *blootvoet* en *ontskiet* geheel ontbreekt. Deze lexicale resten treffen we nog frequent aan in Namaqualand (32%) en iets minder in Mamre (10%) en Genadendal (8%). In Saron zijn deze taalresten in het geheel niet aangetroffen. Een opmerking van een informant in Saron is in dit opzicht daarom extra interessant: "Oumense het *blotvoet* gebruik." Over het totaal bezien scoort de actieve woordenschat (12%) in alle gebieden lager dan de passieve woordenkennis (27%).

Nog hoger dàn de bekendheid met de gevraagde woorden is de bekendheid op gebied van idiomen: gemiddeld 33%. Ook hier weer springt de Moslimgemeenschap er enorm uit met 75%. De respondenten in dit onderzoeksgebied geven exact de Nederlandse betekenis van de uitdrukkingen weer. Andere gebieden scoren ook hoog, maar aanzienlijk minder dan de Moslimgemeenschap: Genadendal 35%, Mamre 29%, Namaqualand 20% en Saron 16%.

Van de items uit dit onderdeel zijn de volgende uitdrukkingen in alle onderzoeksgebieden het meest bekend:

Die kans is verkyk	totaal 63%
Om nie met iemand te kan opskiet nie	totaal 44%

Het minst bekend zijn:

Om van die reën in die drup te kom	totaal 18%
Hy is 'n hele Piet	totaal 19%

In het totaaloverzicht van alle lexicale items levert wederom de Moslimgemeenschap de hoogste score op wat betreft aanwezigheid van lexicale resten: in meer dan de helft van de gevallen (53%) is hier bekendheid met de lexicale items geconstateerd. In Genadendal is dit ongeveer in een derde van de gevallen (31%), voor een kwart in Mamre (25%) en Namaqualand (23%) en het minst in Saron (15%).

Het is opvallend dat juist de Moslimgemeenschap zo hoog scoort hier, terwijl op fonetisch gebied hier nauwelijks Nederlandse taalresten zijn aangetroffen. Een mogelijke verklaring is dat in deze gemeenschap een negatieve connotatie wordt gegeven aan een afwijkende uitspraak, omdat de sprekers er niet van bewust zijn dat die wisseling aan een Nederlandse rest kan worden toegeschreven. Nederlands woordgebruik wordt juist positief gewaardeerd.

In Saron lopen de scores in het fonetische en lexicografische gedeelte meer gelijk: de variëteit die de mensen in Saron spreken vertoont op beide gebieden de minste Nederlandse taalresten.

5.3 De correlatie tussen taalresten en opleiding, leeftijd en geslacht

In deze paragraaf bespreken we de correlatie tussen de aanwezigheid van taalresten enerzijds en de genoten opleiding, leeftijd en het geslacht van de informanten anderzijds. We zijn daarbij uitgegaan van de totaalcijfers uit het onderzoek, zowel op fonetisch als op lexicografisch gebied.

Van de 183 informanten hebben 111 personen een lagere schoolopleiding of minder, onder wie 44 mannen en 67 vrouwen. In totaal hebben 71 informanten een middelbare schoolopleiding of meer genoten: 33 mannen en 38 vrouwen.

In totaal 141 respondenten zijn jonger dan 80 jaar, onder wie 66 mannen en 75 vrouwen. Veertig respondenten zijn 80 jaar of ouder: 11 mannen en 29 vrouwen.

Het totaal aantal mannen is 77 en het totaal aantal vrouwen is 106.

5.3.1 Fonetische taalresten

— Opleiding

Fonetische taalresten komen gemiddeld meer voor bij respondenten met een lagere schoolopleiding of minder (10%) dan bij respondenten met een middel-

bare schoolopleiding of meer (3%). Bovendien komen deze taalresten gemiddeld iets meer voor bij mannen (10%) dan bij vrouwen (9%).

Vrouwen met een middelbare schoolopleiding scoren echter hoger (4%) op de aanwezigheid van fonetische taalresten dan mannen uit deze categorie (1%).

— *Leeftijd*

Bij de groep informanten van 80 jaar of ouder vinden we gemiddeld iets meer fonetische taalresten (8%) dan bij de groep die jonger is dan 80 jaar (7%). Mannen laten in beide leeftijdscategorieën geen verschil zien: beide groepen scoren 6%. Bij de vrouwen is er wel een gering verschil: bij de groep van 80 jaar of ouder hebben we iets meer fonetische resten (8%) geconstateerd dan bij de jongere categorie (7%).

— Geslacht

In het totaal noteren we bij de mannen 182 fonetische taalresten (6%). Bij vrouwen ligt dit percentage iets hoger; in deze groep worden in totaal 299 fonetische taalresten geconstateerd (8%).

De verschillen qua leeftijd zijn zo gering dat er nauwelijks conclusies aan te verbinden zijn. Wel lijkt er een correlatie te bestaan tussen leeftijd en opleiding: respondenten van 80 jaar en ouder hebben in verreweg de meeste gevallen (87%) een lagere schoolopleiding; slechts 13% van de respondenten uit deze groep heeft een middelbare schoolopleiding genoten. Daarbij zij opgemerkt dat deze laatste categorie respondenten alleen uit vrouwen bestaat, hetgeen de correlatie tussen leeftijd en opleiding enigszins lijkt te verzwakken, aangezien bij vrouwen met middelbare schoolopleiding juist meer fonetische taalresten worden genoteerd. Of juist niet: zij zorgen ervoor dat het verschil slechts gering is. Minder fonetische taalresten bij respondenten met een middelbare schoolopleiding zou immers kunnen worden verklaard door de invloed van de leestaal.

Het totale verschil tussen mannen en vrouwen is eveneens gering: slechts 2%. We mogen hier misschien constateren dat de vrouwen in het onderzoeksgebied zich iets behoudender tonen in hun taalgebruik, of dat zij in posities verkeren (minder werk buitenshuis) die hen in staat stellen deze taalresten langer te gebruiken.

Van Schalkwyk (1983: ii) heeft in zijn onderzoek naar de taal van de Rehoboth Basters een soortgelijke bevinding opgedaan: vrouwen in deze taalgemeenschap neigen er meer naar de variatie te behouden, terwijl mannen de initiators zijn van verandering in Rehoboth Afrikaans.

In sociolinguïstische studies uit West-Europa en de Verenigde Staten zien we juist vaak dat vrouwen meer de standaardvariant produceren dan mannen en algemeen wordt aangenomen dat dit taalgedrag te maken heeft met hun taak in de opvoeding en de verwachting dat een goede vaardigheid in de standaardtaal betere toekomstkansen biedt voor opgroeiende kinderen. Van Schalkwyk wijst er in zijn dissertatie dan ook op dat zijn bevindingen die van taalwetenschappers als Labov en Trudgill tegenspreken; zij beweren juist het omgekeerde, namelijk dat vrouwen vooruitstrevender zijn dan mannen waar het

gaat om gebruik van de variëteit of de standaardtaal. Volgens onderzoek van Labov (1972: 243) zijn vrouwen gevoeliger voor de prestigevorm dan mannen en maken zij in nauwkeurig spraakgebruik minder gebruik van gestigmatiseerde vormen dan mannen. Dit zou met name gelden voor vrouwen uit de lagere middenklasse. Hij waarschuwt er echter tevens voor dat men niet mag aannemen dat het een algemeen beginsel is dat vrouwen altijd het voortouw nemen bij linguïstische veranderingen.

Volgens Van Schalkwyk (1983: 174) gelden vooral niet-talige veranderlijken als ouderdom en opleiding als factoren die de uitspraak van de Rehoboth Basters bepalen. Daartegenover leiden niet-talige veranderlijken als geslacht niet tot een beduidend verschil in de uitspraak. Aangezien het effect van het geslacht op de uitspraak bij de Rehoboth Basters heel anders is dan dat uit bij voorbeeld de bevindingen van Trudgill en Labov, concludeert Van Schalkwyk (1983: 175) dat deze correlatie kan verschillen per taalgemeenschap en dat sociolinguïsten hier voorzichtig moeten zijn met generaliseringen.

5.3.2 Lexicon

— Opleiding

In tegenstelling tot de fonetische taalresten, komen de lexicale resten juist meer voor bij respondenten met een middelbare schoolopleiding (36%) dan bij die met een lagere schoolopleiding (23%). Mannen met een middelbare schoolopleiding scoren iets hoger (36%) dan vrouwen (35%). Bij respondenten met een lagere schoolopleiding is het verschil groter: hier noteren we bij mannen 26% en bij vrouwen 22%.

— Leeftijd

Ook qua leeftijd is het beeld op lexicaal gebied totaal anders dan op fonetisch gebied. Respondenten die jonger zijn dan 80 jaar zijn aanmerkelijk beter bekend (30%) met het lexicon uit de vragenlijst dan respondenten van 80 jaar of ouder (19%). Van de jongste groep scoren mannen wederom hoger (32%) dan vrouwen (29%). In de oudere groep is dat beeld precies omgekeerd: daar scoren de vrouwen 20% en de mannen 17%.

— Geslacht

In totaal hebben we bij de mannen 466 lexicale resten geconstateerd (30%). Bij de vrouwen noteren we relatief iets minder, namelijk 559 lexicale taalresten in totaal (27%).

Waar we bij het fonetische gedeelte meer taalresten hebben genoteerd bij vrouwen, ligt de verhouding voor het lexicale gedeelte net omgekeerd: hier scoren mannen drie procent hoger dan vrouwen. Een mogelijke verklaring hiervoor is dat mannen procentueel meer middelbare schoolopleiding hebben genoten dan vrouwen. Immers, respondenten met een middelbare schoolopleiding scoren procentueel hoger op bekendheid met het lexicon dan respondenten met lagere schoolopleiding.

Tevens zou hier weer een mogelijke correlatie kunnen bestaan tussen leeftijd en opleiding in de resultaten. Het is mogelijk dat hoe meer opleiding de respondenten hebben genoten, hoe meer standaard Afrikaans de uitspraak zal zijn, terwijl mensen wellicht deels vanwege hun opleiding bekend zijn met de uitdrukkingen in de idiomen, of eerder de betekenis kunnen herleiden. Een steekhoudende verklaring voor deze resultaten hebben we echter niet.

5.4 Indirecte resultaten uit het onderzoek

Buiten de gegevens waarnaar we hebben gevraagd in de vragenlijsten, heeft het onderzoek een aantal noemenswaardige gegevens opgeleverd zowel op gebied van de uitspraak als op gebied van het woordgebruik. De gegevens zijn verkregen uit de alternatieve antwoorden van de respondenten, die niet in de bovengenoemde resultaten tot uitdrukking komen.

5.4.1 Fonetiek

— *buitekant, buutekant, buidekant*

Vooral in Namaqualand, maar in mindere mate ook in Saron en in de Moslimgemeenschap, treffen we voor *buite* de variant *buitekant* of *buutekant* aan. In Namaqualand komt daarnaast voor: *buidekant* en *buudekant*. De harde, stemloze [t] wordt hier vervangen door de zachte, stemhebbende [d]. Dit verschijnsel van de [d]/[t]-wisseling is al eerder opgemerkt door Links (1989: 22) in de taal van de Kharkams en door Van Schalkwyk (1983: 126) in de taal van de Rehoboth Basters. Hier gaat het echter om het omgekeerde: de stemhebbende [d] wordt in een aantal gevallen vervangen door de stemloze [t]. De vraag is of we in onze voorbeelden niet te maken hebben met gevallen van hypercorrectie: aangezien de respondenten mogelijk weten dat ze vaak ten onrechte een [d] in een [t] veranderen, zijn ze nu bang fouten te maken en spreken een [d] uit waar dit eigenlijk een [t] behoort te zijn. Om deze hypothese te staven is echter nader onderzoek gewenst.

— ontronding van de [œy]

Beide in Saron als in de Moslimgemeenschap treffen we ontronding van de [œy]-klank aan. In Genadendal word de [œy] in een aantal gevallen uitgesproken als [əi]: [brəin], [səikər], [dəim].

— *oomlik, omblik* en *blotvoet*

In Saron (3 maal) en in de Moslimgemeenschap (1 maal) treffen we bij het woord *oomblik* weglating van de [b] aan: *oomlik*. In alle gevallen zijn de respondenten vrouwelijk.

In Namaqualand doet zich nog een ander opmerkelijk taalverschijnsel voor: hier wordt in de woorden *oomblik* en *blootvoet* de lange [o:] verkort tot een [ɔ] (beide 1 keer). Volgens Ponelis (1993: 122) kan dit verschijnsel duiden op

een Nederlandse taalrest. Hij wijst erop dat in sommige Hollandse variaties de oorspronkelijke lange [o:] was behouden (zoals in *woonsdag*), terwijl deze in andere variëteiten is verkort en verhoogd tot [u] (*woensdag*). In het Afrikaans was de lange [o:] verkort voordat hij was verhoogd tot een [u], zoals in *blom* of *genog* in de variëteit van Holland en in het Afrikaans.

— [t]-toevoeging

In alle regio's van ons onderzoeksgebied doet zich het verschijnsel van de [t]-toevoeging voor, met name in het woord *reent* of *reunt*. Zowel Links als Ponelis bespreken dit verschijnsel van de [t]-toevoeging. Links (1989: 25) noemt dit verschijnsel "epentese": "Dié verskynsel, nl. dat 'n eksplosief voor, in die middel of aan die einde van 'n woord gevoeg word, is 'n prominente kenmerk voor Kharkamstaal. Hierdie verskynsel hang saam met 'n historiese neiging wat daar in Afrikaans bestaan. Reeds in Van Riebeeck se dagregister is dit bespeur. Tot vandag toe, byvoorbeeld, het hierdie epentetiese *t* in Afrikaans behoue gebly in woorde soos *geneentheid* en *geleentheid*."

Ponelis (1993: 123) spreekt in dit geval van de "t-paragoge": toevoeging van de [t] in woorden die oorspronkelijk geen [t] kenden, zoals *diakent* of *ervarentheid*. Volgens Ponelis vormt dit verschijnsel een deel van de 17e-eeuwse Hollandse of Amsterdamse spreektaal. Hij merkt hierbij op dat dergelijke kenmerken — die een overeenkomst vormen tussen het Afrikaans en de Nederlandse dialecten — zijn veranderd sinds de vroeg-moderne tijd en dat sommige van hen sindsdien zijn verdwenen. Zo komt de toevoeging van de [t] bij *reën* niet meer officieel voor in het standaard Afrikaans (slechts in tweede instantie), terwijl deze is blijven bestaan in een woord als *ervarentheid*. Het is dus zeer wel mogelijk dat we in het geval van *reent* te maken hebben met een Nederlandse taalrest uit de 17e eeuw.

Opvallend is hier bovendien de [e:]/[ø]-wisseling in het woord *reunt*.

— *stemp/stem*

Voor het woord *seël* treffen we in alle gebieden het alternatief *stemp(s)* of *stem(s)* aan, de verafrikaanste uitspraak van het Engelse *stamp(s)*. Hoewel hier duidelijk sprake is van invloed van het Engels op het lexicon, zoals Links (1989: 68) ook aangeeft bij de bespreking van de Kharkamstaal, is de uitspraak van de oorspronkelijke Engelse woorden hier verafrikaanst; de woorden zijn getransformeerd: *stamps* wordt [stɛmps] of [stɛms].

— *mug*

Een opvallend taalverschijnsel in Saron is de weglating van de [r] in de uitspraak van het woord *murg*. Dit komt hier 16 maal voor, dat is de helft van de gevallen (50%). Desgevraagd spelt een van de respondenten het woord overigens als "m u r g", terwijl in de uitspraak de [r] volledig wegvalt.

5.4.2 Lexicon

— *geontgaan* en *geontskiet*

In Namaqualand treffen we de vormen *geontgaan, geontskiet/geontskoot* en *gevergeet* aan. De *ge-* voor het voltooid deelwoord in deze specifieke voorbeelden is niet algemeen gebruikelijk in het standaard Afrikaans en kan duiden op een Nederlandse taalrest. Volgens Links (1989: 38), die de toevoeging *ge-* ook heeft aangetroffen in de Kharkamstaal, komt deze toevoeging wel voor in het standaard Afrikaans, maar niet zo frequent. Dit taalverschijnsel is volgens hem te verklaren doordat men juist in het Middelnederlands deze vorm al aantrof. Er zou hier dus sprake kunnen zijn van een overblijfsel uit het Middelnederlands.

— *hy* in plaats van *sy*

Een informant uit Namaqualand gebruikt het mannelijk persoonlijk voornaamwoord *hy* als hij verwijst naar een vrouw, in de uitspraak: "Hy het brod geëet" (zie hier overigens ook weer de verkorting van de [o:]). Normaliter wordt in het Afrikaans *sy* gebruikt om een vrouwelijk zelfstandig naamwoord te vervangen. Links (1989: 78) heeft ook in de Kharkamstaal opgemerkt dat het anaforisch verband tussen het geslacht van het persoonlijk voornaamwoord en het zelfstandig naamwoord frequent wordt losgelaten en dat het mannelijk voornaamwoord vaak het vrouwelijk voornaamwoord verdringt.

Of we hier te maken hebben met een Nederlandse taalrest is onduidelijk. Wel is het zo dat in een aantal Nederlandse dialecten (bijvoorbeeld in Noord-Brabant) nog steeds *hij* wordt gebruikt als persoonlijk voornaamwoord dat verwijst naar een vrouwelijk zelfstandig naamwoord. Deze vorm zou ook een overblijfsel uit de 17e eeuw kunnen zijn, waar er geen onderscheid werd gemaakt tussen *hy* en *sy*, tussen de mannelijke en de vrouwelijke vorm.

— *wedeman*

In plaats van *wewenaar* noteren we in Namaqualand (in Buffelsrivier) het woord *wedeman*. Links (1989: 60) heeft in Garies (eveneens Namaqualand) het woord *wewevrou* gevonden. Deze variant bestaat nog in Nederlandse dialecten en zou een taalrest kunnen zijn. Mogelijk is dit ook het geval bij *wedeman*, maar hiervoor hebben we geen evidentie.

6. Slot

De grootste bate van dit soort onderzoek voor de lexicografie ligt misschien niet zozeer in het materiaal zelf als wel in het besef dat er talloze woorden, zegswijzen en idiomen bestaan, die nog steeds gehandhaafd blijven en nog niet zijn opgetekend. Deze variëteiten horen thuis in een omvattend woordenboek als het *WAT*, dat in zijn opnamebeleid niet zozeer gericht is op het opnemen van gespecialiseerde vaktaal maar zich veeleer wil concentreren op spreek- en streektaal.

Het is hierom van belang dat er een breder netwerk wordt gevestigd over het hele land, om aandacht te geven aan het aanwezige materiaal. Gebleken is uit het onderzoek dat er geestdrift bestaat voor deelname aan dit soort projecten. Wellicht is dat niet verwonderlijk, het gaat immers om het optekenen van de taal die nog leeft bij de respondenten, de sprekers. De vraag is echter: voor hoe lang nog?

Taalverandering is een algemeen verschijnsel. Het proces van standaardisering van het Afrikaans, dat onder andere de verdwijning van Nederlandse taalresten uit het Afrikaans insluit, is steeds aan de gang. In sommige variëteiten van het Afrikaans, zoals uit ons onderzoek blijkt, is dit proces verder gevorderd dan bij andere variëteiten. Als deze standaardisering voortduurt, zal dit op een of ander moment voltrokken zijn.

Een vraag die na afloop van het project naar voren is gekomen: wat is de stand van Nederlandse resten in andere geïsoleerde Afrikaanse taalgemeenschappen. Alhoewel er al redelijk veel studies geschreven zijn over de taal van de Rehoboth Basters en de Griqua's, kan er nog heel wat meer onderzoek worden gedaan. Zo liggen het Richtersveld en meer afgelegen zendingsposten nog braak. Het taalgebruik van deze gemeenschappen vraagt om nader onderzoek, voordat ook deze variëtieten verder gestandaardiseerd raken en voor de wetenschap en het nageslacht verloren gaan.

Met dank aan Joep Kruijsen en Dirk van Schalkwyk

Noten

1. Joep Kruijsen heeft dit tevens besproken in zijn lezing over de start van dit project, in april 1998 in Nijmegen.
2. In tegenstelling tot de vragenlijsten is voor dit narratieve gedeelte gewerkt met opnameapparatuur. De opnames zijn in bezit van het Buro van die WAT.
3. In de veronderstelling dat juist de oudere generatie nog bekend is met Nederlandse taalresten, hebben we in de bespreking van de resultaten onderscheid gemaakt in leeftijd. De leeftijdsgrens hebben we gesteld op 80 jaar. Men vergelijke hier ook Rademeyer (1938: 49): uuuitspraken zijn onder oudere geslachten nog zo goed als algemeen, terwijl ze minder gewoon zijn onder jongeren.
4. Bij het lezen van de resultaten dient in acht genomen te worden dat niet alle onderdelen altijd volledig zijn ingevuld op de lijsten. Bovendien hebben mensen soms een geheel andere uitspraak aangegeven dan de mogelijkheden op de vragenlijst.

Literatuur

Botha, R.P., G. Kroes en C.H. Winckler. 1994. *Afrikaanse idiome en ander vaste uitdrukkings*. Halfweghuis: Southern Boekuitgewers.

Kloeke, G.G. 1950. *Herkomst en groei van het Afrikaans*. Leiden: Universitaire pers Leiden.

Labov, W. 1972. *Sociolinguistic Patterns*. Oxford: Basil Blackwell.

Links, T. 1989. *So praat ons Namakwalanders*. Kaapstad: Tafelberg-Uitgewers.

Ponelis, F. 1993. *The Development of Afrikaans*. Frankfurt am Main: Peter Lang.

Rademeyer, J.H. 1938. *Kleurling-Afrikaans. Die taal van die Griekwas en Rehoboth-Basters*. Amsterdam: N.V. Swets & Zeitlinger.

Schönfeld, M. 1970[8]. *Historische grammatica van het Nederlands*. Zutphen: W.J. Thieme.

Stassen, A. 1975[4]. *Zeventiende eeuwse teksten*. Groningen: Tjeenk Willink.

Van Schalkwyk, D.J. 1983. *Fonetiese variasie in die taal van die Rehoboth-Basters*. Ongepubliceerd D.Litt.-proefschrift. Johannesburg: Randse Afrikaanse Universiteit.

Weijnen, A. 1966[2]. *Nederlandse dialectkunde*. Assen: Van Gorcum/Prakke.

Weijnen, A. 1991. *Vergelijkende klankleer van de Nederlandse dialecten*. 's-Gravenhage: SDU.

Lexicography, Terminography and Copyright

Mariëtta Alberts and Michiel Jooste
National Terminology Services, Pretoria, South Africa

Abstract: The focus of this article is on copyright issues with specific reference to lexicography and terminography. Lexicographers and terminographers are in the peculiar position of being both creators of copyrightable products and users of copyrighted products. An inventory of accrued rights, the nature of dictionaries as subjects of copyright, national laws and international conventions, terminographical and lexicographical practice, the copyright status of dictionary elements, as well as infringement pitfalls, is made in order to propose guidelines on the legal position of the compilation and publishing of dictionaries. Electronic publications and dissemination on the Internet is considered and discussed, and contractual agreements protecting mutual rights is offered as a final conclusion.

Keywords: AUTHOR'S RIGHT (COPYRIGHT), COPYRIGHT (AUTHOR'S RIGHT), COPYRIGHT INFRINGEMENT, COPYRIGHT ISSUE, COPYRIGHT LAW, COPYRIGHTABLE PRODUCT, COPYRIGHTED PRODUCT, DATABASE STORAGE SYSTEM, DENOMINATOR, ECONOMIC RIGHT, ELECTRONIC COMMUNICATION NETWORK, FAIR USE, INFRINGEMENT, INTELLECTUAL PROPERTY, INTELLECTUAL PROPERTY RIGHT, LEXICOGRAPHER, LEXICOGRAPHY, MACROSTRUCTURE, MICROSTRUCTURE, MORAL RIGHT, TANGIBLE MEDIUM, TERMINOGRAPHER, TERMINOGRAPHY, TERMINOLOGIST, TERMINOLOGY

Opsomming: Leksikografie, terminografie en outeursreg. In hierdie artikel word gefokus op outeursregkwessies met spesifieke verwysing na die leksikografie en terminografie. Leksikograwe en terminograwe bevind hulle in 'n vreemde situasie deurdat hulle sowel skeppers van outeursregbare produkte is as gebruikers van outeursberegte produkte. 'n Inventaris word opgestel van toegevalle regte, die aard van woordeboeke as onderworpe aan outeursreg, nasionale wette en internasionale konvensies, terminografiese en leksikografiese praktyk, die outeursregstatus van woordeboekelemente, asook van slaggate rakende outeursregskending ten einde riglyne vir die regsposisie van die samestelling en publikasie van woordeboeke voor te stel. Elektroniese publikasies en verspreiding op die Internet word oorweeg en bespreek, en ten slotte word kontraktuele ooreenkomste wat wedersydse regte beskerm, geopper.

Sleutelwoorde: OUTEURSREG (KOPIEREG), KOPIEREG (OUTEURSREG), OUTEURSREGSKENDING, OUTEURSREGKWESSIE (OUTEURSREGVRAAGSTUK), OUTEURSREGWET, OUTEURSREGBARE PRODUK, OUTEURSBEREGTE PRODUK, DATABASISBERGSTELSEL, AANDUIDER, EKONOMIESE REG, ELEKTRONIESE KOMMUNIKASIENETWERK, BILLIKE GEBRUIK, SKENDING, INTELLEKTUELE EIENDOM, INTELLEKTUELE EIENDOMSREG, LEKSIKOGRAAF, LEKSIKOGRAFIE, MAKROSTRUKTUUR, MIKROSTRUKTUUR, MORELE REG,

TASBARE MEDIUM, TERMINOGRAAF, TERMINOGRAFIE, TERMINOLOOG, TERMINOLOGIE

1. Introduction

Information practitioners around the globe are confronted with several issues regarding copyright and the fruits of their labour. The current worldwide discussion of intellectual property rights has been prompted by a shift from fairly clear and predictable copyright laws with regard to printed works, to legal uncertainty in electronic communication networks and database storage systems. The number of recently held seminars, congresses, workshops and published articles on various issues relating to copyright are proof of this.

The focus of this article is on copyright issues with specific reference to lexicography and terminography, involving lexicographers and terminographers as being both creators of copyrightable products, and users of copyrighted products. Two angles of approach are inherently called for in order to determine the legal position of the lexicographer/terminographer in this dual capacity, these angles being best illustrated by the following two questions:

— What rights do lexicographers/terminographers accrue when compiling and publishing dictionaries?
— What may lexicographers/terminographers do when compiling and publishing dictionaries without infringing copyright?

During the course of research for this article, it became clear that general, broad and very often technically vague guidelines regarding the lexicographer's/ terminographer's legal position are abundant, being defined in terms of national copyright laws and statutes (for example the South African *Copyright Act 98 of 1978)*, international agreements on intellectual property rights (for example the *Berne Convention of 1979)* and deductions made from papers presented at various conferences (for example the *Report of the Working Group on Intellectual Property Rights of 1995).*

Language practitioners are positioned between the highly technical rules of copyright law on the one hand and the common practice of their trade of reproducing, quoting and copying from sources on the other hand:

> It is highly unlikely that even the fine-tuned expert definitions found in national and international standards would qualify as ... unique expression. Standard definitions are frequently reused in other standards, in general and technical texts, and in terminology databases ... (Wright 1996: 2)

The principles of "fair use", "right of recognition of authorship", "infringement"

and the multitude of legal formulae and legal terminology governing copyright do not spell out in clear and understandable language what the lexicographer/terminographer should or should not do to prevent infringing or being infringed upon.

To avoid being yet another theoretical voice in the discourse, this article will try to cover several practical problems that lexicographers/terminographers may encounter in the course of compiling a dictionary. These problems stem from the very nature of a dictionary as a specific kind of intellectual property type, because of aspects such as typology, structure and tangible manifestation of the content, the nature of the compilation process — in fact, the essentialia of what distinguishes a dictionary from other printed matter.

It is not the purpose of this article to come up with any revolutionary findings, but rather to make an inventory of the current state of affairs by providing suggestions for specific problems, as substantiated by case law, proceedings of seminars and workshops and articles published in magazines. We hope thereby to illustrate what is permissible within the scope of copyright coverage for lexicographers/terminographers with regard to their products. However, it must be made clear from the start that where copyright is concerned, there are no hard and fast rules as every case will have to be decided by a court of law.

2. What is copyright?

Copyright is, giving a very broad definition, "the right that gives an author or any other entitled person, the sole right to commit certain acts regarding intellectual property of their own creation, especially acts regarding the duplication thereof" (Copeling 1978: 77). Ownership is granted only once the content has been made material by putting it in a tangible medium. The rights accrued by this ownership must be and are protected by law (Copeling 1978: 93).

In South Africa for example, the *Copyright Act 98 of 1978* (as amended), and its regulations, governs this field. Copyright is also territorial (Wright 1996: 2). This means that copyright law technically only extends as far as the law permits it to, and as far as the territory permits the law to rule. In other words, copyright law as it exists in South Africa, is unique to South Africa, since only South Africa is governed by this statute. However, since published works are being used all over the world, there is a need to protect the rights of authors across national borders.

Intellectual property was only legally recognised as a theory in England in 1709 (Galinski 1996: 7). As the concept of intellectual property developed, various agreements were called upon to regulate universal copyright requirements. The *Berne Convention for the Protection of Works of Literature and Art*, completed in 1896, was the first major international agreement on authors' rights. All countries signing the accord are bound to the provisions of that agreement, in order to establish mutual enforcement of copyright between member countries.

The most important convention that binds us today, is still the Berne Convention, as ratified in Paris on 24 July 1971. For political reasons South Africa was not a party to this convention in 1971. Since South Africa was, however, party to the Berne Convention as revised in Brussels in 1948, the provisions of the Paris text had been administratively ratified. This means that we are bound to the provisions of the latter, since South Africa remained a member of the convention and did not indicate otherwise (Copeling 1978: 9).

3. The Berne Convention

The Berne Convention makes provision for the protection of authors' moral and economic rights across national borders. The following provisions are of special interest for the aim of this article:

Article 2 (3)

Translations, adaptations, arrangements ... and other alterations of a literary or artistic work shall be protected as original works without prejudice to the copyright in the original work.

Article 2 (5)

Collections of literary or artistic works ... which, by reason of the selection and arrangement of their contents, constitute intellectual creations, shall be protected as such, without prejudice to the copyright in each of the works forming part of such collections.

Article 3 (3)

The expression "published works" means works published with the consent of their authors, whatever may be the means of manufacture of the copies, provided that the availability of such copies has been such as to satisfy the reasonable requirements of the public, having regard to the nature of the work.

Article 6^{bis} (1) & (2)

Independently of the author's economic rights, and even after the transfer of the said rights, the author shall have the right to claim authorship of the work and to object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his honour or reputation. The rights granted to the author ... shall, after his death, be maintained, at least until the expiry of the economic rights.

Article 8

Authors of literary works protected by this Convention shall enjoy the exclusive right of making and of authorizing the translation of their works throughout the term of protection of their rights in the original works.

Article 9 (1)

Authors of literary and artistic works shall have the exclusive right of authorizing the reproduction of these works, in any manner or form.

Article 9 (2)

It shall be a matter for legislation ... to permit the reproduction of such works in certain special cases, provided that such reproduction does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author.

Article 10 (1)

It shall be permissible to make quotations from a work which has already been lawfully made available to the public, provided that their making is compatible with fair practice, and their extent does not exceed that justified by the purpose ...

Article 12

Authors ... shall enjoy the exclusive right of authorizing adaptations, arrangements and other alterations of their works.

4. Common copyright denominators[1]

Although national laws of countries may show differences, certain common denominators, resulting from said agreements such as the Berne Convention,

1 These denominators are given as a summary and have been obtained from the following sources: Wright 1996; Galinski 1996; Felber 1986; Templeton 1996; South African *Copyright Act 98 of 1978* (as amended); *Report of the AAU Task Force on Intellectual Property Rights in an Electronic Environment* 1994; United States *Copyright Protection Act of 1988*; *General Agreement on Tariffs and Trade (GATT)* 1994; *Grünbuch der Kommission der Europäischen Gemeinschaften* 1995; correspondence with the Chief Editor: Macquarie Dictionary (Australia); the Registrar: Patents, Hallmarks, Authors' Rights and Models (South Africa); and the Publishing Consultant: South African Press Association (South Africa).

common law and civil law are distinguishable. By comparing these, it is thus possible to construe a "universal copyright law" that can be assumed to be applicable irrespective of where a work was copyrighted. These denominators should be acknowledged, developed and nurtured. Once these denominators are globally accepted, they can be applied to the fields of lexicography and terminography in order to establish a standardised lexicographical code of conduct. The latter should promote the reusability and dissemination of information, but at the same time protect authors' rights in a clear and universally recognised set of rules.

4.1 Requirements for copyright to exist

Ideas and information themselves are not protected by copyright. Ideas and information must be concretised, in other words, fixed in a tangible medium, consisting of both content and internal and external form. This would include any carrier or embodiment of the work, such as printed matter, electronic media and audiovisual media, amongst others.

Once an idea is concretised, copyright naturally exists on the product, *and no explicit indication that copyright exists is required.* The work must however be original in character, with character referring both to the original expression and to the arrangement of the knowledge in the work. It should also be noted that originality is defined in very broad terms, requiring only that the work emanates from the author and is not copied — it is thus only applicable to original skill or labour in execution, and not original thought.

4.2 Subjects of copyright

In South Africa, the Copyright Act divides subjects into two broad categories, namely "works that traditionally (are) the subject of copyright" and "works regarded as a medium of communication". Works that traditionally are the subject of copyright may include literary, artistic and musical works. Literary works include, irrespective of literary quality (as long as they are only "written") the following:

- novels, stories and poetical works
- dramatic works, stage directions, film scenarios and broadcasting scripts
- textbooks, treatises, histories, biographies, essays and articles
- encyclopaedias and dictionaries
- letters, reports and memoranda
- lectures, speeches and sermons
- tables and compilations

It should be noted that the act is not very specific in distinguishing between these various types of literary products. These categories should be interpreted very broadly. One can distinguish between works from all areas of literature, the sciences, practical daily life, as well as adaptations, derivative works, translations and collections.

4.3 Authors' rights

Once copyrighted, certain authors' rights are created and include:

(a) Economic rights

— the right of publication
— derivative rights
— the right of use
— the right of access to the original or duplication master
— the right to claim compensation for the licensing of duplication for commercial purposes
— the right to transfer to a third party

(b) Moral rights

— the right of recognition of authorship
— the right to prevent misrepresentation or unauthorised modification

4.4 Infringement on copyright

Infringement occurs when somebody commits an act that is the sole prerogative of the copyright holder without the permission of the copyright holder. Acts of infringement are the translation, reproduction, publishing, performing, broadcasting or adapting of a literary work in any manner or form, without the consent of the copyright holder. Actual copying of the work, or a substantial portion thereof, must take place. A substantial portion depends both on how much of the work is copied and on the quality of the portions copied, the latter being of greater importance than the former.

4.5 Defences to infringement

The South African Act 98 of 1978 makes provision for certain general exceptions provided that there is "fair dealing" such as the inclusion of short excerpts from a copyrighted work in another work for the purposes of criticism or

review, or reporting of current events, provided that the extent of the excerpts shall not exceed the extent justified by the purpose and that the source shall be mentioned, as well as the name of the author if it appears on the work. Proper citation is called for, because reuse without proper citation constitutes plagiarism (Wright 1996: 2). Copyright requires a work to be the original thought and expression of the author, which can be difficult to express in very short excerpts such as dictionary entries. Wright (1996: 2), as already mentioned, argues that "standard definitions are frequently re-used in other standards, in general and technical texts, and in terminology databases with attribution under the provisions of fair use".

When the right of the original owner to the exploitation of the work is compromised by excessive quotation or reuse of copyrighted material, fair dealing becomes questionable. This would definitely be the case if an entire standard was rearranged and incorporated into a terminological collection (see par. 6 below). Wright (1996: 2) further points out that fair dealing would not be questioned when "the extraction of a subset of terms from standard or other works (involve) the inclusion of random texts and definitions documenting the affected concepts such that the resulting new material does not compromise a substantial percentage of the original and the arrangement is significantly different". Since both lexicographers and terminographers are highly dependent on the use of sources to compile dictionaries, the proper incorporation of these sources is of utmost importance. This practice is however easier to apply in lexicography than terminography, as will be explained.

5. Terminography and lexicography

In studying various articles on what is copyrightable and what not, it seems that copyright may be added as yet another category to the list of differences between terminography and lexicography. It is acknowledged that no statute or convention makes any explicit distinction between lexicographical and terminographical collections where copyright is concerned. However, a thorough comparison proves that the coverage of fair use definitely differs in applicability. One may find that the same act of incorporating material from another work is rebutted or proved as infringement in a court of law, depending on the reuse of a work for lexicographical or terminographical purposes.

If one looks at what is not copyrightable, some interesting conclusions can be made. Only original works of authorship are protected, not mere original thought. Trittipo (1996: 369) states that "ideas and facts are true regardless whether any person knows them or not, and they owe their origin to the way the world is, not to any author". One also needs to consider the idea–expression dichotomy which amounts to a rule that if the "idea" and its "expression" is inseparable, copying the expression will not be barred, since protecting the "expression" in such circumstances would confer a monopoly of the "idea" on

the copyright owner (Trittipo 1996: 369).

In the same way, individual words in general language and most terms in special languages are the common property of all speakers, since the idea cannot be separated from the way it is expressed in a single word. In terminology, lengthy phrases are also considered to be terms and cannot be expressed in any other way but in the acknowledged phrase as construed by terminologists and subject specialists. Wright (1996: 2) states that "terminological principles require substantial supportive material in terminological entries in the form of definitions and contextual references". Ideally, these materials should not be original, but rather taken from authoritative sources. Thus, each entry is very likely to contain one, or in many cases, numerous references taken from published, propriety or even standardised works (Wright 1996: 3).

Because terminology deals with the exact coining of concepts, these entries should not be modified to circumvent copyright ownership of the source material, since it would be unethical and terminographically wrong. Terminologists want the word-for-word rendering of standardised definitions and live contexts. It is obvious that within a specific subject field these terms, definitions and contexts are (and are supposed to be) the same. There is no way that these concepts should be allowed to be expressed differently, otherwise it would defy the whole purpose of terminology, namely that of exact and precise communication with a standardised technical vocabulary (terminology) within a specific subject field. An example of this would be the following nuclear term:

> **curie** a unit of radioactivity ... defined as the quantity of any radioactive nuclide in which the number of disintegrations per second is 37.00×10^9. (Jerrard-McNeill 1972[3]: 32)

Two obvious problems are apparent at this stage: terminology *per se* is not copyrightable and the reuse of terminology entries is inevitable. How does one provide a dictionary of, for example, nuclear terms in several languages, and who owns what, if all the terms and definitions thereof are universally descriptive of the same concepts in the subject field — irrespective of language used?

Trittipo (1996: 368) argues that "mere lists such as for parts and associated parts numbers, may not be copyrightable ... when the selection of terms on the list is determined mainly by facts or user expectations, rather than by the list maker's discretion, since a list has to list every relevant or replaceable part". He argues further that where there is no real choice of what to include and select, no copyright exists on such a document. If selection and choice depend on the discretion of the author, copyright does exist. Common terminographical practice frequently involves mere lists of terms in subject domains in various languages. According to this principle, no copyright exists on such lists of terms.

This statement is supported by the decision in *Feist Publications, Inc. vs. Rural Telephone Service Co., Inc.*, 499 US 340 (1991) which ruled that databases that consist of purely factual information (such as telephone directories) cannot

be copyrighted. However, it stated *obiter dicta* that databases that select, organise or arrange these facts in a certain manner (e.g. by using terminographical principles in order to compile a list of bookbinding terms), are copyrightable. Theoretical literature also suggests that within terminology the "smallest meaningful units" are copyrightable, as well as the mechanisms used for accessing and updating these units. These units or data element categories are described in *ISO 12620.2* and are thus already set as a standard by an authoritative body.

One can therefore safely conclude that copyright does exist in every single separate element in any tangible medium in which terminology is expressed. A practical problem results from this fact: dictionaries and lexicographical/terminographical databases are either derivative works, compilations or translations of terminology already excerpted and recorded in other languages. Since copyright exists on the original dictionary or database and it is expected of a terminologist to provide a word-for-word description of the standardised definition of the concept, conflict of interest is inevitable. Wright (1996: 1) argues that there is no limit to a terminologist's right to report a term or any set of terms in a terminological resource, but that "precise unauthorised reproduction of a given set of terms together with their definitions ... without further value-added information or other modification, would in all likelihood constitute an infringement of copyright".

In subject fields with a large vocabulary it does not necessarily pose a problem, since a variable degree of selection is possible and room is left for choice of selection and arrangement. It should be expected that more than one original dictionary within such a subject field may appear (e.g. a dictionary of Commercial Science). In subject fields with a small vocabulary, the copyright owner has little to no protection, since fair dealing provides that if no room is left for the choice of selection at the discretion of the author, no copyright can exist on such a publication.

One may argue that the selection of terms in a subject field involves original thought in execution and is therefore protected, since selection entails the application of certain terminology skills. However, fair dealing would allow a terminologist to use these terms as well as their definitions to a large extent. The concept and its expression in the code of language are also inseparable and it belongs to the respective discourse community, thus it cannot be copyrighted. In addition to this, terminology practice requires the standardised rendering of conceptual information, limiting original expression to a large extent. One is left with the conclusion that ownership of the published material would only be protected as regards layout, typology and aspects not referring to the content and defining of the terminology (vocabulary). In fact, one may go so far as to say that the reuse of the entire macro- and microstructure of a technical dictionary of a minor language, would still be allowed under current rules.

In lexicography, the theoretical arguments seem to be easier, but the issue of copyright infringement is just as complex. As was said earlier, no copyright exists on general words in a language. However, the countless ways in which

one can present, select and arrange these words in a dictionary, ensures that every dictionary can and should be an original work that is protected when it is put on paper (or whatever medium is chosen by the author).

As a result, it would not be fair and reasonable to take the macrostructure of someone else's dictionary and use it as it is, even if one provides one's own microstructure. The headword list, layout of an entry, components of definitions, etc. are essential to a particular dictionary and cannot be copied to another dictionary without infringing copyright (*RJ Romme vs. Van Dale Lexicographie*, 1994). This principle would be applicable to a single dictionary entry too, where room for choice is left in defining and explaining a lemma. If certain definitions seem to correlate or were copied, an infringement defence may be that there can only be so much room to move within a definition. Copyright does cover all the separate dictionary elements, but one cannot prove that someone has infringed copyright if one finds, for example, one illustrative sentence that is the same. One needs to demonstrate a pattern of borrowing.[2]

6. Derivative works and compilations

Both lexicography and terminography depend on other sources for its content. If a completely original work (for example a list of terms compiled by the developers of new technologies requiring neologisms) is not created, dictionaries and products of lexicography and terminology can be divided into derivative works and compilations.

The US *Copyright Protection Act of 1988* defines a derivative work as a work "based upon one or more pre-existing works, such as a translation ... dramatisation, fictionalisation, motion picture version, sound recording ... abridgement, condensation, or any other form in which a work may be recast, transformed or adapted". The Berne Convention also requires that member countries accord to authors the exclusive right of translation (Article 8). The rights to make and to authorise the making of derivative works (such as translations) are among the exclusive rights of an author. According to Trittipo (1996: 369), a derivative work that is created without the authorisation of the original's copyright holder, and without some defence such as fair dealing, is generally not entitled to copyright protection. It is however curious to note that section 2 (3) of the South African *Copyright Act 98 of 1978* states that "a work shall not be ineligible for copyright by reason only that the making of the work ... involved an infringement of copyright in some other work". Thus, an unauthorised transformation of a protected work is still protected under South African copyright

2 Mention must here be made of the advice of Ms S. Butler from the Macquarie Dictionary, Australia, in this regard.

law, and may not be infringed upon by someone else.

An unauthorised translation, transformation or adaptation of an original work should be strongly discouraged, and we are of the opinion that the South African copyright law is lacking in that it does not adequately protect authors of original works by granting protection to such unauthorised adaptations. If there is no copyright on the infringing work, it means that anyone can copy it lawfully, to the prejudice of the author of the original work. This state of affairs is also in contrast with the spirit of Article 8 of the Berne Convention which grants authors the exclusive right of translation and the authorisation thereof.

Once permission is obtained from an author to make a translation of a work, this translation is original because there is significant room for choice in how to translate most things. As Trittipo (1996: 370) states: "a translator's goal is to duplicate an original's meaning ... but a translator also seeks to duplicate tone and feeling, and these are matters of expression".

The majority of dictionaries are compiled solely for the purpose of providing translated equivalents in a target language from words or terms in a pre-existing vocabulary in a source language. All bi- and multilingual dictionaries fall into this category. If a translation of a lexicographical or terminological collection is made without the permission of the original author, the copyright in the translation should vest in the author of the original work, and not in the translation. Since the South African copyright law indicates the contrary, this argument bears no significance and will not be referred to again. In general, once permission has been obtained, the translator becomes the sole author of that product, with the same degree of authors' rights and copyright protection as any author of an original protected work. This rule may however be altered by, and is subject to contractual arrangements made between the various parties involved in the process (Norman 1994: 460).

In the light of the above-mentioned discussion on the unique nature of terminologies in certain subject fields, it appears that a substantial amount of translation of terminologies may be allowed under the principle of fair dealing, since terminology defines concepts (which cannot be protected) and may consist of very small or technically specific vocabularies. Furthermore, terminology practice implies the excerpting of terms, the coining of terms and term equivalents, and the naming and defining of concepts. Terminologists do not translate. Terminologists are always supposed to use the concept as basis and then transfer the exact meaning of that concept into the chosen target language (Cluver 1992: 35). Since a terminologist hardly ever has to express "tone and feeling", but rather provide the exact meaning, it is doubtful that terminologists are subject to authorisation in order to compile their technical dictionaries. One would however infringe copyright if mere translations of terms (and no coining of terms for concepts) are provided, or if mere translations of definitions are given where room for choice of original expression is possible, but ignored. With regard to lexicography, it is obvious that general language dictionaries may not be translated without authorisation from the owner of the source.

Most terminological and lexicographical collections can be viewed as compilations (Wright 1996: 2) which can be defined as "formed by the collection and assembling of pre-existing materials or data that are selected, co-ordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship" (US *Copyright Act Protection of 1988*). Creators of terminological and lexicographical collections can therefore claim copyright protection for their works, taking into consideration the above-mentioned restrictions with regard to the nature of terminological and lexicographical work. Infringement would once again depend on fair dealing. What exactly constitutes fair dealing has been the subject of ongoing debate, and may depend on whether a substantial portion of the source is used in the compilation, whether proper citation is given, whether a pattern of borrowing can be established, etc.

7. Citation

Since general language dictionaries make abundant use of citations in order to explain the context in which words may occur, and technical dictionaries depend on authoritative and standardised sources to capture and define the exact meaning of concepts, both lexicographical and terminological products should pay attention to providing the proper references for these sources. Common practice usually indicates the inclusion of a bibliography of sources as front or back matter. Citations are usually so short that they are considered to be out of copyright. If one needs or wants to use substantial citations, ways of proper references whence the citations are taken, should be adhered to. Usually the citation will be given with the title and date of the source, because this is the important information required to give the person reading the entries access to the complete bibliography. Wright (1995: 256) also supports the idea of a device which refers a reader to a bibliography listing the sources used to compile dictionary entries: "Source identifiers should be short codes that act as pointers that link to targets, i.e. bibliographical entries that occur once in the system. The native procedure for any system is not a serious problem so long as it remains possible to extract bibliographical data and express it as end matter."

8. Copyrightable dictionary elements

The following table may serve as a general guide to determine what elements are copyrightable in lexicographical and terminographical entries:

LEXICOGRAPHY AND TERMINOGRAPHY — GENERAL COPYRIGHT GUIDELINES TO SEPARATE DICTIONARY ENTRIES			
TERMINOLOGY		LEXICOGRAPHY	
COPYRIGHT	NO COPYRIGHT	COPYRIGHT	NO COPYRIGHT
Definitions (if a degree of choice in original expression is available)	All terms	Microstructure	All lemmata
Macrostructure of major languages	Definitions of definition-specific concepts or technically specific languages	Macrostructure	Collocations
Microstructure of major languages	Context indicators of context specific concepts	Definitions	Definitions where no room is left for original expression
Context indicators of major languages	Microstructure of minor languages	Pragmatic information	Grammatical information
Layout	Macrostructure of minor languages	Layout	Fixed expressions
Typology	Translations of terminological vocabularies	Typology	
Example sentences and encyclopaedic information	Grammatical information	Encyclopaedic information and example sentences	

9. The Internet

There are quite a number of on-line dictionaries and terminology databases available on the Internet today. Since copyright and the Internet is a general concern and subject of ongoing debate worldwide, this issue will only be touched on very briefly.

It must be remembered that all works are protected the moment they are written, and no copyright notice is required. Postings (information put on the Internet) are not granted to the public domain, and do not grant any user permission to do further copying except the kind of copying that might be expected in the ordinary flow of the Internet, unless otherwise indicated by the author. If works are used or quoted that are published on the Internet, proper and due reference can be given by mentioning the name of the author and the URL (address of the document), or the e-mail address. If no creation or publishing date is provided in the posting, it is advised that the date of one's using of the posting be given.

It is clear that as far as terminology and lexicography are concerned, existing copyright law also applies to documents on the Internet. As was said ear-

lier, all works should be fixed in a tangible medium before copyright exists in such a publication. This would include any carrier or embodiment of the work, including those published on the Internet. Note that, apart from copying, publishing of another's work as one's own is also an act of copyright infringement.

Thus similarly, terminographical and lexicographical collections will be treated according to the arguments put forward in this article. The extent of protection of a work is not altered by the medium in which it is published. One can therefore not take someone else's dictionary and put it on the Internet without obtaining permission from the author, or create databases through compilation and ignoring the rules of infringement as stated in statutes and legal practice. One cannot typeset a protected work and post it on the Internet (even though one does present it as the work of the author) without the permission of the author, since that would be the same as copying a work and distributing it for all to use and see, resulting in a possible infringement of the author's economic rights. One may argue that one of the main goals of the Internet is to disseminate information as widely and as freely possible. However true this may be, it does not justify the blatant infringing of authors' rights and robbing authors of their livelihood for the convenience of the information society.

The issue of copyright and the Internet is in a very early stage of its evolvement, and it is being worked on and discussed by people all over the world. As far as information on the Internet goes, copyright does exist as explained, but infringements cannot effectively be prevented or sanctioned. It is therefore one's own responsibility to ensure the protection of information granted to the Internet domain, and use the information highway in a climate of mutual trust and ethic responsibility.

Copyright protection is universally recognised as the best form of legal protection for both the old and new categories of works. Copyright therefore still plays a central role in the digital world of computer programs and databases.

GATT, the agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), states clearly that both computer programs, whether in source or object code, and databases are protected by copyright, subject to the fundamental requirement that they be original in the sense that they are the author's own intellectual creation (Gervais 1995: 1).

Article 10 of TRIPS reads as follows:

Computer Programs and Compilations of Data

1. Computer programs, whether in source or object code, shall be protected as literary works under the Berne Convention (1971).
2. Compilations of data or other material, whether in machine readable or other form, which by reason of the selection or arrangement of their contents constitute intellectual creations shall be protected as such. Such protection, which shall not extend to the data or material itself, shall be without prejudice to any copyright subsisting in the data or material itself.

Gervais (1995: 1) clearly states that "what is true of these new categories of works is also true of the more traditional ones, most of which are expressly mentioned in Article 2 of the Berne Convention".

Since the works mentioned in this article will thus continue to be protected by copyright worldwide under the Berne Convention, the TRIPS agreement and the Universal Copyright Convention, one should, according to Gervais (1995: 1), rather determine which rights apply to the information highway (and whether any new rights are needed) and how they are going to be enforced and administered.

Copyright is composed of a bundle of parallel rights and this has implications for copyright on the information highway:

- The right of reproduction authorises or prohibits the making of a copy of a work or a substantial part thereof in any form. Copying was originally intended to cover incidental private copying and "was not conceived as a main mode for disseminating works in the way that is likely to become prevalent on the information highway" (Gervais 1995: 2).
- The right of communication to the public is the act of making a work available in any manner to persons not restricted to specific individuals belonging to the family circle. Any form of transmission of information, whether interactive or not, can be considered as an act of public performance under the Berne Convention (Gervais 1995: 2).
- The (moral) right of integrity recognised by the Berne Convention, as well as the (economic) right of adaptation will be affected when works transmitted over the digital highways are manipulated (changed). Gervais (1995: 2) argues that in cases where changes give rise to a new work, authorisation from the rights holder of the original work is necessary to use the new work.
- Original compilations of works and other data are protected by copyright as explained previously. According to Gervais (1995: 2) this protection still applies when the contents are downloaded even in part via the digital highways. In some cases, copyright protection may be complemented by a *sui generis* right.

Gervais (1995: 2) states that "copyright in its present form covers all forms of exploitation, present and future, on the information highways".

Administration and enforcement of copyright on the information highway

As regards dissemination of protected works by various service providers on the World Wide Web (WWW), these works will have to be administered properly. According to Gervais (1995: 3) database operators will pay and report to collective administration organisations representing rights holders according to negotiated or fixed tariff sister organisations based on the use of data. Central

clearing houses which will have access to information from almost all countries, as well as to specialised licensing sources could search for rights clearance on behalf of the producers of media and perhaps even negotiate on their behalf. The producer need not deal with these matters him-/herself.

The challenge posed to rights holders is to offer a fully digital worldwide administration system for transmission of works to the public and the reproduction of such works. They should also offer single sources for licensing for multimedia productions. This will, according to Gervais (1995: 4) require a uniform subcoding system and global Electronic Data Interchange (EDI) standards.

Gervais (1995: 4) states: "Over the past 109 years (since the inception of the Berne Convention), each technological change (the invention of cinema, of broadcasting, of sound recordings, of computer programs, etc.) has prompted demands to 'start from scratch'. The information superhighways are no exception. Yet in all previous cases, the Convention has continued to apply and will continue to do so. The works involved are protected under the Convention and the national laws of approximately 150 countries. Moreover, the transmission of works over the information highway and their reuse are covered by existing rights. "

10. Concluding remarks

This article aimed to provide language practitioners working in the fields of terminology and lexicography certain guidelines concerning copyright. These should be viewed as mere guidelines, and are not intended to be taken as legal advice. Many bold statements have been made in this article in this regard, but it must be remembered that, since there are no black and white guidelines on copyright as it effects terminography and lexicography, every case brought before a court of law shall be tried on its own merit, and one cannot predict clear verdicts for the multitude of problematical copyright issues that may surface in the course of our work. The best advice for parties working together on projects is to enter into contracts to protect their mutual rights.

Bibliography

Alberts, M. 1990. *'n Bepaling van Afrikaanse vakleksikografiese behoeftes.* Unpublished D. Litt. et Phil. Thesis. Pretoria: UNISA.

Bergenholtz, J. and S. Tarp. 1995. *Manual of Specialised Lexicography.* Amsterdam: John Benjamins.

Cluver, A.D. de V. 1992. *Die verskille en ooreenkomste tussen algemene leksikografie en vakleksikografie in die praktyk.* Pretoria: Nasionale Terminologiediens.

Copeling, A.C.J. 1978. *Copyright and the Act of 1978.* Durban: Butterworths.

Felber, H. 1986. Einige Grundfragen der Terminologiewissenschaft aus der Sicht der allgemeinen Terminologielehre. *Special Language/Fachsprache* 8 (3-4): 110-123.

Galinski, C. 1996. Terminology and Copyright. *TermNet News* 52/53: 7-15.

General Agreement on Tariffs and Trade (GATT). 1994. *Final Act Embodying the Results of the Uruguay Round of Multilateral Trade Negotiations.* Marrakesh: GATT Secretariat.

Gervais, Daniel. 1995. Protection of Copyright on the Information Highway. *TermNet News* 50/51: 1-4.

ISO DIS 12620: 1995. *Terminology-Computer-Applications-Data Categories.* Geneva: ISO.

Jerrard, H.G. and D.B. McNeill. 1972[3]. *A Dictionary of Scientific Units.* London: Chapman and Hall.

Kommission der Europäischen Gemeinschaften. 1995. *Grünbuch. Urheberrecht und verwandte Schutzrechte in der Informationsgesellschaft.* Luxembourg: KEG.

Lehman, B. 1995. The Report on the Working Group on Intellectual Property Rights. *Intellectual Property and the National Information Infrastructure (IITF).* Available as an ASCII file from http://www.iitf.doc.gov or as a PDF file from the U.S. Patent and Trademark Office World-Wide Web site http://www.uspto.gov

Norman, S. 1994. Copyright: Legal Protection of Databases. *IFLA Journal* 20(4): 459-461.

Report of the AAU Task Force on Intellectual Property Rights in an Electronic Environment. 1994. Washington, DC.

RJ Romme vs. Van Dale Lexicografie BV (Hof, Den Haag, April 1, 1993) [1994] NJ 224.

Romeo, James J. 1997. Language and the Internet. *Language International* 9 (1): 20-21.

South African *Copyright Act 98 of 1978* (as amended). Pretoria.

Templeton, B. 1996. *10 Big Myths about Copyright Explained.* http://www.clarinet.com/brad/copymyths.html

Trittipo, M. 1996. A Primer on Translations and Copyright. *Global Vision: Proceedings of the 37th Annual Conference of the American Translators Association. October 30 - November 3, 1996*: 367-371. Colorado Springs.

United States *Copyright Protection Act of 1988.* Washington, DC.

World Intellectual Property Organization (WIPO). 1995. *Berne Convention for the Protection of Literary and Artistic Works.* Paris. *Act of July 24, 1971, as amended on September 28, 1979.* Geneva.

Wright, S.-E. 1995. Copyright Issues affecting Terminology in Electronic Environments. Brunnstein, K. and P.P. Sint (Eds.). 1995. *Intellectual Property Rights and New Technologies: Proceedings of the Knowright '95 Conference*: 253-258. Österreichische Computer Gesellschaft. Vienna/Munich: R. Oldenbourg.

Wright, S.-E. 1996. Intellectual Property Rights and Terminology Management. *TermNet News* 52/53: 1-6.

Consulted

Ms Monica Seeber, Publishing Consultant: South African Press Association (South Africa).

Ms Sue Butler, Chief Editor: Macquarie Dictionary (Australia).

Ms Sue-Ellen Wright, Managing Editor: TermNet News (Austria, USA).

Mr Von Burton Dursham, The Registrar: Patents, Hallmarks, Authors' Rights and Models, Department of Trade and Industry (South Africa).

Using the Predictability Criterion for Selecting Extended Verbs for Shona Dictionaries[1]

Emmanuel Chabata, *ALLEX Project,*
Department of African Languages and Literature,
University of Zimbabwe, Harare, Zimbabwe

Abstract: The paper examines the "predictability criterion", a classificatory tool which is used in selecting affixed word forms for dictionary entries. It focuses on the criterion as it has been used by the African Languages Lexical (ALLEX) Project for selecting extended verbs to enter as headwords in the Project's first monolingual Shona dictionary *Duramazwi ReChiShona*. The article also examines the status of Shona verbal extensions in terms of their semantic input to the verb stems they are attached to. The paper was originally motivated by two observations: (a) that predictability seems to be a matter of degree; and (b) that the predictability criterion tended to be used inconsistently in the selection of extended verbs and senses for *Duramazwi ReChiShona*. An analysis of 412 productively extended verbs that were entered as headwords in *Duramazwi ReChiShona* shows that verbal extensions can bring both predictable and unpredictable senses to the verb stems they are attached to. The paper demonstrates that for an effective use of the predictability criterion for selecting extended verbs for Shona dictionaries, there is need for the lexicographer to have an in-depth understanding of the kinds of semantic movements that are caused when verb stems are extended. It shows the need to view verbal extensions in Shona as derivational morphemes, not inflectional morphemes as some earlier scholars have concluded.

Keywords: DEFINITION, DERIVATIONAL MORPHEME, DICTIONARY, DICTIONARY ENTRY, LEXEME, LEXICOGRAPHY, MORPHOLOGY, PREDICTABILITY CRITERION, SEMANTICS, SHONA, VERB STEM, VERBAL EXTENSION

Opsomming: Die gebruik van die voorspelbaarheidskriterium om uitgebreide werkwoorde te selekteer vir Shonawoordeboeke. Hierdie artikel ondersoek die "voorspelbaarheidskriterium", 'n klassifikasiehulpmiddel wat gebruik word om geaffigeerde woordvorme te selekteer as woordeboekinskrywings. Dit fokus op die kriterium soos dit gebruik is deur die African Language Lexical (ALLEX) Project vir die selektering van uitgebreide werkwoorde as lemmas in die Projek se eerste eentalige Shonawoordeboek *Duramazwi ReChiShona*. In hierdie artikel word die status van Shona se werkwoordelike uitbreidings ondersoek in terme van hul semantiese opname in die werkwoordstamme waarmee hulle verbind is. Die artikel is oorspronklik gemotiveer deur twee waarnemings: (a) dat voorspelbaarheid 'n graadkwessie is; en (b) dat die neiging bestaan het om die voorspelbaarheidskriterium in die seleksie van uitgebreide werkwoorde en betekenisse vir die *Duramazwi ReChiShona* inkonsekwent toe te pas. 'n Ontleding

[1] This paper was presented at the Second International Conference of the African Association for Lexicography, held at the University of Natal, Durban, 14-16 July 1997.

van 412 produktief uitgebreide werkwoorde wat as lemmas in *Duramazwi ReChiShona* opgeneem is, toon dat werkwoordelike uitbreidings sowel voorspelbare as onvoorspelbare betekenisse kan toevoeg tot die werkwoordstamme waarmee hulle verbind is. Die artikel bewys dat dit vir die effektiewe gebruik van die voorspelbaarheidskriterium vir die seleksie van uitgebreide werkwoorde vir Shonawoordeboeke vir die leksikograaf noodsaaklik is om 'n grondige insig te hê in die tipes semantiese verskuiwings wat veroorsaak word deur die uitbreiding van werkwoordstamme. Dit toon die noodsaaklikheid om werkwoordelike uitbreidings in Shona te beskou as afleidingsmorfeme en nie as fleksiemorfeme soos sommige vroeëre vakkundiges besluit het nie.

Sleutelwoorde: DEFINISIE, AFLEIDINGSMORFEEM, WOORDEBOEK, WOORDEBOEKINSKRYWING, LEKSEEM, LEKSIKOGRAFIE, MORFOLOGIE, VOORSPELBAARHEIDSKRITERIUM, SEMANTIEK, SHONA, WERKWOORDSTAM, WERKWOORDELIKE UITBREIDING

1. Introduction

Headword and sense selection is an important stage in dictionary making, for it determines what to include in or exclude from a dictionary, a criterion that influences the usefulness of the dictionary for its target users. The selection process, therefore, needs well-formulated selection principles. The research reported on here is intended to be a contribution towards having clearly-defined principles for the selection of headwords and senses to enter in Shona dictionaries. Towards this goal, the article focuses on the use of the predictability criterion by the ALLEX Project, since, beyond the very general, what it actually refers to has not yet been explicitly explained in the style manuals developed within the Project.

The paper looks at the predictability/unpredictability concepts in general lexicographic practice and discusses the implications of using the predictability criterion for headword and sense selection in Shona lexicography. Specifically, it looks at the use of the predictability criterion for the selection of extended verbs and senses to enter in Shona dictionaries. Attention is given to the application of the criterion to extended verbs since there are contrasting views concerning the semantic input of Shona verbal extensions to verb stems. Earlier scholars treated verbal extensions as if they were inflectional morphemes, thus, as if they add little semantic meaning to base forms. However, this article considers Shona verbal extensions as derivational. Since this is not a traditional treatment of these morphemes, some space is devoted to explaining why these extensions should be considered derivational morphemes. The article also examines the typical kinds of semantic shift caused by the addition of verbal extensions to base forms. The semantic movements discussed are those that were discovered in the analysis of 412 extended verbs defined in *Duramazwi ReChiShona*. In the subsequent sections, therefore, the focus is on the use of the predictability criterion in the general lexicographic context, its use in the context of Shona lexicography, the nature of Shona verbal extensions and the semantic divergences caused by addition of verbal extensions to the 412 verb stems examined in the study.

2. The predictability criterion in the general lexicographic context

The predictability criterion has a long and productive tradition of use in lexicography and is one that has been recommended by a number of lexicographers and lexicologists (for example, Zgusta 1971, Landau 1984, Jackson 1988 and Svensén 1993). According to *The Oxford English Dictionary* (2nd ed.), something that is predictable "is capable of being predicted or foretold". This follows its definition of the verb "predict" which reads "of a theory, observation etc.: to have as a deducible or inferable consequence; to imply". The predictability criterion has been used in compiling dictionaries for many languages, including Shona, mainly as a means for selecting derived word forms for dictionary entries. According to this criterion, if the meaning of a derived word form can easily be traced back to the meaning of the base form, it would be considered predictable and, as a result, excluded from the dictionary. On the other hand, if the meaning of the derived form cannot be traced, or is difficult to trace, from the base form, it would be considered unpredictable and would, therefore, be entered and defined as a headword. Zgusta (1971: 242), for example, argues that if there is a category of words constituted by a uniform derivation, for instance by the same suffix, and if the membership of this class is quite open, that is, if new members of the class is commonly produced and are easily understood, and if the semantic effect of the derivational process is as uniform as its form, it is not necessary to indicate in the dictionary all known members of the class. In this case, the lexicographer would enter and define the derivational morphemes, for example, highly productive prefixes or suffixes, where he/she feels it to be impossible to include all the instances where the prefix or suffix would occur. A note for the dictionary user would then be given to inform him/her of the meanings these morphemes would add to base forms.

From this perspective, it can be noted that if we use the predictability criterion for headword and sense selection, a derived lexeme can be omitted if both its form and meaning are regular and predictable from the derivative formula. Lyons (1977: 515) argues that if we think of the lexicon/dictionary as an appendix to the grammar, and, if we assume, moreover, that we are able to find the main lexical entry for each lexeme indexed by means of its citation form (which may or may not be a stem from which we can generate all the other forms), there is no need to contain these "purely morphological lexical entries". In fact, Lyons (1977: 515) regards these kinds of lexical entries as being "theoretically redundant".

However, as noted by Zgusta (1971: 242), when the predictability criterion is used, it is necessary for the lexicographer to study all the words in which the particular derivational morpheme occurs in order to see whether the semantic effect described by its "summary indication", that is, its definition, is really identical in all cases. He argues that the lexicographer is obliged to check every member of the category which would otherwise be eligible for selection in order to see whether some of the members do not have semantic "specialities" of their own, not shared by the other members of the class. In this regard Zgusta recommends that when the predictability criterion is used, any word

that shows a semantic speciality, should be entered in the dictionary unless it is eliminated for other reasons such as rareness or obsoleteness.

Like Lyons (1977: 515), Zgusta (1971: 128) also notes that many dictionaries do not list all the words that might be derived from base forms in regular or predictable ways. He is, however, quick to note that derivation is not perfectly regular or uniform in all cases. He points out that the problem is that the difference between the base and the derived form is sometimes great, whilst some meaning differences are not so great, but still observable. He (1971: 129) argues:

> The greater the number of words in which the same derivational morpheme causes the same change of the lexical meaning, the smaller will be the inclination of the lexicographer to list all these words ... If a derivational morpheme is not frequent and/or if its modifying effect on the lexical meaning is far from uniform, the similarity to a grammatical function will be incomparably smaller and the lexicographer will be more inclined to indicate the respective words as separate items.

Swanson (1967: 64), however, is of the opinion that the predictability criterion, with its suggestions to make some derivational morphemes dictionary entries while eliminating derived forms as headwords or run-ons, can only be used efficiently and effectively by linguists. He argues that the use of this criterion would make access to headwords more difficult for readers not trained in linguistics. He also argues that the dictionary user cannot be expected to know how to identify morphemes that constitute derived word forms, since words are traditionally set off in the orthography by spaces. Swanson's view is derived from his observations of monolingual speakers and readers of English, a language whose spelling conventions favour a disjunctive system of word division. The orthographic word in the language under study, Shona, is conjunctively spelt in such a way that large amounts of grammatical and derivational information is put together in word form. An example is **vachazoonana** (they will eventually see one another), where the subject marker (**va-**), future (**-cha-**), auxiliary (**-zo-**), verb stem (**-ona**) and the reciprocal extension (**-an-**) are combined to form a single word. The marking of morpheme boundaries is problematic in Shona because of this conjunctive system. Swanson's view was confirmed for Shona by field research on extended verbs (reported in Chabata 1997) which showed that most tertiary students studying Shona structure did not mark morpheme boundaries correctly, nor could they always distinguish between the base form and verbal extension, knowledge that is needed in order to be able to add the meanings of the respective morphemes and to get the predictably extended meaning.

3. The predictability criterion in the context of Shona lexicography

The lexical tradition of Shona shows that, at least for Hannan (1959), extended forms, predictable or not, were selected as headwords. This however seems to be different for Dale (1981: viii) who claims to have used the predictability cri-

terion for selecting extended verbs. For *Duramazwi ReChiShona* (Chimhundu 1996), a monolingual general Shona dictionary which targets O-level students and contains only 15 828 headwords, the predictability criterion was used for selecting derived word forms, including extended verbs. The idea was that unlike bilingual Shona-English dictionaries like Hannan's *Standard Shona Dictionary* which was meant mainly for second-language Shona speakers, *Duramazwi ReChiShona* was being developed for first-language users who were expected to easily understand predictably extended meanings of derived words. Thus in the process of headword selection, a decision was made that for certain categories of headwords, including extended verbs, only those which were commonly used and which had "unpredictable" meanings would be defined (Chimhundu 1992: 30). It was also decided that verbal extensions would be included as headwords, and that these would be defined. This was actually done in summary form in the front matter of *Duramazwi ReChiShona* (1996: xxiii) and in the body of the dictionary as well. The assumption was that if the targeted dictionary user studying Shona structure at secondary school wanted to know the meaning of an extended verb, he/she would have the knowledge needed to "add up" the meanings of parts of predictably extended verbs. For example, if someone wanted the meaning(s) of **-famb-is-a** (1. walk faster. 2. cause someone to walk), he/she would look for the meanings of **-famb-** (walk) and the relevant one of **-is-** (intensive or causative extension) in the dictionary and then "add" them "up".

The method of adding up meanings of the smaller units to get the meaning of the larger construction will in this article be called the analytical approach. This analytical approach is based on two assumptions, that is: (a) that there are lexical units which contain identifiable parts whose meanings, when combined, equal precisely the meaning of the lexical unit as a whole; and (b) that these would be easily identifiable and "addable". When this approach is used, the "theory" which makes predictability possible, assumes conscious knowledge of the derivational processes within the language by its target audience. In the case of extended verbs, the target users would be expected to know the meaning elements brought by the respective verbal extensions to unextended verb stems.

The predictability criterion, in the context of Shona lexicography, is needed for two main reasons. Firstly, it is needed to save space. Derivational processes in Shona are very productive. In fact, most Shona verb stems can produce through derivation a number of other verb stems which at times are analytically predictable. An example is **-taura** (speak/talk) which can be extended to yield highly predictable senses, for example:

-taurisa (+ intensive: speak loudly)
-taurira (+ applied: tell to somebody)
-taurika (+ potential: able to be spoken)
-taudzana (+ causative + reciprocal: cause one another to talk)
-taurirwa (+ applied + passive: be told by somebody)

Inclusion of all these extended forms would take up unnecessary space since their meanings can easily be understood by adding up the meanings of the verb **-taura** and those of the respective extensions.

Secondly, it is needed to avoid redundancy. To include all derived verb forms would be superfluous, since in many cases the meanings of these forms could be traced quite transparently from the base form. As a result, Shona lexicographers have tended to adopt Tsonope's observation that "what is highly predictable is highly deletable" (Chimhundu 1992: 36).

However, despite the fact that it has had a relatively long and productive tradition of use for headword and sense selection within Shona lexicography, the predictability criterion presents problems. For instance, in Chabata (1997) it was shown that what might be predictable to one person, or a group of people, might not be predictable to another person, or another group. To illustrate this, we can take the example **-bikira.** This verb stem is derived by the addition of the applied extension **-ir-** to the root **-bik-** (cook). The sum total of the meanings of **-bik-** and **-ir-** would give us the meaning "cook for (somebody)". This meaning might be predictable to everyone. However, **-bikira** can have, at least for some Shona speakers, although not for all, another more specialised meaning "prepare a love potion". This second meaning might not involve cooking at all; a person could just do good things for his/her partner so that he/she is loved more. The meaning is therefore a meaning that has been transferred from the basic meaning, that is, a metaphorical meaning. Faced with the selection task, this verb might be "predictable" for some lexicographers, but not necessarily for everyone.

Another problem of using the predictability criterion is that despite the fact that any criterion suggests either the presence or absence of something, predictability is a matter of degree. We can have forms that are very predictable, others that are somewhat predictable or somewhat unpredictable, and still others that are very unpredictable. In the above examples we have seen cases where meanings can be somewhat predictable or somewhat unpredictable. However, in addition to these, there are some cases where we can have forms that look like extended verbs that have highly unpredictable meanings. An example of this can be **-kanganwa** (forget) in which, although extension-like elements (that is, the reciprocal **-an-** and the passive **-w-**) seem to be part of the stem, they cannot be related to any relevant roots, that is, **-kang-** or **-kangan-**. Although **-an-** and **-w-** are forms of extensions in Shona, we cannot argue that in this example these are verbal extensions productively extending the verb stem **-kanga**. Instead, they form an indivisible part of the heavily lexicalised verb stem **-kanganwa**. This matter of degree makes the selection of derived forms difficult for the lexicographer using this criterion, for it is usually not easy to determine whether a form is very predictable, somewhat predictable or somewhat unpredictable.

Another problem stems from the treatment Shona verbal extensions have received from some scholars. A number of scholars (for example Fortune 1955, 1957, 1984, Dembetembe 1987 and Harford 1990) have written on Shona verbal extensions and extended verbs. In their analyses they focus on the form of the

extensions, syntactic functions and, to a much lesser extent, the meanings of individual extensions. For example, Dembetembe (1987: 31) says:

> Extensions are distinguished one from another by their shape, syntactic function and meaning. Of these, shape and meaning are obtained subtractively, and function by some form of transformation. Extensions have been identified in this study mainly on the basis of their shape and syntactic function. Meaning has been applied to a lesser extent for the simple reason that translation from Korekore to English is sometimes rather misleading.

Fortune (1955, 1957, 1984) and Dembetembe (1987) deal with the effects of verbal extensions primarily at the morphophonological and the syntactic levels, and not at the semantic level, with the effect that some scholars (for example Harford 1990) actually refer to extensions as inflections, thereby suggesting strongly that these morphemes add little significant meaning to their respective base forms. At the morphosyntactic level, they focus on effects that can be deduced analytically, that is, those that are predictable. For example, they look at the generally predictable argument structures required by verbal extensions like the passive, applied, potential and the causative. Neither the derivational functions of extensions nor their semantic input to verb stems, therefore, have been traditional areas of study in Shona grammar.

However, recent studies in Shona morphology (Mkanganwi 1995, Chabata 1997) have shown that Shona verbal extensions are derivational, and not inflectional morphemes. There are a number of reasons for this. One reason why Shona verbal extensions are regarded derivational is because they usually change the meanings of the verb roots in question in highly significant ways. Because of this, the addition of verbal extensions produces new words that need to be added to the lexicon. If we take the example **-radza** (place a dead body in a grave), we note that the causative verbal extension **-dz-** adds a metaphorical sense to the meaning of the base form **-rara** (sleep). The semantic change is not wholly analysable from the construction since the meaning of the derived form has radically shifted from the meaning of the base form; a completely different event is being described.

Shona verbal extensions are also considered derivational because they typically (but not necessarily) change the syntactic category of the root (base form) to which they apply. In this case, a verb (V) can be derived from or changed into an ideophone (I), a noun (N) or an adjective (A). For example:

che (ideophone of cutting) (I)	> **-che**	**-k**	**-a** (cut) (V)			
		(potential)	(tv)			
shamwari (friend) (N)	> **-shamwari**	**-dz**	**-an**	**-a** (befriend someone) (V)		
		(causative)	(reciprocal)	(tv)		
-pfupi (short) (A)	> **-pfup**	**-is**	**-a** (shorten) (V)			
		(causative)	(tv)			

Mkanganwi (1995: 67) notes that, although some of these extensions do not change the word category of a form, they do move a form into a different syntactic subcategory and involve large meaning changes. For example:

-famb	-a	> -famb	-ir	-a
(walk)	(tv)	(walk)	(applied)	(tv) (1. walk on behalf of 2. take responsibility)

As we can see, both **-famba** and **-fambira** are in the verbal category. The addition of the applied extension has resulted in at least two senses, one of which shows a radical change in the meaning of the base form, and as a result, is unpredictable. Syntactically, whilst **-famba** is a one-argument verb which may take an adjunct, the addition of the applied extension to it results in **-fambira** which is a two-argument verb which may or may not take an adjunct.

Another reason why Shona verbal extensions are regarded as being derivational is the amount of lexical generality they have. Bybee (1985: 84) for example argues that derivational processes are more likely to have lexical restrictions on their applicability. She goes on to note that derivational processes may be applicable only in very restricted semantic, syntactic and phonological domains. To illustrate this, we can take the potential/neuter extension **-ik-** which can be used to extend transitive verbs, but usually not intransitive ones. This extension can for example be suffixed to verb stems like **-ba** (steal) and **-dya** (eat) to result in **-bika** (able to be stolen; stealable) and **-dyika** (able to be eaten; edible) respectively. However, if **-ik-** is suffixed to **-fa** (die) and **-tsva** (burn), both of which are intransitive, the resultant forms, that is, ***-fika** (able to undergo dying) and ***-tsvika** (able to be burnt; burnable) would be unacceptable, despite the fact that the resulting forms were logically possible. Thus, verbal extensions, for example the potential, are less generalisable.

Bybee (1985: 17) also notes that for a morpheme to be generally applicable, it must have only minimal semantic content. With minimal semantic content, the meanings it would add, would be highly predictable. Derivational affixes such as Shona verbal extensions are not applicable to large numbers of stems in precisely the same ways because they have relatively high semantic content.

Shona verbal extensions are also considered derivational morphemes because they can cause large meaning changes to verb forms to which they are attached. The addition of verbal extensions can result in semantic divergences or movements of a number of types which cannot be understood by just adding the respective meanings of the unextended verb stem and that of the verbal extension(s). In the next section we will look at some of the movements that are caused by adding extensions to verb stems.

4. Semantic divergences caused by extending verbs

An analysis of 412 productively extended (to be defined below) Shona verbs discussed here shows that there are a number of types of semantic differences between extended and unextended verbs, most of which cannot be understood

by the analytical approach. The addition of verbal extensions to verb stems can cause a stem which was not very specialised to become more specialised. Specialisation is explained by Ullmann (1964: 228) as follows:

> The net result of the change is that the word is now applicable to fewer things, but tells us more about them; its scope has been restricted, but its meaning has been enriched with an additional feature.

A Shona example is:

> **-dyisa** 1. unpredictable: feed someone with poisoned food. 2. predictable: cause someone to eat; feed someone.

This verb stem has been derived from **-dya** (eat) by the causative extension **-is-**. While **-dya** refers to eating in general and there is no specification of things that are eaten, one derived sense carries a specialised meaning which refers to feeding with poisoned food only. As we can see in this example, the shift in meaning results in the derived meaning of the verb applying to fewer situations than the base meaning, but it yields more information about those situations.

Kastovsky (1990: 78) argues that derivational morphology is usually associated with the process of specialisation of meaning. He notes that this may be due either to the derivational addition of certain semantic components, or to some change in the meaning of the constituents which results from the combination, or both. He also notes that, as a result of specialisation, the overall meaning of the derived form can no longer be deduced from the meanings of its constituents plus the knowledge of the word-formation patterns; rather, additional information is required. If we take the example that has been provided above for **-dyisa**, we would note that the morpheme-by-morpheme analysis of this verb stem would give us only the easily predictable and unspecialised sense, that is, "cause someone to eat; feed someone". The specialised sense cannot be deduced by using this approach.

This point is also noted by Lyons (1977: 524) who argues that the meaning of complex lexemes (which would include extended verbs in Shona) is more specialised than that of the lexemes from which they appear to be derived. He suggests that the reason for this could be that complex lexemes, like simple lexemes, once created and introduced into the language and passed into general currency, may be institutionalised and, by virtue of their use in particular contexts, develop more or less specialised senses.

The addition of verbal extensions to verb stems can also lead to generalisation of meaning. With respect to generalisation, Robins (1990: 344) notes:

> Some words widen the range of their applications or meanings when they come to be used in situational contexts in which they were previously not used or with reference to elements of the contexts with which they were previously not connected.

Although Robins is writing about English here, this generalisation could also be applied to Shona in which the addition of a verbal extension to a verb root may lead to a generalisation of the meaning or reference beyond the confines of the meaning of the base form. An example is **-sunga** (tie). A verb stem **-sungira** is derived from **-sunga** by the addition of the applied extension **-ir-**. The analytic analysis of **-sungira** gives us the combined meanings of **-sunga** and the **-ir-** extension "tie someone". However, the applied extension has also generalised the meaning of **-sunga**. Whilst **-sungira** still has the element of using a rope to tie in some contexts, its meaning has been extended to also refer to the act of "carrying out a traditional prechildbirth ceremony". In this sense therefore, the verb stem applies to more contexts than its literal meaning would suggest.

It is however important to note that specialisation and generalisation seem to function merely as cover terms that have been applied to the different kinds of semantic change brought to verb stems by verbal extensions. The addition of verbal extensions to verb stems seem to cause meaning changes along relatively specific paths of divergence. This was shown by Chabata (1997) in his analysis of extended verbs that had been entered and defined in *Duramazwi ReChiShona*. The primary data was a set of 412 productively extended verbs and their definitions selected from this general Shona dictionary. An extended verb stem was considered to be productively extended if the stem had a literal, basic sense that is commonly used and/or if the extended form, together with its commonly used sense, serves as a base for other extended forms. These were the forms that were considered to be prototypically and productively extended but not lexicalised verb stems.

Out of 15 828 headwords selected in all for *Duramazwi ReChiShona*, 6 634 were verbs. Of these, about 2 000 seem to have been extended in one way or another. Of these 2 000, about 1 588 appeared to be lexicalised forms, that is, meaning shifts had merged in such a way that the sense was no longer traceable back to the senses of the morphemes that made it up. An extended form was considered lexicalised: (a) if its sense was not analytically predictable, for example the relationship between **-femba** (sniff) and **-femb-er-a** (guess); (b) if, although extension-like elements seem to be part of the stem, they can no longer be related to the roots, for example **-kanganwa** and **-kanganisa**, which cannot be related to either **-kangana** or **-kanga**; and/or (c) if the forms cannot be used productively as base forms for other extended verbs. It is however important to note that the "lexicalised" category (through which the data was narrowed) was of necessity an artificial category, since there was no clear separating line between "lexicalised" and "nonlexicalised" forms. Of the 412 that were analysed, about 259 were provided with fully predictable senses. These forms had been included because they were commonly used in everyday speech. There were about 153 verb headwords for which at least two senses had been defined and where one defined sense was not analytically predictable. Attention was then given to the defined sense that was not analytically predictable, in order to understand better how the sense moved away from the basic literal sense. Categories which were called "meaning shifts" or "paths of

divergence" were developed. Some of the "paths" that were discovered include the following:

- Literal to metaphorical — where the description referred to by the unextended verb is transferred to some other description different from but analogous to that to which it is "properly" or literally applicable.
- Inclusive to exclusive — where the addition of the extension would lead the extended form to refer only to one thing or a small group of things rather than everything that the unextended form may refer to.
- Neutral to descriptive — where a neutral verb becomes descriptive and comments on some activity.
- Nonemotional to emotional — where the addition of an extension changes the meaning of the unextended verb from referring to physical activities to refer to activities that are connected with, based upon or appealing to the feelings or passions.
- Reversive — where the meaning of the extended form seems to contradict the meaning of the base form.
- Nonhabitual to habitual — where the addition of a verbal extension to a base form referring to one event or occasion changes the meaning to refer to something that is existing as a settled practice which is constantly repeated.
- Shift of reference — where the addition of an extension leads verbs to change from referring to an action or process to referring to a resultative state that has nothing to do with the named action or process.
- Physical to mental — where the addition of a verbal extension to a verb stem causes a change in reference from physical to mental, the mind and thought.
- Intensity — where the meaning of the extended form shows that something has been done excessively, more than in the unextended sense.

Verb senses were then sorted into these developing categories and it was discovered that although many senses fell into more than one category, there was usually a "primary" or "salient" path along which the sense seemed to have travelled. In some cases, senses seemed to have moved along two prime paths. To illustrate this we can take the example **-pindira**:

> **-pindira** 1. unpredictable: make the wife of a man who is infertile, to bear children, which is usually done by (one of) the man's younger brothers. 2. literal and predictable: enter (a house) for or on behalf of someone.

The addition of the applied extension **-ir-** to **-pinda** (enter a house) has, as we can see from the first sense of this example, shifted the meaning of **-pinda** to a sense that is metaphorical. The sense is metaphorical in that the act of **-pindira** may not even involve the sense of entering a house; a man would just need to have sex with his brother's wife so as to produce children. Besides being meta-

phorical, the meaning is also exclusive because it does not refer to having sex with anyone's wife. In addition, the act is restricted to the wife of one who is infertile. Furthermore, the agent of the action is also restricted only to (one of) the husband's younger brother(s).

The extended verbs were re-sorted according to extension to see if any patterns emerged. These are summarised in table form:

SEMANTIC DIVERGENCES FOR EACH VERBAL EXTENSION

Extension Type	Number of Forms with Unpredictable Senses	Meaning Shift	Percentage of Unpredictable Senses
Causative	26	Metaphorical	77,9
		Exclusivity	22,1
Applied	68	Metaphorical	35
		Description	25
		Emotional	18,3
		Reversive	11,7
		Habitual	10
Perfective	13	Emotional	25
		Exclusivity	75
Potential	6	Shift of reference	100
Passive	4	Metaphorical	100
Reciprocal	7	Mental	59,1
		Emotional	40,9
Repetitive	1	Intensity	100
Intensive	0	0	0
Reversive	0	0	0
Doubly/multiply extended verbs	28	Exclusivity	36,5
		Description	51,1
		Reversive	12,4

The table shows the kinds of meaning shifts that each extension brought to the extended verbs that were studied. The summary of this table indicates that the applied extension can bring a wider variety of meaning differences to stems than any other extension and that at the other end of the scale, the intensively and the reversively extended stems brought with them no meaning shifts at all. Shifts in the directions of metaphorical, emotional and exclusivity spread across the widest variety of extensions and combined extensions.

5. Conclusion

In this article, we have seen that the addition of verbal extensions to verb stems may result in meanings that are either predictable or unpredictable from the

meanings of their parts. We have also seen that the unpredictably extended definitions provided for the examined verb stems diverged in certain specific ways. While it is tempting to think that these divergences "belong" to productively extended verb stems, there is the possibility that these could be typical paths along which derived meanings diverge in general. In either case, additional research is needed. At the same time, developing a deeper sense of how meanings move away from the literal, noncommenting and nonevaluating meanings "predicted" by morphological analysis of productively extended verbs may help definers make more informed choices when they are selecting unpredictably extended verb stems. The definer could for example ask a set of questions about each verb stem — Is this metaphorical, emotional, descriptive, and so on? — in order to select verb headwords using a more delicate set of criteria than the very general predictability criterion.

An exploration of the predictability criterion has shown that this criterion can be used to save space and to avoid redundancy by excluding extended verbs that are analytically predictable, that is, those of which the senses can be understood by adding up the meanings of the base form and those of any extensions found in the stem.

This study has focused on the derivative nature rather than on the grammatical effects of verb extensions, showing that they are highly relevant (in the sense of Bybee 1985), that they have a high semantic content, and that they yield new lexemes that often need separate treatment in the lexicon of the language. We have also seen the need to view verbal extensions as derivational morphemes if the use of the predictability criterion is to be helpful in the selection of Shona extended verbs and senses.

References

Bybee, J.L. 1985. *Morphology: A Study of the Relation between Meaning and Form.* Philadelphia: John Benjamins.

Chabata, E. 1997. *Applying the Predictability Criterion to Extended Verbs: A Study of a Headword and Sense Selection Problem in Shona Lexicography.* Unpublished M.A Dissertation. Harare: University of Zimbabwe.

Chimhundu, H. (Ed.). 1992. *Report on the African Languages Lexical (ALLEX) Project Planning and Training Workshop.* Harare: Department of African Languages and Literature, University of Zimbabwe.

Chimhundu, H. (Ed.). 1996. *Duramazwi ReChiShona.* Harare: College Press.

Dale, D. 1981. *Duramazwi: Shona-English Dictionary.* Gweru: Mambo Press.

Dembetembe, N.C. 1987. *A Linguistic Study of the Verb in Korekore.* Harare: Zambezia, University of Zimbabwe.

Fortune, G. 1955. *An Analytical Grammar of Shona.* Cape Town/New York: Longmans and Green.

Fortune, G. 1957. *Elements of Shona.* Harare: Longmans.

Fortune, G. 1984. *Shona Grammatical Constructions.* Vol. 2. Harare: Mercury Press.

Hannan, M. 1959. *Standard Shona Dictionary.* Harare: The College Press and The Literature Bureau.

Harford, C. 1990. The Applicative in ChiShona and Lexical Mapping Theory. Mchombo, S.A. (Ed.). 1990. *Theoretical Aspects of Bantu Grammar*. Stanford: CSLI Publications.

Jackson, H. 1988. *Words and their Meaning*. London/New York: Longman.

Kastovsky, D. 1990. The Interaction of Semantic and Formal Structures in the Lexicon. Tomaszczyk, J. and B. Lewandowska-Tomaszczyk (Eds.). 1990. *Meaning and Lexicography*. Amsterdam/Philadelphia: John Benjamins.

Landau, S.L. 1984. *Dictionaries: The Art and Craft of Lexicography*. New York: Cambridge University Press.

Lyons, J. 1977. *Semantics*. Vol. 2. New York: Cambridge University Press.

Mkanganwi, K.G. 1995. *Shona: A Grammatical Sketch*. Harare: Department of Linguistics, University of Zimbabwe.

Murray, J.A.H. (Ed.). 1989². *The Oxford English Dictionary*. London: Clarendon Press.

Robins, R.H. 1990. *General Linguistics: An Introductory Survey*. London: Longman.

Svensén, B. 1993. *Practical Lexicography: Principles and Methods of Dictionary-making*. Oxford/New York: Oxford University Press.

Swanson, D.C. 1967. Recommendations on the Selection of Entries for a Bilingual Dictionary. Householder, F.W. and S. Saporta (Eds.). 1967. *Problems in Lexicography*. Bloomington: Indiana University/The Hague: Mouton.

Ullmann, S. 1964. *Semantics: An Introduction to the Science of Meaning*. Oxford: Basil Blackwell.

Zgusta, L. 1971. *Manual of Lexicography*. The Hague: Mouton.

Die makrostrukturele vergestalting van affikse en tegnostamme in Afrikaanse vertalende woordeboeke*

Gerda de Wet, *Departement van Afrikaans en Nederlands, Universiteit van Stellenbosch, Suid-Afrika*

Abstract: The Macrostructural Embodiment of Affixes and Technostems in Afrikaans Translation Dictionaries. The history of Afrikaans dictionaries shows that in the past the focus was chiefly on the inclusion of words as lemmas in the macrostructure. Although the Afrikaans lexicon mainly consists of words, there occur two other types of lexical items in the lexicon as well, namely sublexical items (e.g. affixes and stems) and multilexical items (e.g. expressions and idioms).

In this article the inclusion and presentation of sublexical items in Afrikaans translation dictionaries, that is, standard translation desk dictionaries with Afrikaans and English as language pair, are emphasized. The two dictionaries under discussion are *Groot Woordeboek/Major Dictionary* (MD) and *Tweetalige Woordeboek/Bilingual Dictionary* (BD) which belong to this category. The discussion encompasses a critical evaluation of the inclusion and presentation of sublexical lemmas according to present-day metalexicographical approaches. The theoretical discussion concentrates throughout on practical lexicography and articles are therefore evaluated according to theoretical principles.

From the investigation of the macrostructural embodiment of sublexical lemmas in BD and MD, it becomes clear that a wider variety of stems than affixes are included in these dictionaries. Where affixes are included, they are treated more scantily than stems. There is also a need for the inclusion of comprehensive instructions for the user in both dictionaries.

Keywords: MACROSTRUCTURE, AFFIXES, TECHNOSTEMS, SUBLEXICAL ITEMS, TRANSLATION DICTIONARY, DESK DICTIONARY

Opsomming: Die Afrikaanse woordeboekgeskiedenis toon dat daar in die verlede veral op die opname van woorde as lemmas in die makrostruktuur gekonsentreer is. Alhoewel die Afrikaanse leksikon hoofsaaklik uit woorde bestaan, kom daar egter ook twee ander tipes leksikale items voor, naamlik subleksikale items (bv. stamme en affikse) en multileksikale items (bv. vaste uitdrukkings en idiome).

In hierdie artikel val die klem op die opname en aanbieding van subleksikale items in Afrikaanse vertalende woordeboeke, dit wil sê standaard vertalende handwoordeboeke met Afrikaans en

* Hierdie artikel is 'n aangepaste hoofstuk uit 'n M.A.-skripsie *Die opname en bewerking van subleksikale lemmas in Afrikaanse vertalende woordeboeke* wat in Desember 1997 deur die Universiteit van Stellenbosch aanvaar is.

Engels as taalpaar. *Groot Woordeboek/Major Dictionary* (GW) en *Tweetalige Woordeboek/Bilingual Dictionary* (TW) wat tot dié kategorie behoort, is die woordeboeke onder bespreking. Die bespreking sluit 'n kritiese evaluering in van die opname en aanbieding van subleksikale lemmas aan die hand van bestaande metaleksikografiese uitgangspunte. Die teoretiese bespreking is deurgaans op die leksikografiese praktyk gerig en woordeboekartikels word daarom aan die hand van teoretiese beginsels geëvalueer.

Uit die ondersoek na die makrostrukturele vergestalting van subleksikale lemmas in GW en TW, blyk dit dat 'n groter verskeidenheid stamme as affikse in hierdie woordeboeke opgeneem word. Waar affikse wel opgeneem word, word hulle kariger behandel as stamme. Daar is ook 'n behoefte aan die insluiting van vollediger en omvattender instruksies vir die gebruiker in albei woordeboeke.

Sleutelwoorde: MAKROSTRUKTUUR, AFFIKSE, TEGNOSTAMME, SUBLEKSIKALE ITEM, VERTALENDE WOORDEBOEK, HANDWOORDEBOEK

Affikse word baie swak in Afrikaanse vertalende woordeboeke hanteer. Afgesien van die inkonsekwente opname en bewerking van prefikse, figureer suffikse feitlik nêrens in die leksikografiese bewerkingsproses nie. Hierdie swak hantering van affikse spruit onder meer voort uit die betrokke woordeboeke se woordgerigte benadering en die gevolglike ontoereikende keuse van leksikonitems wat as behandelingseenhede opgeneem word. Die aanname dat affikse lemmastatus moet kry, volg daaruit dat die affigale vorm die enigste optrede van die betrokke leksikale item *is*; daar *is* nie 'n woordvariant wat semanties parallel aan die affiks optree nie. Uit die aard van hul vorm het affikse net 'n subleksikale optrede en gevolglik kan hul opname en bewerking nie aanleiding gee tot die duplisering van inligting nie (Gouws 1991: 117).

Dit blyk verder ook dat daar verwarring bestaan tussen die bewerking van prefikse en tegnostamme wat as beginkomponente van samestellings optree. Van Niekerk (1989: 88-89) onderskei affikse en tegnostamme van mekaar op grond van veral een belangrike kenmerk: affikse verbind altyd met stamme en nooit met mekaar nie, terwyl tegnostamme wél, maar nie noodwendig nie, met mekaar kan verbind, byvoorbeeld **haplo-** en **-grafie** vorm **haplografie** en **brio-** en **-fiet** vorm **briofiet**.

1. Prefikse

In sowel TW as GW word prefikse as lemmas opgeneem, maar op 'n baie klein skaal. Prefikse wat opgespoor kon word, is die volgende: in TW (A-E): **al-, her=, Pan=, mal=, wan=** en (E-A): **cis=, ex-, demi=, extra=, mal=, mis=, non-, post=**, en in GW (A-E): **al-**[2], **her=** en (E-A): **ex=, non-, post-**[3], **quasi-**[1].

1.1 Homonimie by affikse

Die leksikografiese hantering van homonimie by affikse geniet nog minimale aandag. Tans word affikse, waar van toepassing, in sowel TW as GW as homonieme

saam met leksikale lemmas hanteer. So 'n geval in GW is die volgende: **post**[1], **post**[2] en **post-**[3] (E-A):

> **post**[1], (n) stut, paal, styl (deur); stander, pilaar (*myn.*); (v) aanplak; bekend maak.
> **post**[2], (n) pos; posisie, betrekking; wag; fort; poskantoor; poswese, posdiens; pospapier; standplaas [...]
> **post-**[3], na=.

Die probleem met hierdie bewerking is dat subleksikale en leksikale lemmas nooit as lede van een homonimiese paar hanteer kan word nie (Gouws 1989: 138). Homonieme vereis vormlik identiese lemmas en die koppelteken, as deel van die lemma, impliseer dat daar 'n vormlike verskil is.

Affikse kan ook, soos woorde, verskillende polisemiese waardes hê of onderlinge homonimiese verbande toon. Combrink (1990: 38) maak in sy lys voorvoegsels onder andere voorsiening vir meer as een morfologiese optrede van **ge-**. So is daar **ge-**[1] in **(het) geëet, geslaap, gewerk**; **ge-**[2] in **(word/is) geëet, gesê, gedoen**; **ge-**[3] in **('n) gemors, geraas, getjank** en **ge-**[4] in **gebaadjie, gedas, gestewel**. Die leksikograaf moet ondersoek instel of daar 'n semantiese onderskeid tussen die vier vorme van **ge-** is, en indien wel, of die onderskeid polisemies of homonimies van aard is. Indien die onderskeid argumentshalwe homonimies is, behoort hierdie affikse as verskillende lemmas neerslag in die woordeboek te vind. Dit beteken dat **ge-**[1], **ge-**[2], **ge-**[3] en **ge-**[4] as aparte subleksikale lemmas en as lede van 'n homonimiese kategorie opgeneem behoort te word. Die onderskeid in hierdie geval is egter meer kompleks: **ge-**[1], **ge-**[2] en **ge-**[4] is wel poliseme van mekaar omdat die verskillende optredes van **ge-** hier nie betekenismatig genoegsaam verskil om aparte inskrywings te regverdig nie. Die feit dat hulle meerfunksionaliteit openbaar, is nie noodwendig 'n aanduiding dat hulle homonieme is nie. Hierdie drie gevalle is wél sogenaamde "grammatiese homonieme", maar semanties is daar 'n nouer verband tussen hulle. In die geval van **ge-**[3] is daar egter 'n betekenisverskil met die ander drie en dus is dit 'n homoniem van hulle.

Die situasie in die vertalende woordeboeke is tans van so 'n aard dat, indien 'n affiks wél opgeneem word, dit net in een artikel bewerk word en moontlike homonimiese affikse almal onder een makrostrukturele inskrywing hanteer word.

1.2 Polisemie by prefikse

Op grond van hul status as leksikonitems behoort affikse as homonimiese of as polisemiese lemmas opgeneem te word waar van toepassing. Homonimie geld tussen leksikale items van dieselfde vlak, met ander woorde óf tussen subleksikale items óf tussen meerleksikale items óf tussen woorde. Daar is ook nog nie, soos in die geval van homonimie, in die Afrikaanse vertalende leksikografie gekyk na die opname van polisemiese affikse nie.

2. Suffikse

Suffikse kry geen makrostrukturele erkenning in TW of GW nie. Hulle figureer wél as deel van afleidings, maar nie as volwaardige subleksikale items in die sin dat hulle as aparte lemmas opgeneem word nie. Alhoewel alle produktiewe suffikse vir opname in die woordeboek oorweeg behoort te word, kry affikse wat slegs grammatiese waarde het, normaalweg nie lemmastatus nie. Daar is verskillende suffikse wat nie noodwendig opgeneem hoef te word nie; twee sulke tipes is die vervroulikingsuffikse wat (a) persoonsname en (b) soortname vorm.

2.1 Vervroulikingsuffikse wat persoonsname vorm

Suffikse wat vrouepersoonsname vorm, is onder andere die **-ika** in **Laurika**, **-ita** in **Benita**, **-lina** in **Douwlina** en **-ri** in **Anneri**. Tweetalige woordeboeke neem meermale persoonsname op. Vir die Afrikaanse vervroulikingsuffikse by persoonsname is daar egter nie dikwels ooreenstemmende Engelse vertaalekwivalente beskikbaar nie. Die beste oplossing sou wees om telkens die volledige vrouepersoonsnaam aan die Afrikaanse kant te lemmatiseer, gevolg deur die naaste Engelse vertaalekwivalent. Dieselfde metode kan ook aan die Engelse kant toegepas word.

2.2 Vervroulikingsuffikse wat soortname vorm

In die buitetalige werklikheid is daar tans die neiging tot geslagtelike gelykheid. Dit is die gevolg van 'n sterk opkoms van die feminisme. Beyer (1995: 6-7) voer aan dat die vervrouliking van 'n leksikale item in ekstreme gevalle beskou word as seksistiese diskriminasie. Die ideale toestand sou wees dat persoonsname geslagtelike neutraliteit kommunikeer. Die neiging bestaan dus al meer in Afrikaans om die manlike persoonsname vir albei geslagte te gebruik en om die vroulike vorme wat dikwels van die manlike vorme deur middel van suffikse afgelei word, te vermy.

Combrink (1990: 105) noem drie reëls waarvolgens woorde in Afrikaans vervroulik word. Dit is veral reël (c) wat vir hierdie artikel ter sake is.

(a) Afsonderlike vroulike vorme vir sommige lewende wesens, soos **vrou**, **merrie** en **koei**.
(b) Samestellings met [+ menslike] stamme soos **dame(s)-** in **dameskoen**, **meisie(s)-** in **meisiespan**, en **-vrou** in **polisievrou** asook [– menslike] stamme soos **-koei** in **walviskoei** en **-wyfie** in **leeuwyfie** wat vroulikheid impliseer.
(c) Afleidings deur middel van suffikse. Onder hierdie suffikse is **-e** soos in **prinsipale**, **-es** soos in **bewaarderes**, **-ise** soos in **aktrise** en **-ster** soos in **naaldwerkster**.

Uit bostaande drie reëls is dit duidelik dat Afrikaans se vervroulikingsvorme

grootliks morfolgies gevorm word, terwyl dit in Engels meestal op leksikale vlak geskied. Dit skep 'n probleem vir die leksikograaf, deurdat daar nie geredelik vir elke Afrikaanse vervroulikingsuffiks 'n Engelse woord as vertaalekwivalent beskikbaar is nie. 'n Moontlike oplossing vir hierdie probleem sou wees dat die leksikograaf aan die Afrikaanse kant van die woordeboek dié pare geslagsgemerkte leksikale items waarvoor daar in Engels net een vertaalekwivalent is, as aparte lemmas opneem. So 'n voorbeeld sou wees:

A-E
onderwyser teacher
onderwyseres teacher

E-A
teacher onderwyser (n), onderwyser (m), onderwyseres (v).

In voorbeeld A-E geld 'n ekwivalentverhouding van konvergensie. Die omkeerbaarheidsbeginsel lei dan tot voorbeeld E-A wat 'n verhouding van divergensie weerspieël. Dit blyk dus uit bostaande inskrywings dat beide **onderwyser** (wat sowel die manlike as die neutrale vorm verteenwoordig) en **onderwyseres** (die vroulike vorm) in Engels met die ongeslagtelike **teacher** vertaal kan word. Aan die ander kant blyk dit ook dat **teacher** met sowel **onderwyser** as **onderwyseres** vertaal kan word. In hierdie verband moet die leksikograaf duidelik leiding aan die gebruiker gee. Die volgorde waarin die vertaalekwivalente verskaf word, behoort volgens die heersende leksikografiese praktyk gedoen te word, naamlik gebruiksfrekwensie. Die woord wat die meeste gebruik word, behoort eerste gegee te word — en nie noodwendig die ongemerkte item nie. Naas die vertaalekwivalent, moet die geslag van die persoon ook aangedui word. Dieselfde geld vir leksikale items soos die Afrikaanse **eggenoot** en **eggenote** wat albei met die Engelse **spouse** vertaal word. Dit is wel belangrik om affikse waarvoor daar nie vertaalekwivalente is nie, op te neem en hul bewerking so aan te pas dat dit nie net op die weergawe van 'n vertaalekwivalent gerig is nie. In die geval van **eggenote** en **onderwyseres** wat onderskeidelik met die ongeslagtelike **spouse** en **teacher** vertaal word, behoort aangedui te word dat hulle vervroulikingsvorme is, byvoorbeeld:

> **eggenote** spouse (f.)
> **onderwyseres** teacher (f.)

Om elke suffiks in die woordeboek op te neem, sou lei tot onnodige oorlading. Die leksikograaf moet probeer om 'n ondersoek te loods na die mees produktiewe suffikse in Afrikaans en ter wille van volledigheid 'n seleksie van hierdie suffikse insluit. Kempen (1982: 400-588) bespreek 'n groot aantal suffikse wat in Afrikaans voorkom. Nie al hierdie suffikse is meer produktief nie. Suffikse wat egter nog produktief in Afrikaans optree, is onder andere gevalle soos **-aar, -aas, -dig, -erd, -lei, -lik** en **-loos**. Sulke gevalle behoort vir opname oorweeg te word. Suffikse moet per definisie dieselfde bewerking as prefikse kry.

Vir die probleem ten opsigte van die opname van affikse is daar twee moontlike oplossings.

Eerstens: Die leksikograaf kan volledigheidshalwe 'n lys van al die Afrikaanse affikse in die agterwerk van die woordeboek opneem, met hul moontlike Engelse vertaalekwivalente. Gepaardgaande daarmee kan ook 'n lys van Engelse affikse ingesluit word, met hul onderskeie Afrikaanse vertaalekwivalente. Affikse kan egter wél aanspraak maak op lemmastatus, en die lys in die agterwerk moet nié die opname van affigale lemmas in die makrostruktuur van die woordeboek vervang nie. Die opname van 'n afsonderlike lys in die agterwerk lei tot 'n verdere toegangstruktuur. Alhoewel so 'n woordeboek nie meer monotoeganklik is nie, word sy waarde as houer van inligting vergroot.

Tweedens: Subleksikale items kan in kombinasie met stamme behandel word. Die suffiks **-e** wat meervoudsvorme in Afrikaans aandui, is een van die suffikse wat nie noodwendig sal kwalifiseer vir individuele opname in 'n vertalende woordeboek nie. Afrikaans het egter ook 'n **-e** wat vervrouliking aandui. Die leksikograaf moet op die een of ander wyse tussen die twee onderskei. Om hierdie probleem te omseil, kan die leksikograaf in die artikel van 'n lemma wat **-e** as meervoudsvorm neem, die suffiks se funksie soos volg aandui:

boek *-e pl.*

Die meervoud word egter nie altyd slegs deur die toevoeging van 'n meervoudsuffiks gevorm nie. Soms lei meervoudsvorming tot 'n spellingsaanpassing, byvoorbeeld **dieet** en **diëte, betoog** en **betoë**. Soms maak dit die gebruik van afkappingstekens (**hoera** en **hoera's, radio** en **radio's**) en kappies (**wig** en **wîe, brug** en **brûe**) nodig. Daar bestaan ook wisselmeervoudsvorme byvoorbeeld **doktors** en **doktore, vroue** en **vrouens**. Om groter konsekwentheid in die woordeboek te bewerkstellig en dit vir die gebruiker toegankliker te maak, kan die leksikograaf naas die lemma in die enkelvoudsvorm, in dieselfde artikel die meervoudsvorm opneem deur die hele meervoudsvorm uit te skryf:

dieet *diëte pl.*
hoera *hoera's pl.*

3. Tegnostamme

Sowel TW as GW het op groot skaal tegnostamme wat as beginkomponente van samestellings optree, in die makrostruktuur opgeneem, byvoorbeeld in TW (A-E): **aëro=, Anglo:, astro=:, chiro=:, elektro=:, fito=:, fono=:, galvano, halo=, heli=:, hemi=:, hemo=:, hetero=:, hidro=:, higro=:, hiper=:, hipo=:, hippo=:, hipso=:, histo=:, holo=:, homeo=:, mega=:, meta=:, mikro=:, neo=, osteo=:, outo=:, pseudo=, psigo=:, teo=:, termo=:, xeno=:, xero=:, xilo=:** en (E-A): **aero=, Anglo=, astro=:, bio=:, chiro=:, electro=:, exo=:, galvano=:, halo=, heli, helio=:, hema=, haema=, hemi=:, hemo=, hepta=:, hetero=:, hexa=:, hydro=:, hygro=:, hypno=:, hypo=:, hypso=:, holo=:, homo=:, kilo=, litho, mega=:, meta=:, necro=:, neo=:, neuro=:,**

ortho=:, **osteo=:**, **paleo=:**, **pal(a)eo=:**, **penta=:**, **phyto=:**, **pneumo=:**, **psycho=:**, **sapro=:**, en **thermo-:**.

Soos in die geval van suffikse word tegnostamme wat as eindkomponente van samestellings funksioneer, ook nie in die woordeboek opgeneem nie alhoewel daar 'n substansiële aantal in die taal voorkom, byvoorbeeld **-fiet**, **-fiel**, **-foon**, **-fobie**, **-graaf**, **-gram**, **-liet**, **-logie**, **-loog**, **-maan**, **-metrie**, **-niem**, **-paat**, **-skoop** en **-staat**. Wat wél in TW die geval is, is dat hierdie tegnostamme wat as eindkomponente van tegnosamestellings[2] optree, soms in die artikel van 'n tegnostam wat as beginkomponent optree, bewerk word:

> **hidro=: ~chloried, ~chloride** hydrochloride. [...] **~fiel** =*e n.* hydrophile *(chem.)* [...] **~fobie** hydrophobia [...] **~foon** =*one* hydrophone [...] **~graaf** =*awe* hydrographer [...]

Alhoewel eindkomponente wél opgeneem word, is daar geen konsekwente manier om 'n betrokke eindkomponent in die woordeboek op te spoor nie. As 'n gebruiker argumentshalwe die tegnostam **-fobie** wil naslaan, is dit nie onder die alfabetiese lysting van lemmas te vind nie. Dit is ook nie moontlik om onder die beginkomponente vir die toepaslike samestelling te gaan soek waarin die eindkomponent kan optree nie.

Vir sover tegnostamme wat as eindkomponente in tegnosamestellings optree, deel is van die Afrikaanse leksikon, maak hulle aanspraak op lemmastatus. Mits hul produktiwiteit hul opname in die woordeboek regverdig, behoort hulle as lemmas leksikografies volledig bewerk te word, met ander woorde behoort hulle mutatis mutandis dieselfde aandag te geniet as tegnostamme wat as beginkomponente funksioneer. Dit behels onder andere dat hulle voorsien moet word van sowel etimologiese inligting as eenwoordsitate om sodoende die betekenis en gebruik van die tegnostam duidelik te illustreer.

Die aanwending van voorbeeldsinne om die gebruik van 'n woord te toon is normaalweg 'n noodsaaklikheid in die geval van leksikale lemmas. In die geval van 'n tegnostam waar dit veral gaan om die optrede van die betrokke tegnostam in verbinding met 'n ander stam, is eenwoordsitate van die grootste belang. Dit sou egter nuttig wees as die leksikograaf elk van die eenwoordsitate ook met ten minste een voorbeeldsin kan voorsien om sodoende die betekenis van die tegnostam binne die samestelling waarin dit optree, duideliker te illustreer.

By 'n artikel waarin 'n tegnosamestelling soos **hidrochloried** voorkom, verwys Wiegand (1983: 432-437) na die inskrywing **hidro=** as 'n lemmadeel of deellemma in 'n lemma-eksterne nesingang. Die bewerkingsteken is die volledige lemma, naamlik **hidrochloried**. **~chloried** bestaan uit 'n lemmadeel en 'n leksikografiese plekhouersimbool — die tilde. 'n Lemmadeel is deel van 'n lemma. Dit tree op in die lemmatisering van samestellings. Die lemmadeel lei 'n groep neslemmas in, soos in die geval van **hidro=**. Die lemma wat bewerk word, is die lemmadeel, bv. **hidro=**, plus die tweede stam, bv. **~chloried**. Die lemma is dan **hidrochloried**.

Die lemmas in 'n woordeboek is gewoonlik alfabeties en vertikaal gerig. Wanneer samestellings in een artikel opgeneem is en as sublemmas hanteer word

én hierdie sublemmas is alfabeties ingeskryf, praat Wiegand van nislemmatisering. Elke lemma in so 'n groep is 'n nislemma. In so 'n nis is daar geen semantiese of morfologiese implikasies nie. 'n Voorbeeld van nislemmatisering in TW is die geval **hidro=**:

hidro=: **~chloried** [...] **~dinamies** [...] **~dinamika** [...] **~ëlektries** [...] **~fiel** [...]
hidroulies [...]

Indien van die alfabetiese patroon afgewyk word en van die sublemmas in 'n artikel weens semantiese redes bymekaar geplaas word, noem Wiegand dit neslemmatisering. 'n Voorbeeld van neslemmatisering in TW is die geval ***Noord=:***

Noord=: **~~Afrika** [...] **~~Afrikaans** [...] **~~Amerika** [...] **~~Amerikaans** [...] **~~Europa** [...] **~~Frankryk** [...]
noorde [...]

Hier behoort die lemma **noorde** volgens 'n streng alfabetiese volgorde voor **Noord-Europa** te staan. Maar weens die semantiese implikasies word **Noord-Europa** in die artikel van *Noord=:* opgeneem.

4. Omkeerbaarheid en die opname van prefikse en tegnostamme

In 'n tweerigting- vertalende woordeboek moet daar aangedui word wat die wedersydse ekwivalentverhoudinge tussen die leksikale items van die betrokke tale is (Gouws 1989: 162). Dit beteken dat 'n lemma A in die brontaal X, wat 'n leksikale item van die brontaal is, met sy vertaalekwivalente B en C wat leksikale items van die doeltaal Y is, in taal Y as een van die moontlike vertaalekwivalente van die lemmas B en C opgeneem word.

Omdat subleksikale lemmas wel volwaardige lemmas is, moet die leksikograaf konsekwent wees in sy bewerking van die verskillende tipes lemmas. In hierdie geval behoort omkeerbaarheid, soos toegepas in die artikels van leksikale lemmas, ook toegepas te word in die artikels van subleksikale lemmas.

4.1 Prefikse

Die volgende artikel kom in TW voor:

al- all-, pan-

Die eerste vertaalekwivalent *all-* word nie soos 'n prefiks aan die A-E kant as lemma opgeneem nie, maar as vetkursiefgedrukte soekelement is dit slegs 'n lemmadeel in 'n lemma-eksterne nesingang, naamlik

all=: **~~absorbing** allesoorheersend, allesbeheersend [...]

Die lemma **al-** se vertaalekwivalent *pan-* het nie 'n ooreenstemmende Engelse lemma **pan-** nie. TW maak wel voorsiening vir die lemma **Pan-** met sy vertaalekwivalente *Al-* en *Pan-*, terwyl daar nie 'n Afrikaanse **Al-** as lemma opgeneem is nie. **Pan-** word ook nie as lemma aan die A-E kant opgeneem nie, maar wel **Pan=:**. Leksikonitems is dus onvolledig verteenwoordig in die makrostruktuur van die woordeboek. Dit kan verwarrend wees vir die gebruiker wanneer 'n vertaalekwivalent verskaf word wat op sy beurt nie weer as lemma opgeneem is nie.

Nog voorbeelde in TW waar die omkeerbaarheidsbeginsel by prefikse nie toegepas is nie, is **her=** en **post=**. Aan die A-E kant word **her=** as affigale lemma gegee met sy vertaalekwivalent *re=*. Omgekeerd word **re-** egter nie as affigale lemma aan die E-A kant opgeneem nie. So ook word **post=** as lemma aan die E-A kant opgeneem met sy vertaalekwivalent *na=*. Aan die A-E kant word daar egter net vyf deellemmas ***na=:*** in die ingangsposisie van groter lemmaneste gegee.

Dit blyk dus dat die omkeerbaarheidsbeginsel ten opsigte van affigale lemmas nie werklik in TW toegepas word nie.

4.2 Tegnostamme

Net soos in die geval van prefikse, is TW nie konsekwent wat betref die wedersydse opname van tegnostamme as subleksikale lemmas en as vertaalekwivalente nie. Die volgende lys is 'n seleksie van tegnostamme aan die A-E kant van TW, waarvan die Engelse weergawes ook aan die E-A kant as lemmas opgeneem word:

A-E	E-A
aëro=	**aero=**
chiro=:	**chiro=:**
elektro=:	**electro=:**
fito=:	**phyto=:**
galvano=:	**galvano=:**
heli=:	**heli=:**
mega=:	**mega=:**
penta=:	**penta=:**
psigo=:	**psycho=:**
termo=:	**thermo=:**

Dit gebeur wel soms dat die vertaalekwivalent van 'n tegnostam wat as lemma aan die X-Y kant opgeneem is, nie aan die Y-X kant as lemma gegee word nie, of omgekeerd, byvoorbeeld **centi=:**, **hexa=:**, **homo=:**, **necro=:** en **paleo=:** word aan die E-A kant opgeneem, maar nie **senti=:**, **heksa=:**, **homo=:**, **nekro=:** en **paleo=:** aan die A-E kant nie, terwyl **mikro=:**, **outo=:**, **pseudo=:**, **teo=:**, **xeno=:**, **xero=:** en **xilo=:** aan die A-E kant vermeld word, maar **micro=:**, **auto=:**, **pseudo=:**, **theo=:**, **xeno=:**, **xero:** = en **xilo=:** nie aan die E-A kant nie.

'n Probleem wat onder andere voorkom by die metode om soekelemente te gebruik, is dat die omkeerbaarheidsbeginsel dikwels nie by samestellings wat uit

twee tegnostamme gevorm is en 'n tegnosamestelling as vertaalekwivalent het, gehandhaaf word nie. Dit gebeur deurdat die leksikale item wat as vertaalekwivalent van die tegnostam verskyn, nie as sodanig lemmastatus kry nie. Die eerste stam tree in verbinding met 'n ander stam as deellemma op. So word daar byvoorbeeld aan A-E kant van TW **fito=**: **~chemie** opgeneem met sy vertaalekwivalent *phytochemistry*. Die woord **phytochemistry** word nie as lemma aan die E-A kant vermeld nie, maar die vertaalekwivalent *phyto=* word wel as subleksikale lemma aangebied waaronder ander tegno-eindkomponente met hul vertaalekwivalente gegee word: **phyto=**: **~biology** [...] **~genesis** [...] **~pathology** [...]. 'n Soortgelyke geval word aan die A-E kant aangetref waar die subleksikale lemma **meta=**: **~bisulfiet** vermeld word saam met die vertaling *metabisulphite*. Aan die E-A kant word nêrens 'n lemma **metabisulphite** aangetref nie, maar wel 'n subleksikale lemma **meta=**: opgeneem waaronder ander tegno-eindkomponente saam met hul vertaalekwivalente gegee word: **meta=**: **~bletics** [...] **~genesis** [...] **~xylem** [...]. Gevalle waar hierdie stelsel gebruik word, kom redelik dikwels voor.

Wat inderwaarheid hier plaasvind, is dat die bewerking nie op die tegnostam self gerig is nie, maar op die samestelling waarin die tegnostam optree. Die woordeboekgebruiker moet uit die vertaalekwivalent van die betrokke samestelling aflei wat die vertaalekwivalent van die tegnostam is. Die tegnostam word dus nie as subleksikale item hanteer soos die veronderstelling is nie, maar as deel van 'n samestelling wat sy eie lemmastatus het. By gevalle soos die samestellings met **peri-** in TW sou 'n subleksikale lemma **peri-**: opgeneem kon gewees het, maar hier word aan al die samestellings hul eie lemmastatus gegee. 'n Tegnostam **peri-** ontbreek dus aan sowel die A-E as die E-A kant. In GW word **peri-** aan die E-A kant as subleksikale lemma aangebied, maar al die samestellings waarin dit voorkom, word nie daaronder behandel nie. Hulle kry ook, soos in TW, hul eie lemmastatus. Die behandeling verskil dus nie van dié aan die A-E kant waar **peri-** nie as subleksikale lemma opgeneem word nie.

Soos by leksikale en affigale lemmas, behoort die tegnostam eerstens individueel bewerk te word en dus van ('n) vertaalekwivalent(e) voorsien te word. Hierdie vertaalekwivalent(e) behoort dan weer aan die alternatiewe kant van die woordeboek as subleksikale lemma(s) opgeneem te word deur dit (hulle) vormlik met 'n koppelteken te merk. Dit behoort verder dieselfde mikrostrukturele behandeling as ander lemmas te kry en dán eers kan 'n stelsel van nes- of nislemmatiese rangskikking toegepas word.

Dit gebeur ook in TW dat 'n tegnostam as subleksikale lemma opgeneem word, maar dat sy vertaalekwivalent, wat wel as subleksikale item in die doeltaal optree, nie as sodanig aangebied word nie, byvoorbeeld **bio=** aan die A-E kant wat met *bio* vertaal word. 'n Verdere inkonsekwentheid wat soms voorkom, is die aanbieding van tegnostamme as leksikale lemmas sonder aanduiding van hul subleksikale status. So 'n voorbeeld is **galvano** wat aan die A-E kant as leksikale lemma gegee word met *galvano* as vertaling, maar in werklikheid subleksikale status het soos blyk uit die tegnosamestellings wat daaronder gegee word: **~grafie**, **~meter**, **~plastiek**, e.s.m. Aan die E-A kant word **galvano=**: slegs as subleksikale lemma opgeneem. Die inskrywing **galvano-**: behoort as subleksikale lem-

ma aan sowel die Afrikaanse as die Engelse kant aangebied te word, gevolg deur die ooreenstemmende vertaalekwivalente.

Selfs al word die omkeerbaarheidsbeginsel toegepas, kom daar nog steeds inkonsekwenthede voor. Aan die E-A kant word **aero=** opgeneem met *aëro=* as vertaalekwivalent en dit word gevolg deur lemmadele soos **~bat** ensovoorts. Aan die A-E kant kry **aëro=** niè 'n vertaalekwivalent nie. Dus, waar **aero=** as behandelingseenheid optree, verval dié funksie by **aëro=** en word dit as 'n lemma-eksterne ingang opgeneem.

Dit blyk dus dat, wat die makrostrukturele bewerking van tegnostamme betref, die omkeerbaarheidsbeginsel, soos ook in die geval van affigale lemmas, nie werklik in TW toegepas word nie.

5. Die verwysing na subleksikale lemmas in die voorwerk van GW en TW

5.1 Kriterium vir die opname van subleksikale items

Die hoofkriterium vir die opname van subleksikale items in Afrikaanse vertalende woordeboeke is, soos tevore genoem, hul status as leksikale items. Daarom verdien hul hantering reeds aandag in die voorwerk van die woordeboek.

5.2 Die voorwerk

Die voorwerk van 'n woordeboek is veronderstel om die woordeboekgebruiker te help om die inligting wat hy/sy soek, maklik te vind en te gebruik. Hausmann en Wiegand (1989: 330) dui in hul artikel oor die verskillende onderdele van die woordeboek aan dat die woordeboek hoofsaaklik verdeel kan word in die voorwerk, die agterwerk en 'n sentrale woordelys wat die makrostruktuur en die mikrostruktuur van die woordeboek omvat. Die voor- en agterwerk is meestal opsioneel, terwyl die sentrale woordelys en die toeligting twee noodsaaklike tekste is. Die tekste wat in die voor- en agterwerk voorkom, hoef nie in 'n spesifieke volgorde opgeneem te word nie. Die voorwerk bestaan gewoonlik uit die titel, die inhoudsopgawe, gebruiksintruksies of voorwoord en die grammatika van die woordeboek. Die agterwerk bevat meestal verskillende byvoegsels.

5.2.1 Die voorwoord

Een van die funksies van die voorwoord wat Carstens (1995: 142) onder andere noem, is naamlik om "aan die gebruiker 'n sleutel te gee tot die struktuur en die metataal van die woordeboek". Die voorwoord en toeligting van die tipografie van TW en GW sal in die lig van hierdie funksie bespreek word.

Die voorwoord van GW bestaan uit 'n uiters bondige paragraaf. Daar word slegs aangedui dat die veertiende uitgawe "grondig hersien en aansienlik uitgebrei (is)" en dat "meer idiomatiese uitdrukkings bygekom (het)". Behalwe die ver-

wysing na die insluiting van "nuwe woorde ... uit die daaglikse gesproke taalgebruik" en "eiename" word daar geen verantwoording gegee van die opnamebeleid nie. Dit ís so dat die deursneewoordeboekgebruiker dikwels nie eens die voorwoord lees nie, maar net in die sentrale teks die leksikale inligting soek wat hy/sy benodig. Nogtans behoort die leksikograaf nie om dié rede die voorwoord onvolledig aan te bied nie.

Daar is tans die strewe in die leksikografie om leksikale, multileksikale en subleksikale lemmas so ver moontlik volgens dieselfde kriteria te behandel en gevolglik is dit onwenslik dat enigeen van dié lemmatipes 'n skraler bewerking as die ander kry. Die afwesigheid van 'n verduideliking aangaande die hantering van veral subleksikale lemmas wat tradisioneel nie genoegsame aandag in woordeboeke gekry het nie, kan impliseer dat hierdie lemmas glad nie in die woordeboek figureer nie.

In die voorwoord behoort ook gewys te word op die problematiek rondom die onderskeid en gebruik van affikse en tegnostamme. Die woordeboek moet hier 'n enkoderende funksie vervul deur die gebruiker in staat te stel om die vertaalekwivalente wat verskaf word, in die regte konteks te gebruik. Alhoewel sowel affikse as tegnostamme subleksikale items is, verbind hulle nie met dieselfde stamme in woordvorming nie. Affikse, byvoorbeeld, verbind nie met tegnostamme nie en tegnostamme kan met mekaar verbind, maar ook met gewone stamme. In die niealfabetiese gedeelte van die woordeboek behoort morfologiese inligting óf in die toeligtende aantekeninge óf elders in die voor- en/of agterwerk te verskyn (Gouws 1989: 233). In hierdie minigrammatika behoort die teikentaal se fleksiestelsel kortliks bespreek en patroonmatige gevalle aan die hand van toepaslike voorbeelde geïllustreer te word. Svensén (1993: 230-231) wys daarop dat die voorwoord 'n uitvoerige bespreking in die vorm van "instructions for use" moet bied wat hulp met die ontsluiting van die woordeboek kan verleen. Hierdie gebruikersinstruksies moet 'n verduideliking insluit van die omvang van die mikro- en die makrostruktuur van die woordeboek asook die verskillende inligtingskategorieë wat hanteer word. Hy beklemtoon verder dat sowel die tipografiese konvensies as die leksikografiese simbole wat in die woordeboek voorkom, in die gebruikersinstruksies verklaar moet word.

Alhoewel TW 'n heelwat vollediger voorwoord as GW het, ontbreek daarin ook heelwat inligting. Daar word kortliks verwys na die tipografiese merkers in die woordeboek, asook vermeld waar die gebruiker 'n verduideliking van die stelsel sal aantref. 'n Tekortkoming is weer eens die versuim om die aard en omvang van die makrostruktuur te omskryf. Daar word nie aangedui watter tipes leksikale items opgeneem is of op grond waarvan hulle ingesluit is nie. Die redakteur vermeld wel dat heelwat vernuwing moes plaasvind, waaronder die aanbieding, tipografie en voorkoms van die makrostruktuur, maar omtrent die opnamebeleid word niks gesê nie.

Soos GW het TW ook in sy makrostruktuur (maar op 'n groter skaal as GW) gebruik gemaak van die opname van subleksikale items. Nêrens in die bewerking van die subleksikale lemmas is egter 'n aanduiding gegee omtrent die dekoderende en enkoderende funksie daarvan nie. Die gebruiker kan die indruk kry, uit die stilswye van sowel GW as TW, dat slegs dié subleksikale lemmas wat in die

woordeboek voorkom, produktief is in die taal of enigsins belangrik genoeg is om van kennis te neem.

5.2.2 Toeligting van die tipografie

Die gebruik van struktuurmerkers in woordeboeke dra by tot 'n grootskaalse besparing van ruimte en 'n verhoging van die teksdigtheid van die woordeboek. Hoe hoër die teksdigtheid van die woordeboek, hoe minder eksplisiet en toeganklik is die inligting wat aangebied word. Die gevolg is dat daar hoë eise aan die gebruiker gestel word, en té hoë eise kan lei tot swak kommunikatiewe oordrag. Daar is egter geen probleem met die gebruik van struktuurmerkers op sigself nie, solank hulle ondubbelsinnig en duidelik gedefinieer is én as sodanig aangewend word. Wiegand (1989: 428) verwys na twee tipes struktuurmerkers, naamlik tipografiese en nietipografiese struktuurmerkers. Hy beskou vetdruk en kursief onder andere as tipografiese merkers en tekens soos die dubbelpunt, komma en aanhalingstekens as nietipografiese merkers. 'n Aantal nietipografiese merkers word in sowel GW as TW gebruik, waaronder die tilde (~), die dubbelpunt (:) en die dubbele koppelteken (=). Die dubbele koppelteken en die koppelteken word om die beurt gebruik om subleksikale status in die woordeboek aan te dui. Dit geskied egter op 'n arbitrêre basis en verdien verdere aandag.

5.3 Die gebruik van die dubbele koppelteken in GW en TW

Subleksikale lemmas verteenwoordig op vormlike vlak by uitstek woorddéle. Hulle kan uitgeken word aan die aanwesigheid van 'n koppelteken as deel van die lemma — hetsy daarvoor hetsy daarna. Hierdie stelsel moet duidelik omskryf word in 'n woordeboek se toeligtingsteks sodat die gebruiker maklik kan onderskei tussen die verskillende tipes leksikale items.

Die dubbele koppelteken word in sowel GW as TW gebruik om woorddele of samestellings wat aanmekaar geskryf moet word, te onderskei van dié woorddele of samestellings wat met koppeltekens geskryf moet word. Die gebruik van die dubbele koppelteken is egter teenstrydig met die huidig aanvaarde manier van spel en skryf. Volgens die agste uitgawe van die AWS (1991: 34) word die koppelteken onder andere "gebruik as weglatingsteken in byvoorbeeld *Moeders- en Vadersdag* en as onderbrekingsteken aan die einde van 'n getikte of geskrewe reël, byvoorbeeld *ver-gadering*".[3] Maar om moontlike verwarring by die woordeboekgebruiker uit te skakel oor wanneer 'n woord aanmekaar geskryf word en wanneer 'n koppelteken gebruik word, kan die dubbele koppelteken bo die koppelteken aanvaar word, mits eersgenoemde slegs vir hierdie doel aangewend word en nie nog ander funksies verrig nie.

Die probleem is nou dat daar in sowel GW as TW ook van die dubbele koppelteken gebruik gemaak word om as deel van 'n lemma op te tree en sodoende sy status as subleksikale item te merk. GW en TW gebruik albei die dubbele koppelteken om nie net stamme en tegnostamme nie, maar ook affikse as subleksikale lemmas te merk. In hierdie gevalle vervul die dubbele koppelteken 'n twee-

ledige funksie: dit dien om 'n lemma as subleksikale item aan te dui én om die affiks as vasgeskrewe aan die stam te toon. Sowel GW as TW neem slegs prefikse as subleksikale lemmas op. TW voer egter nog 'n verdere gebruik van die dubbele koppelteken in, naamlik om fleksie-affikse aan te dui, byvoorbeeld:

ding *=e*
dioptrie *=trieë*
diploma *=s*
direk *=te =ter =ste*
diskonteer *ge=*

'n Ander voorbeeld van dubbelsinnigheid kom voor in die artikel van die neslemma **long** (A-E) in TW:

long [...] **~ont=**
steking pneumonia [...]

waar die dubbele koppelteken aandui dat **ontsteking** as een woord geskryf moet word. In hierdie geval merk die dubbele koppelteken nie **ont-** se subleksikale status nie.

5.3.1 Die inkonsekwente gebruik van subleksikale merkers in TW en GW

Die nietipografiese merker wat vir die aanduiding van affikse gebruik word, onderskei die affigale lemma vormlik van dié van die leksikale en die multileksikale lemma.

TW verwys in die E-A toeligtende aantekeninge na die optrede van 'n woorddeel as "trefwoord":

aero= aëro=. **~bat** — woorddeel gebruik as trefwoord
kunsvlieër [...]

Dit is streng genome 'n kontradiksie in terme waaruit onder andere dié woordeboek se woordgerigte benadering blyk. Meer gepas is Wiegand se terminologie, naamlik dat **aero=** 'n lemma-eksterne nesingang is met **~bat** die lemmadeel en sy plekhouersimbool. In die voorbeeld word die dubbele koppelteken gebruik om die subleksikale status van **aero=** aan te dui. In hierdie geval het **aero=** 'n dubbele funksie. Naas sy optrede as lemmadeel, tree dit ook op as lemma met *aëro=* as vertaalekwivalent. Voorbeelde van die inkonsekwente markering van subleksikale lemmas kom in sowel TW as GW voor. So word daar in TW, naas gevalle met die dubbele koppelteken, byvoorbeeld **her=** en **wan=** (A-E), en **arch=** en **extra=** (E-A), ook soms gevalle met die koppelteken, byvoorbeeld **al-** (A-E), en **ex-** en **non-** (E-A), aangetref. In GW kom daar ook, naas inskrywings met die koppelteken, soos **al-**[2] (A-E), en **post-**[3] en **non-** (E-A), ook meermale inskrywings met die dubbele koppelteken, soos **aller=**[3] (A-E) en **quasi=** (E-A), voor. Dit gebeur ook in sowel TW as GW dat die lemma soms 'n dubbele koppelteken bevat, maar die vertaalekwivalent slegs 'n koppelteken of geen teken nie (of omgekeerd), byvoor-

beeld in TW: **al-** all-, pan- (A-E), **all=** (E-A) en **eks-** ex, **oud-**[2] ... ex- (A-E); **ex=**[2] ... oud=, ... eks= (E-A), en in GW: **semi-** semi-, half- (A-E), **semi**[2] half=, semi= (E-A) en **post-**[3] na= (E-A).

Alhoewel die gevalle met die koppelteken volgens die huidige Afrikaanse spel- en skryfwyse korrek gebruik word, lyk dit volgens die stelsel wat in sowel TW as GW geld, asof die subleksikale lemmas waarin 'n koppelteken voorkom, daarmee geskryf moet word. Dit is juis nié die geval nie.

Nóg die voorwoord nóg die gebruiksaanwysings van óf GW óf TW maak voorsiening vir die gebruik van die koppelteken as deel van die lemmas van subleksikale items soos blyk uit die beperkte verwysings na die koppelteken en dubbele koppelteken in sowel GW ("~- waar 'n koppelteken gebruik word, bv. **see:** ~ **=eend** vir **see-eend** [...] = aan die end van 'n reël dui aan dat die betrokke woorddeel vas aan die volgende geskryf moet word") as TW: ("Die dubbele koppelteken (=) [is ingevoer] as afbrekingsteken by toevallige breuke aan die einde van 'n reël vir die koppeling van woorde wat sonder 'n koppelteken geskryf word. [...] Die koppelteken (-) binne 'n reël of aan die einde daarvan dui aan dat daar werklik 'n koppelteken moet wees.").

Wat in albei hierdie woordeboeke gebeur, is dat een tipografiese merker tegelykertyd twee of selfs méér funksies moet verrig. Dit is belangrik dat die leksikograaf 'n stelsel invoer wat konsekwent gehandhaaf en ondubbelsinnig deur die gebruiker geïnterpreteer kan word.

Die probleem is verder dat nie net affikse nie, maar ook tegnostamme (én stamme) deur middel van die dubbele koppelteken gemerk word.

5.3.2 Die inkonsekwente markeringstelsel van tegnostamme in TW en GW

Tegnostamme, net soos ander stamme, kan nie onafhanklik funksioneer nie. Dit word dus veronderstel dat hulle ook as subleksikale lemmas opgeneem behoort te word. Hulle kan weliswaar met mekaar verbind om tegnosamestellings te vorm, byvoorbeeld **higro-** en **-fiet** wat saam **higrofiet** vorm, maar hulle kan nie afsonderlik funksioneer nie. Om aan te dui dat tegnostamme[4], onder andere, met ander stamme moet verbind om as woord te kan funksioneer, word hulle tans soos volg in die makrostruktuur van TW opgeneem:

> **astro=: ~fisika** astrophysics. **~fisies** a. physical **~fisikus** a. physicist. [...]
> **higro=: ~fiet** hygrophyte. **~fobie** hygrophobia. **~graaf** hygrograph [...]

Die dubbele koppelteken merk die subleksikale status, terwyl die dubbelpunt die "soekelementstatus" van die betrokke tegnostam aandui, met ander woorde in die geval van **astro=: ~fisika** en **higro=: ~fiet** dien **astro=:** en **higro=:** as die soekelement vir die samestellings **astrofisika** en **higrofiet** onderskeidelik. Hierdie soekelement is die ingang tot die lemmanes of -nis wat lemmas bevat en bestaan uit die soekelement plus die daaropvolgende stam.

Hierdie stelsel waarvolgens tegnostamme in sowel TW as GW hanteer word, word nie in enigeen van dié woordeboeke se voorwoorde of gebruiksaanwysings

verduidelik nie. Soos in die geval van die affikse, word die tegnostamme ook nie in hierdie woordeboeke konsekwent gemerk nie.

Verder: In sowel TW as GW word tegnostamme aangetref wat met 'n dubbele koppelteken of koppelteken gemerk of selfs ongemerk is, van vertaalekwivalente voorsien is en soms ook, maar soms nie nes- of nisingange van lemmas vorm, byvoorbeeld in TW: **Anglo, bio=** en **pseudo=** (A-E), en **aero=, halo=** en **hema=** (E-A), en in GW: **hipo=** en **pseudo-** (A-E), en **hydro=** en **poly=** (E-A). Hierdie tegnostamme tree as selfstandige behandelingseenhede op waarby vertaalekwivalente dus buite samestellingsverband gevoeg word.

Vergelyk in TW:

> **pseudo=** pseudo=, bogus, professed, pretended.
> **aero=** aëro=. **~bat** kunsvlieër (kunsvlieënier), aërobaat. [...]

en in GW:

> **hipo=** under, below.
> **poly=** veel=.

Al sou die bestaande stelsel van soekelemente ook konsekwent deurgevoer word, sou dit nog nie die probleem oplos om tipografies te onderskei tussen die aangawe van affikse en van tegnostamme (én stamme) as lemmas nie. Wat egter belangrik is, is dat die leksikograaf 'n gekose stelsel bondig, maar volledig in die voorwoord of gebruiksaanwysings sal verduidelik sodat die gebruiker die inligting in die woordeboek probleemloos kan ontsluit.

5.4 Skematiese voorstelling van die artikeluitleg

Anders as in GW waar 'n lemma geïsoleer word van sy konteks en die struktuurmerkers dan verduidelik word, word in TW 'n skematiese voorstelling gebruik om die artikeluitleg van die woordeboek te verduidelik. In die kolom langsaan hierdie skema van die artikel word 'n verduideliking van die funksie van die betrokke struktuurmerkers en ander mikrostruktuurelemente gegee.

6. TW se stelsel van lemmadele in vetdruk en in kursief

TW se stelsel van lemmadele word nie net baie swak in die toeligting uiteengesit nie, maar word ook nie konsekwent in die woordeboek gehandhaaf nie. TW onderskei tussen twee tipes lemmadele, naamlik (a) dié in vetdruk en (b) dié in kursief, byvoorbeeld:

(a) **mikro=: ~analise**
(b) ***agter=: ~aan***

In die voorwerk van die A-E kant word hierdie metode soos volg verduidelik:

(a)	**mikro=**: **~analise** micro-analysis. **~analisator, ~analiseur** microprobe. **~balans.**	part of compound word used as head-word, grouping series of compounds
(b)	**eland** =*e* eland (*SA*); elk (*Eur.*); moose (*Am.*) **~hond** elkhound ***elands=***: **~boon(tjie)** [...] **~doring.**	*elands=*: italics indicate that headword follows a related previous headword **(eland / ~boon(tjie)** = **elandsboon(tjie)** in previous line **~hond** = **elandhond**

In die E-A kant se voorwerk word soos volg na die lemmadele verwys:

air: **~pageant** vliegskou-(ing), =vertoning. **~par-tition** lugskot, =afskor-ting. **~pas=.**	trefwoord kursief om te toon dis 'n ver-volg

Dit is opvallend dat dié verduidelikings oor die voorkoms en gebruik van lemma-dele betreklik vaag is. Eerstens bied die A-E kant 'n vollediger beskrywing aan as die E-A kant. Tweedens word die lemmadeel wat in vetdruk is, net aan die A-E kant verduidelik en nie aan die E-A kant nie.

By nadere ondersoek blyk dit dat die lemmadele in vetkursief hoofsaaklik beperk is tot leksikonitems wat as stamme optree. Verder, soos verduidelik in die gebruiksaanwysings, dien hierdie stamme as 'n vervolg op 'n leksikale lemma wat vroeër, meer spesifiek direk voor die eerste van 'n moontlike reeks soekele-mente in die makrostruktuur opgeneem is, byvoorbeeld:

agter [...] (leksikale lemma)
agter=: **~aan** [...]
agter=: **~deel** [...]
agter=: **~ingang** [...]
agter=: **~mekaar** [...]
agter=: **~saal** [...]
agter=: **~veld** [...]

Nog voorbeelde is ***invals=:*** onder **inval**, ***blare=:*** onder **blaar** en ***Gods=:*** onder **God**. Die gebruik van die dubbele koppelteken word nie aan die E-A kant geïl-lustreer nie, alhoewel dit tog wel ook daar voorkom, byvoorbeeld ***be=:*** **~~all** en ***logo=:*** **~gram**. Die rede hiervoor is Afrikaans se konjunktiewe en Engels se dis-junktiewe spelling waarby die produkte van kompositumvorming onderskeidelik vas en los geskryf word.

Verder blyk dit ook dat dié lemmadele wat in vetdruk verskyn, hoofsaaklik tegnostamme is, of anders gestel, as daar na die tegnostamme — soos in 3 hierbo aangebied — gekyk word, blyk dit dat tegnostamme (volgens die lemmadeel/soekelementmetode) hoofsaaklik in vetdruk opgeneem word omdat hulle nie

voorafgaande leksikale lemmas het nie. Voorbeelde soos **aëro=: ~baat, chiro=: ~graaf, heli=: ~blad** (A-E) en **homo=: ~centric, macro=: ~cephalic** en **ortho=: ~centre** (E-A) staaf dit. Ook in hierdie geval kom daar egter uitsonderings voor, byvoorbeeld **bisam=: ~hert, kaboe=: ~koring** en **kabouter=: ~haai** waar die lemmadele nie tegnostamme is nie, maar wel in vetdruk aangebied word, omdat hulle soos tegnostamme ook nie voorafgaande leksikale lemmas het nie.

Alhoewel hierdie stelsel van soekelemente in TW nie baie helder in die gebruiksaanwysings verduidelik word nie, blyk dit dat dit tog redelik konsekwent in die woordeboek deurgevoer word sodat die gebruiker dit sonder veel probleme kan volg.

7. Slot

In hierdie artikel is die opname en aanbieding van subleksikale lemmas in Afrikaanse vertalende woordeboeke ondersoek.

Weens die tot onlangs toe nog woordgerigte benadering van woordeboeke, word subleksikale lemmas baie swak opgeneem en aangebied in Afrikaanse vertalende woordeboeke. Dit dra onder andere daartoe by dat die woordeboeke nie baie gebruikersvriendelik is nie en nie 'n goeie enkoderende funksie verrig nie. Dit blyk verder dat stamlemmas volledíger opgeneem word as affigale lemmas — moontlik weens stamme se ooreenkoms met hul woordvariante. Affigale lemmas het nie ooreenstemmende woordvariante nie.

'n Verdere leemte wat uit die ondersoek blyk, is dat die voorwoorde van sowel TW as GW nie genoegsame gebruikersleiding verskaf nie. Die makrostruktuur van die woordeboek word glad nie omskryf nie. Die woordeboekgebruiker is dus in die duister oor watter tipes leksikale items in die makrostruktuur opgeneem is en watter nie. Die gebruik van struktuurmerkers word ook baie swak verduidelik.

Die leksikograaf moet besluit oor die opname al dan nie van produktiewe, minder produktiewe en onproduktiewe affikse en stamme in vertalende woordeboeke en in die voorwoord aandui wat die beleid rondom die opname van hierdie leksikonitems is.

Dit blyk dus uit hierdie artikel dat subleksikale lemmas in sowel TW as GW nie voldoende opgeneem en aangebied word nie.

Aantekeninge

1. Die aanduiding van die koppelteken teenoor die dubbelkoppelteken is die gebruik van die onderskeie woordeboeke om subleksikale status aan te dui.
2. 'n Tegnosamestelling is 'n samestelling waar die eerste of beide stamme van die samestelling 'n tegnostam is, byvoorbeeld **biometrie, argeologie** en **geograaf**.
3. In die sewende uitgawe van die AWS (1964) is daar wel voorsiening gemaak vir die gebruik van die dubbele koppelteken as skeidingsteken aan die einde van 'n reël.
4. Ook ander stamme wat 'n nes of nis inlei, word so aangebied.

Bronnelys

Woordeboeke

Bosman, D.B., I.W. van der Merwe en L.W. Hiemstra. 1984[8]. *Tweetalige Woordeboek/Bilingual Dictionary*. Kaapstad: Tafelberg.

Eksteen, L.C. 1997[14]. *Groot Woordeboek/Major Dictionary*. Kaapstad: Pharos.

Ander bronne

Beyer, H.L. 1995. *Die leksikografiese hantering van morfologies gemerkte geslagsopposisiepare in Afrikaanse woordeboeke, met spesifieke verwysing na die Verklarende Handwoordeboek van die Afrikaanse Taal*. Ongepubliseerde M.A.-skripsie. Universiteit van Stellenbosch.

Carstens, A. 1995. 'n Kritiese beskouing van HAT[3]. *Lexikos* 5: 138-165.

Combrink, J.G.H. 1990. *Afrikaanse morfologie*. Pretoria: Academica.

Gouws, R.H. 1989. *Leksikografie*. Pretoria/Kaapstad: Academica.

Gouws, R.H. 1991. Die leksikografiese hantering van woordgroepstamme. *Lexikos* 1: 113-127.

Hausmann, F.J., O. Reichmann, H.E. Wiegand en L. Zgusta (Reds.). 1989-1991. *Wörterbücher: Ein internationales Handbuch zur Lexikographie / Dictionaries: An International Encyclopedia of Lexicography / Dictionnaires: Encyclopédie internationale de lexicographie*. Berlyn/New York: Walter de Gruyter.

Hausmann, F.J. en H.E. Wiegand. 1989. Component Parts and Structures of General Monolingual Dictionaries: A Survey. Hausmann, F.J., O. Reichmann, H.E. Wiegand en L. Zgusta (Reds.). 1989-1991: 328-360.

Kempen, W. 1982. *Samestelling, afleiding en woordsoortelike meerfunksionaliteit in Afrikaans*. Goodwood: Nasou.

Svensén, Bo. 1993. *Practical Lexicography*. Oxford: Oxford University Press.

Taalkommissie. 1991. *Afrikaanse Woordelys en Spelreëls*. Kaapstad: Tafelberg.

Van Niekerk, A.E. 1989. *Die leksikografiese hantering van komposita*. Ongepubliseerde M.A.-skripsie. Universiteit van Stellenbosch.

Wiegand, H.E. 1983. Was ist eigentlich ein Lemma? Ein Beitrag zur Theorie der lexicographischen Sprachbeschreibung. Wiegand, H.E. (Red.). 1983a: 401-474.

Wiegand, H.E. 1983a. *Studien zur neuhochdeutschen Lexikographie III*. Germanistische Linguistik 1-4/82. Hildesheim: George Olms.

Wiegand, H.E. 1989. Der Begriff der Mikrostruktur: Geschichte, Probleme, Perspektiven. Hausmann, F.J., O. Reichmann, H.E. Wiegand en L. Zgusta. (Reds.). 1989-1991: 409-462.

Synonymy in the Translation Equivalent Paradigms of a Standard Translation Dictionary*

Phillip Adriaan Louw, *Department of Afrikaans and Dutch, University of Stellenbosch, South Africa*

Abstract: The norm in current canonical translation dictionaries with Afrikaans and English as the treated language pair is an undiscriminated grouping of partially synonymous translation equivalents. These are separated by commas as sole markers of synonymy. Lexicographers should reject this practice and embrace the view that absolute synonyms are just as rare as absolute equivalents. In most cases members of a target language synonym paradigm will be partial synonyms demanding some form of contextual guidance in order to distinguish them from other equivalents in the paradigm.

This article will focus on the motivation for the indication of partial target language synonymy. Two particular motivations will be discussed, as well as ways in which equivalent discrimination can be implemented.

The first motivation arises from a group of problematic phenomena that effect contextual divergence between the source and target language. Stylistic and register divergence should necessitate contextual guidance. Lexicographical labels are the most frequently used discriminators, but in South African dictionaries they are applied too sparingly and inconsistently. Other possible discriminators will also be discussed.

The most problematic motivation for the indication of partial synonymy is however different equivalents for various usages of a lemma. Ways in which equivalent discrimination can be implemented in these cases, will be discussed in detail.

Lastly, it will be shown that without a new, more effective method of indicating and ordering target language synonyms, none of the major changes that are pleaded for, will bear fruit.

Keywords: ABSOLUTE SYNONYMY, CONTEXTUAL GUIDANCE, EQUIVALENT DISCRIMINATION, LEXICOGRAPHIC LABELS, PARTIAL SYNONYMY, POLYSEMY, SENSES, STANDARD TRANSLATION DICTIONARY, SYNONYMY, TARGET LANGUAGE SYNONYMS, TARGET LANGUAGE SYNONYM PARADIGM, TRANSLATION EQUIVALENT PARADIGM, USAGES OF THE LEMMA

Opsomming: Sinonimie in die vertaalekwivalentparadigmas van 'n standaard vertalende woordeboek. Die norm in die huidige kanonieke vertalende woordeboeke met Afrikaans en Engels as die behandelde taalpaar is 'n ongediskrimineerde lysting van

* This paper was presented at the Second International Conference of the African Association for Lexicography, held at the University of Natal, Durban, 14-16 July 1997.

gedeeltelik sinonieme vertaalekwivalente. Hulle word deur kommas as die enigste merkers van sinonimiteit geskei. Leksikograwe behoort hierdie praktyk te verwerp en die siening te aanvaar dat absolute sinonieme net so skaars is as absolute ekwivalente. In die meeste gevalle sal lede van 'n doeltaalsinoniemparadigma gedeeltelike sinonieme wees wat die een of ander konteksleiding benodig om hulle van die ander lede van die paradigma te onderskei.

Hierdie artikel sal op die motivering vir die aanduiding van gedeeltelike doeltaalsinonimie fokus. Twee spesifieke motiverings sal bespreek word asook wyses waarop ekwivalentdiskriminasie geïmplementeer kan word.

Die eerste motivering is die gevolg van 'n groep problematiese verskynsels wat kontekstuele divergensie tussen die bron- en doeltaal veroorsaak. Stilistiese en registerdivergensie behoort konteksleiding te noodsaak. Leksikografiese etikette is die diskriminators wat die frekwentste gebruik word, maar in Suid-Afrikaanse woordeboeke word hulle te min en te inkonsekwent aangewend. Ander moontlike diskriminators sal ook bespreek word.

Die mees problematiese motivering vir die aanduiding van gedeeltelike sinonimie is egter verskillende vertaalekwivalente vir verskillende gebruike van 'n lemma. Wyses waarop ekwivalentdiskriminasie by hierdie gevalle geïmplementeer kan word, sal in detail bespreek word.

Laastens sal aangetoon word dat geen van die groot veranderings wat bepleit word, vrugte sal dra sonder 'n nuwe, meer effektiewe metode om sinonieme te merk en te orden nie.

Sleutelwoorde: ABSOLUTE SINONIMIE, BETEKENISONDERSKEIDINGS, DOELTAALSINONIEME, DOELTAALSINONIEMPARADIGMA, EKWIVALENTDISKRIMINASIE, GEBRUIKE VAN DIE LEMMA, GEDEELTELIKE SINONIMIE, KONTEKSLEIDING, LEKSIKOGRAFIESE ETIKETTE, POLISEMIE, STANDAARD VERTALENDE WOORDEBOEK, SINONIMIE, VERTAALEKWIVALENTPARADIGMA

Standard translation dictionaries with Afrikaans and English as treated language pair present the South African lexicographer with a unique challenge. Within the current dictionary culture and corresponding market, there is scarcely room for dictionaries aimed at a single language group. The demand is for a standard bilingual bidirectional translation dictionary addressing two language groups. Such a dictionary must be a practically viable linguistic aid as well as a cultural product that reflects the changing faces of the standard varieties of Afrikaans and English.

The defiance of the traditional typological plea for a single user addressing (following Ščerba's vital distinction between active (production, encoding) and passive (receptive, decoding) dictionaries), necessitates a fresh perspective on all the structures of the dictionary. Its most profound implications are however on a microstructural level. Current representatives in this class, such as *Tweetalige Woordeboek / Bilingual Dictionary* (henceforth BD) and *Groot Woordeboek / Major Dictionary* (henceforth MD) have not met these challenges adequately. Louw and Gouws (1996) have shown that innovative changes to the addressing procedures in these dictionaries can help the microstructure to become an effective key to communicative equivalence, as it should be. Yet the practical nature of such changes (which must, on a microstructural level, be manifested in con-

sistently applied systems of discrimination) has not been fully addressed. The implications for the treatment of some of the most salient lexicographical hurdles e.g. homonymy, polysemy and synonymy will require illumination. Synonymy, perhaps the most problematic and badly dealt with hurdle, will be discussed in this article.

Source language polysemy vs. target language synonymy

It is essential that any lexicographer should, as part of the translation dictionary's organisation, make a demarcation between source language polysemy and target language synonymy and make this demarcation accessible to the dictionary's users. The current system dictates that semicolons separate single translation equivalents or different target language synonym paradigms that can replace a lexical item in its different senses. Commas separate different target language synonyms within a given target language synonym paradigm or list of synonyms. This system is often confusing to many users who struggle to gauge this creative appropriation of everyday punctuation. Furthermore both meaning (especially sense) discrimination and equivalent discrimination that can show up contextual differences between target language synonyms, is applied too sparingly and inconsistently.

The result is that a distinction that should be crystal clear is muddled instead. Long lists of equivalents are given and the less competent user has little chance of knowing the significance of conventions (which are not explained in the explanatory introduction). The user also has little chance of making the right choice. It is small wonder that Kromann et al. (1991: 2724) commented in particular on the undiscriminated listing of quasi synonyms by calling this "one of the ancient and deadly sins of translation lexicography in bidirectional dictionaries".

Target language synonymy

Whereas discrimination between senses can be relatively easily maintained by means of a combination of sense discrimination and translation complements, equivalent discrimination in target language synonym paradigms is a more difficult matter. In order to come to a conclusion about the demands made on a lexicographer, target language synonymy and especially partial synonymy will be discussed in this paper. Examples from *Groot Woordeboek / Major Dictionary*, one of the bilingual desk dictionaries currently filling the gap the absence of a standard translation dictionary has left, will be examined. Firstly, reasons for the listing of target language synonyms will be surveyed.

Reasons for the listing of target language synonyms

Martin (1967: 156) gives two reasons for the listing of synonymous translation equivalents. Firstly a dictionary must "suggest to the translator a range of choi-

ces". Bogaards (1994: 613) expounds on this point of view by stating that a range of target language synonyms should be given, so that the user can be sure of finding "the one element that best fits the context". It is especially the failure of lexicographers of translation dictionaries to meet the practical requirements of this aim that has been severely criticised by metalexicographers and translators. See for example Martins' (1967: 156) own criticism, "sometimes the uncritical heaping of near synonyms is simply an evasion of responsibility on the part of the dictionary maker: unable to (or too little informed) to make up his own mind, he shifts the burden of choice to the user of the dictionary". If a general standard translation dictionary takes this procedure too far, it could waste valuable space. The lexicographer has to make sure that every synonym given is truly a functional translation equivalent for the lemma and make sure that the context of equivalence is clear.

Secondly, Martin states that "target language synonyms must be listed in order to give a clearer picture of the semantic spectrum of every item". His definition might however lead to a faulty assumption that there is necessarily some semantic difference between target language synonyms. This is only true though if the target language item is itself polysemous. Translation equivalents listed in a target language synonym paradigm (even if they themselves are polysemous lexical items), do not represent different senses of the lemma and can therefore not "give a clearer picture of a semantic spectrum of the lemma". Target language synonyms must display the full usage and contextual spectrum of the lemma. This approach necessitates a different view of synonymy as a whole, with the concept of partial synonymy as the point of focus.

Absolute vs. partial synonymy

In this article the view that there are few if any absolute synonyms in a language, is supported. On this point see Al-Kasimi (1977: 63) and Louw and Nida (1988: xv). The term *partial synonymy* is then used to show that contextual differences do exist between target language synonyms. Within the boundaries of this model, I shall focus on two of the possible motivations for the indication of partial synonymy in a target language synonym paradigm. The status quo in MD will be critically analysed and suggestions will be made on how to deal with this issue more adequately.

Contextual differences that require labelling

The first motivation for the indication of partial synonymy in a target language synonym paradigm still implies a relation of lexical divergence between the lemma and the target language forms. The lemma or specific sense of the lemma can be replaced by different translation equivalents, because there are two or more contextual nuances implicit in the source language form. It is even

possible that the target language may have a separate lexical item for each of these nuances.

Examples of such divergence which may require labelling (or even more contextual guidance) are stylistic and register divergence. The target language synonyms are therefore semantically equivalent to each other, but differ in style or register. Ideally they reflect a stylistic or register variation that is not implicit in the source language item, because this lemma is a neutral form that defies contextual boundaries and can accordingly be used in different registers and stylistic contexts. In translation dictionaries with Afrikaans and English as treated language pair, for example, this is however not always the case.

The ideal is approached in a case such as **pa** in the Afrikaans-English side of MD. **Pa** gets four synonymous translation equivalents, **pa**, **father**, **dad** and **daddy**. Each of these can be used in a slightly different context. The variation exists on the level of style (e.g. **father** vs. **dad**) and on the level of register (e.g. adult language in **father** vs. child language in **daddy**). MD rightly lists these translation equivalents as synonyms. The user here has to deal with a source language lexical item, **pa**, that can be used in different stylistic contexts and registers and has to make the right choice of equivalent to fit the source language context. The target language is so nuanced that various words can be used in this stylistic and register spectrum. The source language does however also contain several words (e.g. **vader**, **pappa**, **pappie**, **outop**, etc.) to cover this spectrum. The neutral term (in this case **pa**) can be translated with target language forms that belong to different styles and registers, but the more marked lexical items should be translated only with target language forms that are equivalent in style and register. This is not always the case, as will be shown in the discussion of pseudodivergence.

Where MD's information transfer also fails is in the clarification of the context of equivalence. No contextual guidance is given in the above-mentioned target language synonym paradigm. Gouws (1989: 204) discusses the usefulness of stylistic and other labels in detail. In this case, labels could have provided the necessary equivalent discrimination. Neither *Tweetalige Woordeboek/ Bilingual Dictionary* nor MD have a consistently applied system of stylistic labels with the translation equivalent as address. This is a weakness that depreciates the value of both these dictionaries as tools of empowerment in the search for communicative equivalence.

The same lack of consistency is found when there are differences of register caused by jargon. Afrikaans does not have an accessible medical term as partial synonym for **pitsweer**. As a result the lemma **pitsweer** in MD is given two synonymous translation equivalents: **furuncle** and **boil**. No label is however given to distinguish **furuncle** as a medical term from the more generally used term **boil**. For source and target language speakers alike, this procedure should be unacceptable. A consistently applied system of both lemmatically addressed and nonlemmatically addressed labels should be a priority in a truly innovative standard translation dictionary.

Pseudodivergence

Furthermore, should communicative equivalence be the primary aim of a standard translation dictionary, only truly functional equivalents should be given. As pointed out before, this is often not the case. In MD, for example, translation equivalents are given that are not functional equivalents of the lemma. In the article with the lemma **urinate**, the lexical item **fluit** (which also means "whistle" in Afrikaans) is presented as an unmarked translation equivalent and therefore as an absolute synonym for **urineer**. It is obvious though that **fluit** is a marked lexical item, which if chosen in any formal context, could have embarrassing results for any decoding or re-encoding English-speaking dictionary user not fully proficient in Afrikaans. The situation is even worse in a case such as the treatment of **bullshit** in MD, where a mixed presentation of vulgar and standard terms in one target language synonym paradigm can cause problems for any user.

> **bull: ~shit**, (vulg.), stront, kak, onsin, nonsies, nonsens ...

Far more care should be taken when dealing with sensitive lexical items (especially profanities, sexist language, etc.) and even standard language forms that refer to sensitive issues.

Further possible study

The discussion of this motivation for the indication of partial synonymy has focused on stylistic and register variation. Dialectic, temporal and other differences have not been discussed but should form part of any detailed future study dealing with partial target language synonymy. A study of this magnitude would also have to include an examination of the second motivation: different translation equivalents for different usages of the lemma.

Different translation equivalents for different usages of a lemma

From within lexical semantics there has always been a special effort to determine the precise meaning of each specific lexical item. This has led to a well-reasoned distinction between the senses (referring to semantic nuancing) of a lexical item as opposed to the usages thereof. These usages encompass contributing values from the extralinguistic context. Differences exist that cannot be shown up by means of a strict semantic analysis of the lexical item.

This distinction is well catered for in the Afrikaans and English monolingual lexicography. Whereas polysemy is indicated by means of a numerical, article internal system, different usages are given different letters of the alphabet as indicators. In cases where these usages fall within the scope of specific

senses, the letters combine with the numerical sense indicators.

Even though not consistently applied in monolingual dictionaries (possibly because of lack of space), this system is very relevant to translation dictionaries. The anisomorphism of languages often leads to a situation where a certain language has one item with different usages, but another language has a number of lexical items to display the same range of factors contributing from the extralinguistic context. These lexical items are usually presented as target language synonyms in a translation equivalent paradigm.

MD follows this principle in its treatment of the lemma **tjank.** Among others it lists **yelp, howl** and **whine** as synonymous translation equivalents. The differences between these items can rightly be described as subtle contextual nuances that reflect the different usages of **tjank**. MD yet again does not explicate the context by means of additional contextual guidance to its users. Consequently, this entry is of little use to source and target language users alike. The situation is worsened by the inclusion of **blub** and **bleat** in the target language synonym paradigm, when they obviously represent separate senses and not separate usages of the lemma.

The veiling of information created by the undiscriminated heaping of **yelp, howl** and **whine** can easily be avoided by giving contextual guidance to lead the user to the equivalent he/she needs. Only two types of discrimination that illuminate the context are discussed here. Firstly, a discriminator akin to the source language entries used as sense discrimination can be employed. This must preferably not be a full sentence, but ideally a phrase or a word that precedes the translation equivalent. If possible, this discriminator must also be presented in a different typeface or -size to the lemma or the translation equivalent. This will make the inner search route easier by countering confusion in the user's mind with regard to the inner access structure of the dictionary. In the article with the lemma **tjank**, for example, these discriminators could be used. "*kort en hard* ~" can be inserted in front of **yelp**, "*lank en hard* ~" in front of **howl** and "*lank en saggies* ~" in front of **whine**. The tilde represents the lemma. With these discriminators the necessary information is correctly given early on in the source language user's inner search route.

These entries have the lemma as primary address. They help to guide the dictionary user towards the correct usage of the lemma, thereby facilitating the choice of the correct equivalent. Yet the addressing structure is more complex than it originally appears to be. The discriminator itself is the primary address of the translation equivalent, which is then connected to the lemma only by means of a secondary lemmatic addressing procedure. Though primarily addressed to the lemma the discriminator, because of its interposition, provides valuable equivalent discrimination. The source language speaker therefore gets the best of both the source and the target language worlds.

This type of discrimination is particularly valuable in cases where one translation equivalent encompasses two or more of the lemma's usages. The discriminators can then for example be separated by "en"/"and". If the

lexicographer thinks a clearer distinction should be made, the translation equivalent can be listed more than once with separate discriminators. In most cases, however, this type of relisting which clashes with Haas's (1967: 46) "compactness desideratum", should not be necessary. See the treatment of **kring** in MD. **Circle** can act as translation equivalent for the lemma in several of its senses and source language usages. A possible source language usage discriminator that precedes **circle** could be: **a** en **b** "*sirkelvormige lyn*" en "*persone, diere wat 'n sirkel vorm*". The phrases used here are taken from the HAT (Odendal 1994: 584).

The first type of discriminator targets the source language user. The needs of the mother tongue speaker of the target language should, however, not be ignored. A system of context words or phrases in the target language acting as translation complements can be useful equivalent discrimination. They can be used either in isolation, or in combination with the previously mentioned source language discriminators. The context word or phrase is then addressed directly to the translation equivalent and the translation equivalent to either the lemma or to the preceding source language discriminator.

This system is introduced in the treatment of the lemma **die**[2] in MD. The translation equivalent **sneuwel** is discriminated from its synonymous partner **doodgaan** by means of the context phrase *(op slagveld)*. In this instance two usages of the lemma are lexicalised as two items in the target language. The first of these two items, **doodgaan**, is entered unmarked, because of its status as the primary equivalent. The omission of contextual guidance at the primary equivalent can only be practised within the South African context where a reasonably high level of bilingualism exists. The lexicographer of any future standard translation dictionary must keep the extent of his/her target user group carefully in mind, before this practice can be accepted uncritically. The omission of contextual guidance cannot benefit less competent bilingual users, but in opposition to this, valuable space could be saved by this omission.

However, the greatest problem in this dictionary seems to be the lack of consistency in the application of a workable system. Context words and phrases highlight usage divergence only sporadically in MD. The omission of guidance on usage creates problems in the case of sexually sensitive terms. If one proceeds from the politically and linguistically more correct presupposition that gender differences implicit to a single lexical item are usages rather than senses of that particular item, a lemma such as **teacher** should, for example, be given three equivalents, each with a separate discriminator: **onderwyser** (neutraal t.o.v. geslag), **onderwyseres** (vroulik) and **onderwyser** (manlik). The order in which these equivalents are listed, is based on intuitive frequency and is therefore not a final one. In MD the given equivalent **onderwyser(es)** receives no contextual guidance and does not reflect the usage trilogy. Furthermore, the use of brackets is never explained in the introduction or user's guide. As another example of this inadequacy, see MD's treatment of **Parisian**, where a similar usage trilogy exists.

The discussion of two types of discrimination should not at all be seen as an attempt at exhausting the field of study. Lexicographical examples, longer definientia, illustrations and menus are all options. Their functional value as discriminators between equivalents replacing different usages of a lemma still needs to be explored. The study of discrimination within target language synonym paradigms is still very new.

Indication and ordering

The discussion of different usages of the lemma raised another crucial issue besides discrimination. It is also necessary to find a good, consistently applied system of indicating meaning and contextual relationships in the microstructure. It is unacceptable, for example, to list near-absolute synonyms together with translation equivalents for the different usages of the lemma and separate them all by means of commas. Equivalents for different usages of the lemma lie on the border of partial synonymy because they cover large contextual differences not traditionally regarded to be of a semantic nature. In fact it is debatable whether one should for example call **howl**, **yelp** and **whine** synonyms at all, as treated in MD. The best approach seems to be the appropriation of the letter system from monolingual dictionaries along with the numerical system denoting senses. The letters will create new units with nonsemantic boundaries within which finer distinctions within the category of partial synonymy can be made. Semicolons can then for example be used to separate synonyms that display differences that can be shown up by means of labels. Commas can separate near-absolute synonyms. The reinterpretation of synonymy as a contextual instead of a strictly semantic phenomenon precludes the possibility of absolute synonymy in Afrikaans and English. No provision should be made for absolute synonyms. It is important that the lexicographer explains these changes made to the inner access structure as well as the changes to the article structure in an introduction or user's guide.

Within this system, equivalents will have to be ordered in structured sets according to empirical methods. All additional information aimed at the equivalents will have to be integrated within these sets or units. As has been shown elsewhere in the world, this is the only way to achieve functional equivalence and facilitate communicative success.

Conclusion

In this article only two motivations for the listing of partial synonyms were discussed. A better treatment of these and of issues such as target language polysemy is needed in future dictionaries. I have focused on the need for equivalent discrimination in the target language synonym paradigms. The lack of consistently applied systems of discrimination is one of the contributing factors

to the often failing information transfer in BD and MD. There is also a need to find better methods of ordering and indicating relations within target language synonym paradigms.

More careful attention must in future be paid to making the nature and extent of partial synonymy clear to the user. This plays an important part in an innovative approach that can make the microstructure a truly effective key to unlocking communicative equivalence.

Bibliography

A. Dictionaries

Bosman, D.B. et al. 1984[8]. *Tweetalige Woordeboek/ Bilingual Dictionary*. Cape Town: Tafelberg.

Kritzinger, M.S.B. et al. 1986[13]. *Groot Woordeboek/ Major Dictionary* . Pretoria: J.L. van Schaik.

Louw, J.P. and E. Nida. 1988. *Greek-English Lexicon of the New Testament Based on Semantic Domains*. Vol. 1. New York: United Bible Societies.

Odendal, F.F. 1994[3]. *Verklarende Handwoordeboek van die Afrikaanse Taal* (HAT). Doornfontein: Perskor.

B. Other sources

Al-Kasimi, A.M. 1977. *Linguistics and Bilingual Dictionaries*. Leyden: E.J. Brill.

Bogaards, P. 1994. Synonymy and Bilingual Lexicography. Martin, W. et al. (Eds.). 1994: 612-617.

Gouws, R.H. 1989. *Leksikografie*. Cape Town: Academica.

Haas, M. 1967. What Belongs in a Bilingual Dictionary? Householder, F.W. en Sol Saporta (Eds.). 1967: 45-50.

Hausmann, Franz J. et al. (Eds.). 1989-1991. *Wörterbücher. Ein internationales Handbuch zur Lexikographie/Dictionaries. An International Encyclopedia of Lexicography/Dictionnaires. Encyclopédie internationale de lexicographie*. Berlin: Walter de Gruyter.

Householder, F.W. and Sol Saporta (Eds.). 1967. *Problems in Lexicography*. Bloomington: Indiana University.

Kromann, Hans-Peder, Theis Riiber and Poul Rosbach. 1991. Principles of Bilingual Lexicography. Hausmann, Franz J. et al. 1989-1991: 2711-2729.

Louw, P.A. and R.H. Gouws. 1996. Lemmatiese en nielemmatiese adressering in Afrikaanse vertalende woordeboeke. *South African Journal of Linguistics* 14(3): 92-100, August 1996.

Martin, S.F. 1967. Selection and Presentation of Ready Equivalents in a Translation Dictionary. Householder, F.W. en Sol Saporta (Eds.). 1967: 153-159.

Martin, W. et al. (Eds.). 1994. *Euralex 1994 Proceedings*. Amsterdam: Free University Press.

Ščerba, L.V. 1995. Towards a General Theory of Lexicography. *International Journal of Lexicography* 8(4): 314-350, Winter 1995.

The Structure of an Afrikaans Collocation and Phrase Dictionary*

Anna Nel Otto, *Department of Afrikaans,*
Vista University, Port Elizabeth, South Africa

Abstract: In this article an Afrikaans collocation and phrase dictionary for mother-tongue speakers (primary target group) as well as advanced learners (secondary target group) is discussed. The position which such a dictionary occupies among other dictionary types is pointed out. A motivation is also given for the inclusion of idioms and other fixed phrases in the proposed dictionary. The three key approaches with regard to the interpretation of the term *collocation* are examined, i.e. the text-oriented approach of Halliday and Hasan (1976), the statistically-oriented approach of Sinclair (Collins Cobuild) and the significance-oriented approach of Hausmann (1984). The arguments in this article favour Benson et al.'s (1986) implementation of the significance-oriented approach. Statistical evidence could be used to examine the usage frequency of collocations and phrases. The advantages and/or disadvantages of these approaches are considered. Three types of words and their treatment in the dictionary are discussed: those which have a very wide range of combination, those which have selectional restrictions imposed by general semantic features, and those of which the range of combination is restricted by certain other words. It is argued that only the last two types should be included in this dictionary. As one of the target groups is unsophisticated learners with a limited grammatical background, the ideal would be to enter lexical collocations both at their bases and at the collocators. To save space however, more information such as examples could then be provided at the bases only. Grammatical collocations should be entered at the bases, i.e. nouns, verbs and adjectives. The division of the dictionary articles into two components to meet the needs of both intended target groups, is discussed.

Keywords: LEXICOGRAPHY, COLLOCATION DICTIONARY, LEXICAL COLLOCATIONS, GRAMMATICAL COLLOCATIONS, TRANSITIONAL COLLOCATIONS, IDIOMS, FREE COMBINATION, PROTOTYPE, SELECTIONAL RESTRICTIONS, BASE, COLLOCATOR

Opsomming: Die struktuur van 'n Afrikaanse kollokasie- en frasewoordeboek. In hierdie artikel word 'n Afrikaanse kollokasie- en frasewoordeboek vir sowel moedertaalsprekers (primêre teikengroep) as gevorderde aanleerders (sekondêre teikengroep) beskryf. Die plek wat so 'n woordeboek inneem naas ander woordeboektipes word uitgewys. 'n Motivering word ook gegee vir die insluiting van idiome en ander vaste uitdrukkings in die voorgestelde woordeboek. Die drie hoofbenaderings met betrekking tot die interpretasie van die term *kollokasie* word ondersoek, d.i. die teks-georiënteerde benadering van Halliday en Hasan (1976), die statisties-georiënteerde benadering van Sinclair (Collins Cobuild) en die betekenis-georiënteerde bena-

* An earlier version of this article was presented at the Second International Conference of the African Association for Lexicography, held at the University of Natal, Durban, 14-16 July 1997.

dering van Hausmann (1984). Die voor- en nadele van hierdie benaderings word oorweeg. Die argumente in dié artikel gee voorkeur aan Benson et al. (1986) se toepassing van die betekenisgeoriënteerde benadering. Statistiese gegewens sou gebruik kon word om die gebruiksfrekwensies van kollokasies en frases te ondersoek. Drie tipes woorde en hul hantering in die woordeboek word bespreek: dié wat met 'n baie wye reeks woorde kan verbind, dié waarvan die seleksie beperk word deur algemene semantiese kenmerke en dié wat slegs met sekere ander woorde kan verbind. Daar word geargumenteer dat net laasgenoemde twee tipes in dié woordeboek opgeneem word. Aangesien een van die teikengroepe ongesofistikeerde aanleerders met 'n beperkte grammatiese kennis is, is die ideaal dat leksikale kollokasies sowel by hul basisse as by die kollokators opgeneem word. Om ruimte te bespaar, kan meer inligting soos voorbeelde dan slegs by die basisse verskaf word. Grammatiese kollokasies behoort by die basisse, d.i. selfstandige naamwoorde, werkwoorde en adjektiewe, opgeneem te word. Die verdeling van die woordeboekartikels in twee komponente om in die behoeftes van al twee die bedoelde teikengroepe te voorsien, word bespreek.

Sleutelwoorde: LEKSIKOGRAFIE, KOLLOKASIEWOORDEBOEK, LEKSIKALE KOLLOKASIES, GRAMMATIESE KOLLOKASIES, OORGANGSKOLLOKASIES, IDIOME, VRYE KOMBINASIE, PROTOTIPE, SELEKSIEBEPERKINGS, BASIS, KOLLOKATOR

Introduction

Knowles (1997: 72) says: "It is a well-known but regrettable fact that very, very few language communities possess satisfactory collocations dictionaries ... The normal unavailability of collocations dictionaries is a great pity because that is exactly what advanced learners need, and indeed, what many native speakers hanker after too. In fact, it is not stretching things too far to say that first-class collocational control is the hallmark of the true L2 expert; collocational control is, of course, normally the last linguistic subsystem to be mastered by L2 learners who proceed to an advanced level."

Where some languages have several collocation dictionaries, Afrikaans has none. Several Afrikaans phrase dictionaries do however exist. In this article an Afrikaans collocation and phrase dictionary, which is presently being compiled, is discussed.

The place of the collocation dictionary among other dictionary types

According to Hausmann et al. (1989: XLII, XLIII) one can, in theory, differentiate between the following major syntagmatic dictionaries: the dictionary of syntactic patterns, the dictionary of collocations, the dictionary of set expressions and idioms, the dictionary of proverbs, the dictionary of quotations and the sentential dictionary. In practice, however, it is often difficult for the lexicographer to decide whether a certain word combination is a collocation or an idiom since certain collocations contain semantic specialized constituents. Cowie (1981: 230) comments in this regard: "Restricted collocations and idioms

are sufficiently related in terms of specialization of sense (of the part in the one case, of the whole in the other)." As the difference between collocations and idioms in this particular case is merely one of degree, this type of collocation can, within the cognitive approach, be regarded as nonprototypical idioms. Compare the following examples from Carstens (1992: 4): **flou** verskonings, **vuil** grappe, 'n **koue** blik, **onverteerde** feite. Benson (1989) uses the term "transitional collocation" for this category. If the lexicographer experiences problems with these distinctions, how can he/she expect the user to know whether one should look up a certain word combination in a collocation dictionary or in an idiom dictionary? This does not suggest that there is no need for separate dictionaries with regard to certain target groups — compare Benson (1989: 5) who believes that idioms should be entered in idiom dictionaries and important idioms in general-purpose dictionaries. According to him transitional collocations and technical collocations should be entered in collocation dictionaries. He (1990: 25-31) also maintains that our existing monolingual dictionaries should change and suggests the development of two types of monolingual dictionaries. The first is a monolingual decoding dictionary (MDD) that would include the largest possible number of "difficult" words and that would devote minimum space to collocations and the core vocabulary of a language. The second is a monolingual general-purpose dictionary (MGPD), intended for native speakers and learners who seek help with decoding and encoding language. "The learner who does not wish to use a learners' dictionary would find the MGPD ideal," Benson (1990: 27-28) argues. "Its decoding capability would be considerable, but, of course, would be less than that of the MDD. The encoding capability of the MGPD would be very strong, but it still could not compete with a specialized combinatory dictionary as a handbook for the production of texts." Every dictionary is written within a specific time and social framework for a specific target group. One could specifically compile a practical collocation dictionary (primarily) for advanced learners of Afrikaans (cf. Hausmann 1979, 1985), but which also contains the most frequently-used idioms and other fixed phrases or a theory-oriented collocation dictionary (containing only collocations) for linguists, language practitioners and lexicographers (cf. Mel'čuk and Žolkovskij 1984: 43, 73). A third option was chosen for the dictionary which is presently being compiled. The dictionary will contain collocations and phrases and will be directed not only at mother-tongue speakers as primary target group but also at advanced learners as secondary target group.

The meaning of the term *collocation*

The term *collocation* should however first be defined, since it gives rise to different interpretations.

For Firth (1957) *collocation* refers to a co-occurrence relation between individual lexical items, such as for example **dark night** and **you silly ass**. A certain

vagueness in the use of the term by Firth has given rise to a number of different interpretations, which can prototypically be identified as three key standpoints (cf. Herbst 1996: 380), namely

(a) a text-oriented approach (cf. Halliday and Hasan 1976),
(b) a statistically-oriented approach (cf. the Cobuild Project of Sinclair) and
(c) a significance-oriented approach (cf. Hausmann 1984: 398).

Herbst (1996: 380) evaluates the different approaches to collocation and comes to the following conclusions.

The text-oriented approach to collocation amounts to not much more than saying that in a text about coastal walking there is a certain likelihood for words such as **coast, sea, path, climb** or **steep** to occur as well. This kind of likelihood of co-occurrence of lexical items, however, seems to be determined to a greater degree by extralinguistic than by linguistic factors. The interpretation of collocation employed by Halliday and Hasan can probably be ignored. Hasan herself has shown that the usefulness of such an approach is limited (Herbst 1996: 383). It must also be doubted whether there is much point in using collocation for any kind of co-occurrence of two lexical items.

A purely statistical view of collocation as advocated by Sinclair seems problematical for a number of reasons. Firstly, there are the general problems involved in any kind of corpus analysis, especially regarding the representative nature of the material analysed. However, computer-assisted analysis may help overcome this problem. In this regard Smadja (1993) also suggests that a computer could be used to get a representative database. He points out that several approaches have been proposed to retrieve various types of collocations from the analysis of large samples of textual data. These techniques automatically produce large numbers of collocations along with statistical figures that reflect the relevance of the associations. None of these techniques provides functional information along with the collocations. Also, the results produced often contain improper word associations, i.e. not true collocations. Smadja (1993: 143-177) describes a set of techniques based on statistical methods for retrieving and identifying collocations from large textual corpora. These techniques produce a wide range of collocations and are based on some original filtering methods that allow the production of richer and higher-precision output. These techniques resulted in a lexicographical tool, *Xtract*. A lexicographical evaluation of *Xtract* shows that 80% of the identified collocations are correct. Church and Hanks (1990) and Church et al. (1991) emphasize the importance of human judgement used in conjunction with these tools. For the proposed dictionary the compiler's own database containing data (mainly from Afrikaans magazines) as well as the database of the publishers will be used.

The second problem that Herbst (1996: 383) points out with regard to a purely statistical view of collocation is that positional statements such as those produced by Sinclair (in Cobuild) are of limited value if one disregards the con-

text. Greenbaum (1974/1988: 115) illustrated, for example, that the occurrence of particular adverbs is determined by a number of factors. It must be doubted whether a purely statistical kind of analysis is able to accommodate the complexity of such factors.

Finally, there is the problem of the limited power of statistical statements. Is **dark night** for instance a significant collocation because nights tend to be dark and not bright?

The significance-oriented approach makes provision for gradience. Any attempt to define collocation in this narrow sense can thus only be aiming at defining a kind of prototype of collocation, by recognizing the gradience character of the distinction between collocations and free combinations on the one hand and between collocations and idioms on the other hand.

The macrostructure of the proposed dictionary

For the proposed dictionary on collocations and phrases the compiler decided to focus not only on the lexical and grammatical collocations as used by Benson et al. (1986), but also on semantic collocations.

There are words which have a very wide range, others where the selectional restrictions can be described through general semantic features, and words of which the range is restricted to certain other words. Svensén (1993: 102) uses the term *semantic collocations* for the second type of words.

The Afrikaans verb **aanvaar** (accept) can for example co-occur with the following nouns: **aanbod, argument, besluit, benoeming, beroep, betrekking, erfenis, gevolge, geskenk, gesag, hulp, jou lot, mosie, nederlaag, ooreenkoms, pos, resolusie, slagspreuk, siekte, uitdaging, skenking, uitnodiging, uitspraak, verantwoordelikheid, voorstel, wet, die bevel oor die regiment, ultimatum, die laste van die lewe, teenslae, smart, pyn** ...

The question arises: Should one include words like aanvaar which have a fairly wide range and if so, how should one treat this particular type of collocation?

The following options could be considered:

(1) one does not include collocations of this type,
(2) one includes the collocations just as one finds them in the data collection, or
(3) one includes the collocations and uses a system where one indicates that certain words act as hyponyms and/or one uses selectional restrictions where possible.

As regards option (1), these collocations should be included for a number of reasons:

(a) We live in a multilingual country where a large percentage of the people who speak Afrikaans, are not mother-tongue speakers of Afrikaans.

(b) Collocations are usually not directly translatable, e.g.

aanvaar bevel oor iemand/verantwoordelikheid — assume command/control/responsibility
aanvaar erfenis — enter upon inheritance
aanvaar pos — accept position
aanvaar aanbod/hulp — take up, accept offer/help
aanvaar die gevolge (van jou dade) — face (the consequences)
aanvaar nederlaag — take (a defeat)
aanvaar aanname, argument, besluit, dokument, mosie, oreenkoms, resolusie, uitspraak, voorstel, ens. — accept
aanvaar jou lot, die laste van die lewe, siekte, pyn, smart, teenslae — come to terms with

(c) Sometimes a verb has a synonym or synonyms, but one cannot use the synonym or synonyms in all contexts in the place of this verb, cf. **aanvaar** and **aanneem**:

bevel aanvaar /*aanneem
verantwoordelikheid aanvaar/*aanneem
pos aanvaar/*aanneem
aanbod/hulp/raad aanvaar // aanbod/hulp/raad aanneem
aanvaar die gevolge van jou dade/*aanneem
aanvaar nederlaag/*aanneem
aanvaar jou lot/pyn/smart, ens. /*aanneem

With regard to the use of synonyms mention should be made of a small part of the research that was conducted with 20 first-year students. Sixteen students had Xhosa as their mother tongue and Afrikaans as their third language. Three students had English as mother tongue and only one had Afrikaans as mother tongue.

The students had to choose between the two synonyms **behaal** and **bereik** in five sentences:

(1) Ek het die dorpie teen sononder (behaal, bereik).
(2) Sy het 'n oorwinning (behaal, bereik) oor haar teenstander.
(3) Ek is so bly dat ek my doel (behaal, bereik) het.
(4) Sy het groot sukses (behaal, bereik) met die kweek van hierdie rose.
(5) Hy het nou die hoogste sport in sy loopbaan (behaal, bereik).

In sentence (1) 15 students chose the correct synonym, i.e. **bereik**, and 5 students chose the wrong one. In sentence (2) 14 students chose the correct synonym, i.e. **behaal**, and 6 students chose the wrong word. In sentence (3) 13 students chose the correct word, i.e. **bereik**, and 7 students chose the wrong word.

(Of course **behaal 'n doel** is possible in Afrikaans, but then it is used in the context of sport.) In sentence (4) 14 students chose the correct word, i.e. **behaal**, while 6 chose the wrong one. In sentence (5) 10 students chose the correct word, i.e. **bereik**, and 10 students the wrong one.

Although **bereik** can combine with a wide range of words, one can therefore argue that one should include both **bereik** and **behaal** in the dictionary and have cross-references between them to help the user to choose the right word.

(d) Another motivation for including collocations is that one can present antonyms, again by using cross-referencing, e.g. **aanbod/uitnodiging aanvaar/van die hand wys** vs. **argument aanvaar/verwerp**, etc.

With regard to option (2), one could not consider this option because it implies that no other collocates exist other than those listed in the dictionary, which is not true. Compare the following sentence with **aanvaar:** "Die vooruitsig op 'n bleskop aanvaar Cora Marie (a cancer patient) nou gelate." One of the meaning distinctions of **aanvaar** could be: "berus in" (come to terms with) with the selection restriction "iets MOEILIKS of NEGATIEFS", followed by the most frequently used collocations. Therefore, one should include collocations with a wide range, provided that one combines this option with option (3): to use selection restrictions and/or hyponyms. An example of a hyponym could be **verantwoordelikheid aanvaar** which could then be replaced by e.g. **pligte, toesig**, etc.

One should, however, be very careful when deciding on the wording of selection restrictions. Carstens (1992: 4) states for instance that the verb **pleeg** (commit) is only selected in the presence of the meaning feature [+MISDAAD]/[+CRIME]. One does not however use **pleeg** only in the presence of the meaning characteristic [+MISDAAD], cf. Ek het ook 'n paar versies ('n skildery) gepleeg (*HAT*: 803). Furthermore, **pleeg** is not often used in combination with words indicating crime. Compare:

***molestasie pleeg** vs. **molesteer**
***aanranding pleeg** vs. **aanrand**
***verkragting pleeg** vs. **verkrag**
***roof pleeg** vs. **beroof**
***inbraak pleeg** vs. **inbreek, inbraak vind plaas**
***smokkelary pleeg** vs. **smokkel, smokkelary vind plaas**
***'n verkeersoortreding pleeg** vs. **begaan**

A third category of collocations includes words of which the range is restricted by other words, e.g. **dawerende applous, die onderspit delf**, etc.

Where should collocations be entered in the dictionary?

Hausmann's approach

Hausmann breaks down lexical collocations into a base and a collocator (1985: 119-121). In verb + noun collocations such as **brand stig** the noun is the base, and the verb is the collocator. In adjective + noun collocations such as **dawerende applous** the noun is once again the base, and the adjective is the collocator. In adverb + verb collocations such as **haarfyn beskryf** the verb is the base, and the adverb is the collocator. In adverb + adjective collocations such as **blakend gesond** the adjective is the base, and the adverb is the collocator.

In theory this works well with sophisticated learners who know the difference between the different parts of speech. However, apart from the fact that most users do not read the front matter of dictionaries, many learners struggle with parts of speech and even if they know the difference between for example a noun and a verb in theory, they sometimes do not know whether an individual word is a noun or a verb because they do not know the meaning of the particular word. For unsophisticated learners with a limited grammatical background, the ideal would be to have these collocations entered both at the base and the collocator. To save space however, more information, such as examples could perhaps be provided only at the bases.

Hausmann does not refer to grammatical collocations. However, on the basis of his approach we can, following Benson (1986: 6), assume that:

(a) if a grammatical collocation contains a noun, the noun is the base – **vertroue in, neem 'n eed dat hy dit sal doen, plesier om te werk, op jou stukke/ gemak**;
(b) if a grammatical collocation contains an adjective, the adjective is the base – **oortuig dat, geheg aan**;
(c) if a grammatical collocation consists of a verb and a preposition, the verb is the base – **dink aan, iemand herinner aan, jou vergryp aan**;
(d) if a grammatical collocation consists of a verb and a second verb in the infinitive, the first verb is the base — **besluit om iets te doen, geniet om iets te doen**.

The microstructure of the proposed dictionary

The dictionary article will be divided into two interactive components.

The first component

In the first component combinations will be placed under the different polysemous senses of the lemma (i.e. the base of the combination) without examples. In the case of collocations no definitions will be provided since collocations are

by definition transparent constructions (cf. Gouws 1989: 232).

Transitional collocations and idioms will however be provided with definitions, and labels will be used to indicate nonstandard forms.

Fixed expressions where the lexical base does not semantically relate to any of the listed senses of the corresponding lemma, will be included under a separate expression component.

Polysemous senses will be ordered according to parts of speech. The primary model which will be followed is the *BBI Dictionary of English Word Combinations*.

The second component

The lack of adequate examples and the unusual nature of some of the examples in the existing Afrikaans standard dictionaries are often pointed out by dictionary reviewers and metalexicographers (cf. Lombard 1992: 148-164). In this dictionary the current situation will be rectified. In the second component, typographically distinguished from the first, there will be an example for every combination mentioned in the first component. This will have an encoding function, especially for the secondary target group. Sentences from spoken and written Afrikaans will reflect real Afrikaans as it is currently used. The ideal will therefore be to use as many citations as possible; however, verbal illustrations will be used when the need arises to illustrate more than one information type in the same sentence (cf. Gouws 1989: 233). There will be a direct relation between the examples in the second component and the labels used in the first component.

The arrangement of combinations in the first component of the article will most probably require that the potential users should use their linguistic intuition; a strict alphabetical arrangement (by using secondary keywords) of examples in the second component should make lighter demands on the dictionary reference skills of the secondary users, for whom this component is especially intended.

A preliminary example is provided below:

kaf n 1. [OMHULSELS, DOPPE] ♦ 'n baal/hoop kaf; gebaalde/ongebaalde kaf; fyn/growwe kaf; stoppels kaf. ♦ die kaf van die koring/korrels/koringkorrels skei; die kaf uit die koring wan. 2. [ONSIN, BOG] ♦ 'n klomp/spul kaf; blote/louter(e)/pure kaf. ♦ kaf praat/kwytraak/verkoop/skryf/publiseer; nie/geen/g'n kaf duld nie; nie/geen g'n kaf vat nie (*informeel*); deur die kaf sny; iets as kaf afmaak. ✿ die kaf en/van die koring/korrels skei, die koring/korrels en/van die kaf skei, tussen die kaf en (die) koring/korrels onderskei, tussen die koring/korrels en (die) kaf onderskei [DIE GOEIE/WAARDEVOLLE VAN DIE SLEGTE/WAARDELOSE SKEI].

The following is part of the data on which the above article is based:

Die oudstaatspresident het 'n uiteensetting van sy aandeel in die hervomingsproses gegee wat selfs sy felste kritici nie sonder meer as kaf kan ***afmaak*** *nie. 0 Sowat 3 000* ***bale*** *kaf is in die brand verwoes. 0 Om ... as 'n oplossing vir die uiters moeilike konflik-situasie in Suid-Afrika aan te bied is* ***blote*** *kaf. 0 Diegene wat Hawke ken, weet ook dat hy geen kaf van enige speler, hoe groot die naam ook al,* ***duld*** *nie. 0 Die uitgetrapte graan en* ***fyn*** *kaf is na die windkant van die vloer gestoot om uitgewan te word. 0 Fischer se getuienis was die grootste* ***klomp*** *kaf wat die hof in 'n lang tyd gehoor het. 0 Die los koring is met houtgaffels teen die wind gegooi om die* ***koringkorrels*** *en kaf te skei. 0 Oor die jare heen is soveel goeie én slegte Amerikaanse programme uitgesaai, dat die kaf nie meer van die* ***koring*** *geskei kan word nie. Joernaliste het 'n rol te speel om die publiek te help om die kaf van die* ***koring*** *te skei, maar waarom publiseer hulle dan soveel kaf? 0 Hy het nie tussen die kaf en die* ***koring*** *onderskei in sy aanval nie. 0 Wanneer dit by aansprake op nuwe ontwikkelings kom, is die* ***korrels*** *maar dun tussen die kaf gesaai. 0 My dogter is tans besig met studie vir 'n M Sc-graad. 'n Aansienlike deel van haar studiemateriaal is net in Duits (wat sy nie ken nie) beskikbaar. Dit bring mee dat ek lywige stukke uit Duits moet vertaal. Natuurlik doen ek dit graag, maar dit sou soveel beter gewees het as sy dit self kon lees en dadelik die* ***korrels*** *van die kaf kon skei. 0 Dis tyd dat al die kaf wat oor die veiligheidsmagte gepraat word, van die* ***korrels*** *geskei word. 0 Hy kry alle vervelige kafpraters voor stok deur die kaf wat hulle in opgeblase taal* ***kwytraak****, in 'n splinternuwe konteks te plaas. 0 Mediabase het gesê hulle het nog "nooit sulke* ***loutere*** *kaf" gehoor nie. 0 Daar word baie kaf* ***gepraat*** *wanneer ouens se tonge by die 19de putjie los raak. 0 Enigiemand wat hierdie feit nie wil aanvaar nie, loop die gevaar om oor die Hugenote se aankomsgeskiedenis kaf te* ***praat*** *of te skryf, soos reeds meermale gebeur het. 0 Joernaliste het 'n rol te speel om die publiek te help om die kaf van die koring te skei, maar waarom* ***publiseer*** *hulle dan soveel van die kaf 0 As daar deur al die kaf na die kern van die saak heengesny word, blyk dit dat die sittende president met 'n paar hoofsaaklik prosessuele beperkings al elf lede van die hof sal kan aanstel. 0 "Ek daag mense uit wat volhou met dié* ***spul*** *kaf om in die openbaar die teendeel te beswys," het mnr. Peter Hendrickse gesê. 0 Saans het ons die* ***stoppels*** *kaf uit ons hare geskud. 0 Want kyk, die Chinese* ***vat*** *self nie kaf van liberale kabouters nie. 0 Hoe sal Suid-Afrika teen 2000 lyk? Soos 'n immergroen boom? Of soos kaf wat in die* ***wind*** *wegwaai? Dit sal baie afhang van hoe ernstig gelowiges in Suid-Afrika hul godsdiens gaan neem.*

With acknowledgements to Jana Luther

Conclusion

There is an important place in Afrikaans lexicography for a specialized collocation and phrase dictionary from which both mother-tongue speakers and advanced learners of Afrikaans can benefit. This dictionary should be compiled according to theoretical criteria, but the specific needs and skills of the target users should be taken into account.

Bibliography

Benson, M., E. Benson and R. Ilson. 1986. *The BBI Combinatory Dictionary of English.* Amsterdam: John Benjamins.

Benson, M. 1989. The Structure of the Collocational Dictionary. *International Journal of Lexicography* 2: 1-14.

Benson, M. 1990. Collocations and General-purpose Dictionaries. *International Journal of Lexicography* 3: 23-34.

Carstens, A. 1992. Kollokasies: vrye verbindings of lekseme? *Suid-Afrikaanse Tydskrif vir Taalkunde* 10: 1-11.

Church, K. and P. Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16: 22-29.

Church, K., W. Gale, P. Hanks and D. Hindle. 1991. Using Statistics in Lexical Analysis. Zernik, Uri (Ed.). 1991. *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon:* 115-164. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cowie, A.P. 1981. The Treatment of Collocations and Idioms in Learners' Dictionaries. *Applied Linguistics* 11: 223-235.

Firth, J.R. 1957. Modes of Meaning. *Papers in Linguistics, 1934-1951*: 190-215.

Gouws, R.H. 1989. *Leksikografie.* Pretoria/Kaapstad: Academica.

Greenbaum, S. 1974/1988. Some Verb-intensifier Collocations in American and British English. Greenbaum, S. 1988. *Good English and the Grammarian:* 113-124. London: Longman.

Halliday, M.A.K. and R. Hasan. 1976. *Cohesion in English.* London: Longman.

Hausmann, F.J. 1979. Un dictionnaire des collocations est-il possible? *Travaux de linguistique et de littérature* 17: 187-195.

Hausmann, F.J. 1984. Wortschatzlernen ist Kollokationslernen. *Praxis des neusprachlichen Unterrichts* 31: 395-406.

Hausmann, F.J. 1985. Kollokationen im deutschen Wörterbuch. Bergenholtz, H. and J. Mugdan (Eds.). 1985. *Grammatik im Wörterbuch*: 118-129. Tübingen: Max Niemeyer.

Hausmann, F.J., O. Reichmann, H.E. Wiegand and L. Zgusta (Eds.). 1989. *Wörterbücher/Dictionaries/Dictionnaires.* Berlin/New York: Walter de Gruyter.

Herbst, T. 1996. What are Collocations: Sandy Beaches or False Teeth? *English Studies* 4: 379-393.

Knowles, F.E. 1997. Collocability in Languages for Special Purposes (LSPs): Some Preliminaries. *Lexikos* 7: 70-93.

Lombard, F.J. 1992. Voorbeeldmateriaal in woordeboeke. *Lexikos* 2: 148-164.

Mel'čuk, I.A. and A. Žolkovskij. 1984. *Explanatory Combinatorial Dictionary of Modern Russian.* Vienna: Wiener Slawistischer Almanach.

Odendal, F.F. 1994. *Verklarende Handwoordeboek van die Afrikaanse Taal (HAT).* Midrand: Perskor.

Sinclair, J. (Ed.). 1987. *Collins COBUILD English Language Dictionary.* London: Collins.

Smadja, F. 1993. Retrieving Collocations From Text: Xtract. *Computational Linguistics* 19: 143-177.

Svensén, B. 1993. *Practical Lexicography.* Oxford/New York: Oxford University Press.

Zur Digitalisierung historischer Wörterbücher

Sven Dummer, Frank Michaelis and Michael Schlaefer,
Deutsches Wörterbuch, Akademie der Wissenschaften in Göttingen, Deutschland

Abstract: The textual and structural characteristics of printed historical German dictionaries call for special requirements in converting these works into computer-readable form. The diverse treatment of the articles requires a great deal of follow-up manual work since the often narrative structure of the texts limits automatic processing. The following article describes a series of experiments with Moriz Heyne's "Deutsches Wörterbuch" which were conducted to illustrate the limitations (but also the possibilities) of converting an historical dictionary into electronic media.

Keywords: ELECTRONIC DICTIONARY, HISTORICAL DICTIONARY, ELECTRONIC TEXT ENCODING

Zusammenfassung: Die textuellen und strukturellen Eigenschaften gedruckter historischer deutscher Wörterbücher stellen besondere Bedingungen für die Umsetzung dieser Werke in eine maschinenlesbare Form. Die differenzierte Erfassung der Artikel erfordert einen großen Anteil an manueller Nacharbeit, da die vielfach narrativen Textstrukturen eine automatische Bearbeitung nicht erlauben. Der folgende Beitrag beschreibt eine Reihe von Experimenten mit Moriz Heynes Deutschem Wörterbuch, die mit den Ziel durchgeführt wurden, Grenzen (aber auch Möglichkeiten) einer Umsetzung historischer Wörterbücher ins elektronische Medium zu veranschaulichen.

Stichwörter: ELEKTRONISCHES WÖRTERBUCH, HISTORISCHES WÖRTERBUCH, ELEKTRONISCHE TEXTKODIERUNG

Das Angebot elektronisch nutzbarer Fassungen von Drucktexten hat auf verschiedenen Ebenen in den letzten Jahren eine zunehmende Bedeutung gewonnen. Die z. B. auf CD-Rom verfügbaren Korpora literarischer Texte und Zeitungstexte besitzen inzwischen einen beachtlichen Umfang. Dabei haben die kommerziellen Angebote in ihrem Volumen und vom Niveau ihrer Aufbereitung vielfach die noch vor wenigen Jahren als innovativ geltenden linguistischen Textkorpora wie z. B. die des Instituts für deutsche Sprache in Mannheim überholt. Ein Förderprogramm der Deutschen Forschungsgemeinschaft "Retrospektive Digitalisierung von Bibliotheksbeständen" (1998) läßt für die kommenden Jahre eine wesentliche Verbreiterung der elektronischen Textbasis im wissenschaftlichen Sektor erwarten. Für sprach- und literaturwissenschaft-

liche Forschungen bedeutet eine solche Entwicklung eine qualitative Verbesserung der Forschungsvoraussetzungen insbesondere auf der Ebene der Korpuserstellung und der systematischen Textanalyse.

Die digitale Aufbereitung von Wörterbüchern ist im kommerziellen wie im wissenschaftlichen Bereich in den letzten Jahren ebenfalls verstärkt genutzt worden. Die Frage danach, welcher Wert dieser Art der Wörterbuchaufbereitung zukommt, läßt sich u. a. von den Benutzungsbedingungen elektronischer Wörterbücher ausgehend beantworten. Die praktische Nutzung elektronischer Wörterbücher ist im Unterschied zur üblichen Wörterbuchbenutzung an Arbeitsplätze mit EDV-Ausstattung und eine angemessene technische Verfügbarkeit des Materials gebunden. Die Benutzung elektronischer Wörterbuchfassungen wird unter den z. Z. geltenden technischen Bedingungen allein aus zeitökonomischen Gründen nur dann zu erwarten sein, wenn der Aufwand zur Erreichung des gedruckten Wörterbuchs und zum Nachschlagen von Wörterbuchinformation im Vergleich zur rechnergestützten Nutzung des Wörterbuchs erkennbar höher liegt. Um etwa in einem einbändigen gegenwartssprachlichen Wörterbuch den Artikel *lachen* mit einer punktuellen Frage zur Sprachproduktion nachzuschlagen, erweist sich die Benutzung eines gedruckten Handexemplars am Arbeitsplatz gegenüber dem Starten eines Rechners, dem Einlegen einer CD und der Durchführung einer entsprechenden Suche als der entschieden einfachere und zeitsparendere Weg. Anders dagegen sind die Arbeitsbedingungen zu beurteilen, wenn man im Rahmen einer Arbeitssequenz sehr häufig in einem oder mehreren umfangreichen Wörterbüchern nachzuschlagen hat und die Suchergebnisse arbeitsökonomisch festhalten möchte. Vor allem gilt dies für systematische Wörterbuchbenutzungen, z. B. zur Ermittlung von Wortbildungsreihen, bestimmten Synonymen usw. Bei solchen systematischen Suchen wird man nur mit sehr hohem Leseaufwand im gedruckten Wörterbuch zum Ergebnis kommen. Maschinenlesbare Wörterbücher können demgegenüber für diese Fragestellungen eine effiziente Unterstützung bedeuten.

Unter dem Gesichtspunkt der Literaturversorgung ist die digitale Wörterbuchform vor allem dann von Vorteil, wenn sie sehr umfangreiche oder alte, in den Bibliotheken nur begrenzt vorhandene Werke erschließt. Mit den digital verfügbaren Werken entsteht eine bibliothekarische Situation, die nicht nur den Weg zu verschiedenen Bibliotheken oder gar Fernleihbestellungen erspart, sondern eine generell höhere Arbeitseffizienz vor allem bei intensiver Wörterbuchbenutzung durch verbesserte Literaturversorgung schafft.

Außer in den bislang skizzierten Benutzungssituationen bei der Sprachproduktion oder Sprachbeschreibung kommt digitalisierten Wörterbüchern eine sehr wichtige Rolle in der Lexikographie und in der metalexikographischen Forschung zu. Strukturen von Wörterbüchern und damit letztlich auch deren Aussagewert lassen sich umfassend überhaupt nur mit angemessen aufbereiteten maschinenlesbaren Wörterbuchversionen beurteilen.

Zusammenfassend sind drei Gesichtspunkte zu nennen, unter denen

gegenwärtig digitale Wörterbuchversionen für den Benutzer von Interesse sind: zum einen geht es um die Verbesserung der Literaturversorgung, zum zweiten geht es um die Erschließung effizienter Zugriffe insbesondere bei systematischen Wörterbuchnutzungen, und zum dritten geht es um eine Verbesserung der Grundlagen für Erforschung, Planung und Durchführung von Wörterbüchern.

Die genannten Zielvorstellungen sind von sehr unterschiedlicher Auswirkung auf die Wahl der Digitalisierungsstrategien. Generell lassen sich hier die Möglichkeiten der Image-Erschließung und der sogenannten Volltexterschließung unterscheiden. Die Image-Erschließung beruht auf dem scannergestützten automatischen Erfassen eines authentischen Textbildes. Eine Buchseite z. B. ist dann analog zur xerographischen Kopie als kleinste operationale Einheit verfügbar. Ein inhaltlicher Zugriff ist bei diesem Verfahren nur in dem Umfang möglich, in dem er durch nachträgliche Indizierung der Seitenzahlen, Überschriften oder anderer inhaltlicher Einheiten erschlossen wird. Die Volltexterfassung erschließt demgegenüber jedes Einzelzeichen eines Textes und erlaubt die Suche nach allen im Text vorkommenden Zeichen oder Zeichenkombinationen. Die beiden Verfahren können unter den entsprechenden typographischen Voraussetzungen über das Bindeglied automatischer Texterkennungsprogramme (OCR) kombiniert werden. Komplizierte Druckbilder und viele Frakturschriften sind jedoch gegenwärtig mit Hilfe der automatischen Texterkennung nur sehr unvollkommen zu bearbeiten (Retrospektive Digitalisierung 1998: 46-48). Hier würde die Digitalisierung stets ein Abschreiben und Korrigieren älterer, nur in Buchform vorliegender Werke einschließen und damit eine gegenüber dem Scannen erheblich höhere finanzielle Investition bedeuten.

Da eine image-orientierte Wörterbuchbenutzung nichts wesentlich anderes bietet als die Lesbarkeit des authentischen Textes am Bildschirm, kann sie zwar als Lösung für eine verbesserte Literaturversorgung betrachtet werden. Da sie aber keine strukturierten Suchzugriffe erschließt, muß bei der Digitalisierung von Wörterbüchern zum Zweck systematischer Nutzung oder Analysen die Volltextversion als Standardlösung gelten. Eine Image-Digitalisierung von Wörterbüchern kann nur dann Priorität besitzen, wenn dies bei häufig genutzten Werken durch deren geringe Exemplardistribution oder durch konservatorische Interessen begründet ist. Angesichts der Investitionshöhe für eine Volltexterfassung müssen entsprechende Digitalisierungsvorhaben außer nach ihrem wissenschaftlichen oder praktischen Nutzen spezifisch danach beurteilt werden, mit welchem Aufwand welcher Grad an verbesserten bzw. erweiterten Nutzungsmöglichkeiten zu erreichen ist. Dazu ist kurz auf die bei digitalen Wörterbuchversionen angewandten datentechnischen Aufbereitungsmodi einzugehen.

Die auf dem Markt befindlichen maschinenlesbaren Wörterbücher (Milan 1998) stimmen darin überein, daß sie die Artikeltexte zeichen- und formatgetreu darstellen können. Die Möglichkeiten der rechnergestützten Zugriffe (Textretrival) beschränken sich meist auf die Zeichenebene bzw. zeichen-

abhängige Segmentbildungen. Die Suche kann so z. B. nur eingeschränkt auf die Grobsegmente wie "Stichwort" und "Artikeltext" durchgeführt werden, wenn andere zeichenabhängige Segmente nicht identifizierbar sind. Alle klassischen Suchoperationen werden von dem Programm implementiert: einfache Suche nach Zeichenfolgen, Suche nach Zeichenfolgenmustern (regulären Ausdrücken) und boolschen Operatoren wie UND, ODER, NICHT. Handelt es sich um eine Suche nach einer einfachen Zeichenfolge, so ist vielfach die interaktive Auswahl in einem Wortindex möglich.

Für einen qualifizierten systematischen Wörterbuchzugriff sind darüber hinaus jedoch auch weitere Suchkriterien zu erschließen, und zwar solche, die einen gezielten Zugriff auf Inhaltsstrukturen eines Wörterbuchartikels wie Belege oder Bedeutungsbeschreibungen erlauben.

An den skizzierten Zusammenhängen wird zweierlei deutlich. Zum einen erfordert eine systematische Wörterbuchnutzung die Möglichkeit, gewisse Artikelsegmente (Stichwort, Beleg) anhand bestimmter Kriterien auswählen zu können. Zum zweiten ist es erforderlich, daß diese Segmente in einer vom Artikel losgelösten Form dargestellt werden können. Besonders bei umfangreichen Artikeln, wie sie zum Beispiel das Grimmsche Wörterbuch zu bieten hat, ist diese Reduzierung der als Ergebnis gelieferten Textmenge nicht nur ein wünschenswerter Komfort, sondern eine für die systematische Benutzung notwendige Voraussetzung.

Ein häufiger gewählter Weg, einen Text für differenzierte elektronische Zugriffe vorzubereiten, ist die Textkodierung mittels einer Auszeichnungssprache wie zum Beispiel SGML (Standard Generalised Markup Language). Dabei werden Textsegmente mittels Einklammerung in "Tags" gebildet — was man als "Markup" oder "Text-Auszeichnung" bezeichnet. Die "Tags" entsprechen einem Element des für diesen Text entworfenen Strukturmodells, das in der sogenannten "Document-Type-Definition" (DTD) vereinbart wurde. In der DTD werden die notwendigen und optionalen Elemente festgelegt und ihre Abhängigkeiten zueinander beschrieben. Ferner ist es möglich, für jedes Element Attribute zu vereinbaren. Ein Element "Stichwort" könnte beispielsweise durch ein Attribut "Wortart" näher bestimmt werden.

Die Umsetzung lexikographischer Druckprodukte ins elektronische Medium erfolgt bisher offensichtlich durchweg ohne besondere lexikographische Bearbeitung. Die zugrundeliegenden Strukturmodelle werden auf der Basis dessen entworfen, was technisch mit geringem Aufwand machbar erscheint. Es zeigt sich sehr deutlich, daß zwar durch die Digitalisierung vorhandene lexikalische Datenbestände in ein neues Medium übertragen werden, daß aber die lexikographischen Implikationen der Textversion auch in der elektronischen Version vielfach bestimmend bleiben. Die Digitalisierung bewirkt eine technische Zugriffsverbesserung, nicht die Erstellung neuer Datenbestände und nur sehr begrenzt den Zugriff auf neue lexikographische Organisationsstrukturen. Art und Umfang des verbesserten Zugriffs hängen daher maßgeblich von der gliederungs- und drucktechnischen Aufbereitung des vorhandenen Wörterbuchmaterials ab. Zeigt das Printprodukt konsequente lexikographische Struk-

turen, die sich in der typographischen Textoberfläche angemessen spiegeln, bestehen günstige Voraussetzungen für automatisch erschließbare inhaltliche Zugriffe. Exemplarisch sei hier auf Wörterbücher wie den Robert Électronique (1991) verwiesen. Die Möglichkeit, Artikelstrukturen wie die Gliederungshierarchie auszufiltern oder modulartig isolierbare Paradigmen wie z. B. das aller Stichwörter oder aller Belege zu erstellen, das Vorhandensein entwickelter Variantensuchmöglichkeiten für Wortformen sowie Exportmöglichkeiten für gewünschte Ausschnitte schaffen für die Benutzer auf der datentechnischen Ebene wünschenswert günstige systematische Arbeitsvoraussetzungen, auch wenn man sich vieles, vor allem die Exportmöglichkeiten und die Menüoberflächen, wesentlich komfortabler vorstellen könnte. Andere maschinenlesbare Versionen von Wörterbüchern wie z. B. die des Duden-Universalwörterbuchs (o. J. Version 1.1) bleiben trotz relativ günstig scheinender Strukturbedingungen im Drucktext mit nur sehr beschränkten Such- und Filtermöglichkeiten bei der systematischen Nutzung eher unbefriedigend.

Enthält ein gedrucktes Wörterbuch typographische Polysemien, gering strukturierte, diskursive Artikelbestandteile, implizite bzw. elliptische Darstellungsformen oder metasprachliche Varianten, wird der typographieabhängige maschinelle Zugriff erschwert bzw. durch die Uneindeutigkeit der Informationsklassen so unscharf, daß es nicht mehr sinnvoll ist, eine solche Version ohne Überarbeitung zu benutzen. Offensichtlich aus solchen Gründen ist bei der Erstellung der elektronischen Version des Wörterbuchs von H. Paul in 9. Auflage (1992) auf Strukturierungen weitgehend verzichtet worden. Eine systematische Benutzung der digitalen Version dieses Wörterbuchs ist dadurch nur mit erheblichem Umstand möglich. Der erreichte Standard bleibt gegenüber dem Beispiel des Robert Électronique kaum diskutabel.

Der Zustand der maschinellen Version des Paulschen Wörterbuchs wirft die Frage auf, ob und in welcher Weise die typanalogen wortgeschichtlichen deutschen Wörterbücher überhaupt sinnvoll zu digitalisieren sind. Als Repräsentanten dieses Wörterbuchtyps werden neben dem Paulschen Deutschen Wörterbuch die Werke von J. und W. Grimm (1854-1971) sowie von D. Sanders (1860-1865), F. L. K. Weigand (1909-1910) und M. Heyne (1890-1895) berücksichtigt. Mit Ausnahme des Paulschen Wörterbuchs und den ersten Teilen des Grimmschen Wörterbuchs sind diese Wörterbücher in neuerer Zeit nicht bearbeitet worden. Als materialreiche Hilfsmittel für philologische und sprachwissenschaftliche Arbeit erscheinen sie trotz ihres teilweise nicht unbeträchtlichen Alters und unverkennbarer wissenschaftsgeschichtlicher Bindungen immer noch unverzichtbar. Sie schlagen im Bereich der Verständnissicherung die Brücke von der Gegenwart in ältere Sprachzustände und bieten mit Belegen und Verwendungsbeispielen Anschauung und Materialgrundlage für weitergehende Fragestellungen. Ferner erlauben sie, die einzelnen Wörter und Wortverwendungen u. a. in semasiologischen, etymologischen, kulturgeschichtlichen und morphologischen Zusammenhängen zu betrachten. Die Benutzungsintensität im wissenschaftlichen Bereich ist daher relativ hoch einzuschätzen. Unter je spezifischen Vorstellungen von synchroner oder geschicht-

licher Systematik sind diese Wörterbücher primär für den Zugriff auf Informationen zu einzelnen Wörtern angelegt. Dargestellt wird in der Regel das in der einzelnen Wortgeschichte Spezifische. Das durchaus auch heute noch mit Gewinn nutzbare Inhaltspotential dieser Wörterbücher wird in den Druckversionen nachteilig durch eine im wesentlichen von Vorstellungen des 19. Jahrhunderts geprägte atomistische lexikographische Perspektive bzw. Benutzungserwartung beeinflußt. Die lexikographische Strukturkonsistenz ist ebenso wie die metasprachliche Konsequenz und die Ausführung von Vernetzungen in allen Werken bestenfalls ansatzweise entwickelt. Der Anteil frei umschriebener, elliptischer bzw. impliziter Information erweist sich als hoch. Diskursive Tendenzen überlagern oder durchkreuzen die unterschiedlich entwickelten Gliederungsansätze ebenso, wie die offensichtliche Veralterung vieler beschreibungssprachlicher Formulierungen Barrieren für einen systematischen Zugriff darstellen. Wortbildungsbezogene oder textbezogene Fragestellungen, die vergleichende Suche nach bestimmten Bedeutungsmerkmalen oder die Suche nach Wörtern und Verwendungsweisen u. a. m. sind lektüregestützt nur mit einem ganz erheblichen Such- und Interpretationsaufwand realisierbar. Nicht nur im Fall des Grimmschen Wörterbuchs stößt man bei diesem Verfahren durchaus auch an die Grenze der vernünftigen Relation von Aufwand und Ergebnis. Man befindet sich daher in der unglücklichen Situation, daß zwar ein respektabler Fundus an sprachgeschichtlichen Informationen zur Verfügung stünde, daß sich aber dieser Fundus strukturbedingt einer systematischen Nutzung der Printprodukte gegenüber abweisend verhält.

Die Voraussetzungen für eine Digitalisierung, mit der diese atomistisch-einzelartikelbezogene Benutzungsbarriere überwindbar wäre, erweisen sich angesichts der skizzierten inhaltlichen und formalen Textstrukturen als sehr kompliziert. Mit der bloßen elektronischen Spiegelung von Artikeloberflächen ist eine ernstzunehmende qualitative Verbesserung der Nutzungssituation für historische Wörterbücher nicht zu erwarten. Die Digitalisierung müßte hier kombiniert mit einer Restrukturierung durchgeführt werden. Darunter wird ein Komplex von lexikographischen Eingriffen verstanden, der fehlende oder defekte Segmentbildungen nachträgt bzw. ersetzt und Zugriffsebenen kennzeichnet, ohne die eine elektronische Kodierung der Artikeloberfläche weitgehend ineffizient bleiben muß. Legt man etwa die Standards zugrunde, die aus dem Robert Électronique abzuleiten wären, müßten bis zu 20 Hauptebenen mit zahlreichen Substrukturen segmentiert und klassifiziert werden. Eine solche Bearbeitung ist weder allein auf sprachwissenschaftlicher Grundlage noch allein mit informatischer Kompetenz durchzuführen, sondern erfordert zwingend einen interdisziplinären Ansatz.

Im Rahmen einer Arbeitsgruppe, in der sich Mitarbeiter der Arbeitsstelle Göttingen des Grimmschen Wörterbuchs zusammengefunden haben (H. Albrand, K. Casemir, S. Dummer, U. Härtel, F. Michaelis, M. Schlaefer, M. Schulz), konnten diese interdisziplinären Voraussetzungen geschaffen werden. Die Arbeitsgruppe hat verschiedene Experimente zur Erprobung von Möglichkeiten retrospektiver digitaler Erschließung historischer Wörterbücher durchgeführt.

Dazu wurden Teile des Wörterbuchs von M. Heyne digital erfaßt. Für die Auswahl dieses dreibändigen Wörterbuchs können u. a. seine gegenüber den einbändigen historischen Wörterbüchern höhere Stichwort- und Informationsdichte, die Bearbeitung von einer Hand und die vielfachen Strukturanalogien gegenüber dem Grimmschen Wörterbuch, aber auch gegenüber den anderen genannten historischen Wörterbüchern angeführt werden. Unter dem Blickwinkel der Übertragbarkeit der experimentellen Befunde auf typanaloge Wörterbücher schien das Heynesche Wörterbuch daher am besten geeignet.

Der Text des Heyneschen Wörterbuchs wurde mit einem herkömmlichen Textverarbeitungsprogamm authentisch abgeschrieben. Die benötigten Sonderzeichen waren verfügbar und nur in seltenen Fällen eigens herzustellen. Versuche, den gescannten Text mit Texterkennungsprogrammen (OCR) automatisch umzusetzen, mußten als zu fehleranfällig abgebrochen werden. Der digitalisierte Text wurde in Annäherung an die Originaltypographie formatiert. Angesichts der häufigen typographischen Wechsel ist die Formatierung ebenso wie die Texterfassung zeitaufwendig und fehleranfällig, was besonders sorgfältige Korrekturen erforderte. Das bedeutet etwa eine Verdreifachung der Investitionskosten gegenüber einer in anderen Fällen möglichen automatischen Texterkennung. Trotz der Beobachtungen einer Reihe typographischer Inkonsequenzen wurde zur Simulierung realistischer Arbeitsbedingungen das vorgefundene System beibehalten. Voraussetzungen für eine Konvertierung in ein Nur-Textformat bestehen. Eine exemplarische Gegenüberstellung des Drucktextes und des formatierten maschinenlesbaren Textes zeigen die anschließenden Ausschnitte.

Digitale Fassung

Aal, m. der bekannte Fisch; altes gemeingerm. Wort, ahd. mhd. âl, dunkler Herkunft. Plur. die aale, wenig gebräuchlich die äle: schleimecht fisch und ael Garg. 103; (lasz sie sich wenden wie aele in einer reusze Goethe im Götz, später in aale geändert). Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz. — aal auch von Aufgußtierchen aalähnlicher Form, essigälchen, kleisterälchen. — Zusammensetzungen: **Aalfang,** m. Fang der Aale. — **aalglatt,** ein aalglatter mensch. — **Aalquabbe, Aalraupe,** f. aalähnlicher Fisch. — **Aalreuse,** f. Reuse zum Aalfang. — **Aalstecher,** m. Gabel zum Anspießen der Aale beim Fang. — **Aaltierchen,** n. Aufgußtierchen.

Kopie des Originalartikels

Aal, m. der bekannte Fiſch; altes gemeingerm. Wort, ahd. mhd. âl, dunkler Herkunft. Plur. die aale, wenig gebräuchlich die äle: schleimecht fisch und ael Garg 103; lasz sie sich wenden wie aele in einer reusze Goethe, Götz, ſpäter in aale geändert). Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz. — aal auch von Aufgußtierchen aalähnlicher Form, essigälchen, kleisterälchen. — Zuſammenſetzungen: Aalfang, m. Fang der Aale. — aalglatt, ein aalglatter mensch. — Aalquabbe, Aalraupe, f. aalähnlicher Fiſch. — Aalreuſe, f. Reuſe zum Aalfang. — Aalſtecher, m. Gabel zum Anſpießen der Aale beim Fang. — Aaltierchen, n. Aufgußtierchen.

Die digital verfügbaren Textteile des Wörterbuchs wurden im weiteren analysiert und strukturiert. Resultate dieser Bearbeitung können hier nur exem-

plarisch angeführt werden. Die Beispiele aus dem makro- und mikrostrukturellen Spektrum sollen die Problematik des vorgefundenen Datenbestandes und den zur Kodierung erforderlichen Arbeitsaufwand verdeutlichen.

Die Makrostruktur des Heyneschen Wörterbuchs bietet auf der Stichwortebene neben den abgesetzten Stichwörtern erster Ordnung den Typ unabgesetzter Stichwörter zweiter Ordnung, denen Kompositionsstichwörter zur Einleitung von Nestartikeln, z. T. mit elliptischem Bestimmungswort, gleichgeordnet sind. Ferner wurden Verweisstichwörter von unterschiedlichem Status ermittelt. Unter den Stichwörtern erster und zweiter Ordnung bilden Präfixstichwörter und unmarkierte Homographen jeweils besondere Gruppen. Zu vielen Stichwörtern werden Varianten gebucht. Nicht selten handelt es sich dabei jedoch um eigenständige Wortbildungen. Der Stichwortstruktur sind auch nichtlemmatisierte Weiterbildungen im Artikelfuß zuzuordnen. Die folgende Tabelle listet einige der üblichen Vorkommen im Stichwortbereich auf.

Aal	Einzelstichwort 1. Ordnung
Jahren, jähren	Stichwortvarianten in der Stichwortgruppe
Hohle, f., in älterer Spr. = höhle	Stichwortvariante mit historischer Einordnung im Einleitungsteil
abängstigen (abängsten 17. Jh.)	Stichwortvariante als andere Wortbildungsform mit historischer Einordnung im Einleitungsteil
abfordern (**abfodern,** s. fordern)	Stichwortvariante mit Verweis auf Grundartikel
abkappen, ... abkappen	unmarkierter Homograph
Aalquabbe	Stichwort 2. Ordnung, Kompositionsgruppenwort
=brief	elliptisches Kompositionsgruppenwort
allerwelts=	Präfixstichwort
Angeklagte, m. s. anklagen.	Verweiswort, Verweisstichwort
auch für äsen, s. d.	versteckter Stichwortverweis im Artikeltext
abgelebt, abgelegen, f. ableben, abliegen.	Verweisstichwortgruppe
anderweitige hilfe, thätigkeit, nahrung, vorteile.	nichtlemmatisierte Weiterbildungen im Artikelfuß

Eine Identifikation der Elemente "Stichwort" bzw. "Stichwortverweis" nach typographischen Signalen oder artikelstrukturellen Positionsmerkmalen ist

nach diesem Befund nicht sicher möglich, sondern setzt eine kompetenzgestützte metalexikographische Entscheidung voraus.

Als zweites Beispiel der lexikographischen Strukturierung wird die Erstellung eines mikrostrukturellen Grundmodells der Heyneschen Artikel vorgestellt. Dem Artikelmodell kommt im Restrukturierungsverfahren die Aufgabe zu, Ordnungsrahmen für verschiedene Informationsklassen zu setzen und damit die Möglichkeit zu schaffen, artgleiche Angaben nach ihrem Status innerhalb des Artikels zu gewichten. Solche artgleichen Angaben liegen z. B. mit den Bedeutungsumschreibungen im Artikelkopf und im Artikelkern vor. Mit dem Bezug auf den jeweiligen Artikelteil läßt sich die scheinbare Klassenübereinstimmung jedoch differenzieren. Im Artikelkopf wird im Verständnis des Heyneschen Konzepts eine panchronische Grundbedeutung angegeben, im Artikelkern erscheinen im weiteren Sinn Segmente des polysemen historischen Inhaltsspektrums, die durch Belege oder explizite Angaben anderer Art ihre geschichtliche Konkretion erhalten. Weiterhin trägt die mikrostrukturelle Artikelmodellierung dazu bei, formale Bezugsebenen für die zunächst inhaltlich bestimmten Module zu schaffen und so deren lexikographische Operationalisierbarkeit zu unterstützen.

Das hier entwickelte Modell stützt sich vor allem auf entsprechende Analysen der Grimm-Lexikographie. Es werden autonome Artikel und abhängige Artikel unterschieden. Autonome Artikel enthalten idealtypisch einen Artikelkopf mit Stichwort, Wortartangabe und der Angabe einer generalisierenden Bedeutungsangabe, die im einzelnen einen sehr variablen Status besitzen kann. Dem Artikelkopf schließt sich fakultativ ein Einleitungsteil vorwiegend mit etymologischen, wortgeschichtlichen oder formalen Beschreibungen der Worteinheit an. Der für diesen Artikeltyp obligatorische Artikelkern könnte auch parallel zum Grimmschen Wörterbuch vielfach als "Bedeutungsteil" beschrieben werden, wenn man akzeptiert, daß Bedeutung hier in einem sehr ambivalenten Verständnis verwendet wird und in die Bedeutungsbeschreibung zahlreiche andere Beschreibungsebenen integriert werden. In einem dem Bedeutungsteil folgenden fakultativen Fußteil der Artikel finden sich Aufzählungen, Verweise und Vergleiche.

Aal, m. der bekannte Fisch;	Artikelkopf
altes gemeingerm. Wort, ahd. mhd. âl, dunkler Herkunft. Plur. die aale, wenig gebräuchlich die äle: schleimecht fisch und ael Garg. 103; (lasz sie sich wenden wie aele in einer reusze Goethe im Götz, später in aale geändert).	Einleitungsteil
Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz. — aal auch von Aufgußtierchen aalähnlicher Form,	Artikelkern (Bedeutungsteil)
essigälchen, kleisterälchen. —	Artikelfuß
Zusammensetzungen: **Aalfang,** m. Fang der Aale. —	Wortbildungsgruppe

aalglatt, ein aalglatter mensch. — **Aalquabbe, Aalraupe,** f. aalähnlicher Fisch. — **Aalreuse,** f. Reuse zum Aalfang. — **Aalstecher,** m. Gabel zum Anspießen der Aale beim Fang. — **Aaltierchen,** n. Aufgußtierchen.	

Dem Fußteil ließen sich auch Heynes Kompositionsgruppen zuordnen. Da diese Kompositionsgruppen unbeschadet ihrer Einleitung mit der Überschrift *Zusammensetzungen* jedoch nicht durchgängig Komposita enthalten und nicht alle Artikel auch als abhängige Artikel beschreibbar sind, wird für das im weiteren verwendete Artikelmodell die Wortbildungsgruppe als eigenes Segment behandelt, in dem Artikelmikrostruktur und Makrostruktur der Stichwortreihe verzahnt sind.

Die abhängigen Artikel sind dann deutlich zu erkennen, wenn sie in einer Wortbildungsgruppe angeschlossen werden oder außer dem Stichwort und der Wortartangabe nur rudimentäre Angaben enthalten und nur unter Bezug auf das Stichwort des vorangehenden autonomen Artikels letztlich verständlich sind:

Aaltierchen, n.	Artikelkopf
	Einleitungsteil
Aufgußtierchen.	Artikelkern (Bedeutungsteil)
	Artikelfuß
	Wortbildungsgruppe

In vielen Fällen verschwimmt jedoch diese Grenzziehung. Es erscheint daher am ehesten sinnvoll, alle Artikel, die einer Überschrift *Zusammensetzungen* folgen, den abhängigen Artikeln zuzuordnen. Die Bedeutungsangaben sind nicht sicher in Analogie zu den selbständigen Artikeln zuzuordnen, so daß eine Mehrfachansetzung erwogen werden könnte.

Die mikrostrukturelle Analyse wurde unter dem Blickwinkel der verschiedenen Beschreibungsebenen für Struktur, Status und Verwendung der als Stichwörter angesetzten Sprachzeichen weitergeführt. Eine als Ergebnis dieser Analyse vorgenommene Modellierung des vorgefundenen Informationsprofils kann für das weitere Vorgehen folgende Schichten unterscheiden:

Angaben zu Zeichenkategorien und Zeichenstrukturen:	
Stichwort	Aal
Wortart	m.
Wortbildung	nur noch in Zuss. bergab, hügelab; Zusammensetzungen: Aalfang
Grammatik (Formbildung, Flexion, Syntax)	Plur. in der alten Spr. wie Sing.
Bedeutung	Unsinnlich, zur Bezeichnung eines Schwankens
Betonung	Konfékt

Angaben zum Ursprung, zur Herkunft des Zeichens:	
Etymologie	altes gemeingerm. Wort (...) dunkler Herkunft
Angaben zur Stellung des Zeichens, Zeichengebrauchs im deutschen Diasystem:	
Sprachsoziologie	in beschränktem Gebrauche auch in kaufmänn. Sprache
Stilistik	vielfach gemeines Schimpfwort; In dichterischer Freiheit bei BÜRGER
Sprachgeographie	nach der Zeit nur mundartlich (schwäb., schweiz.) und in Quellen die von der Mundart beeinflußt
Sprachgeschichte	Präp. mit Dativ, = von, jetzt von diesem verdrängt
Angaben zur Zeichenverwendung:	
Phraseologie	Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz.
Frequenz	wenig gebräuchlich die äle
Usualität	das Wort ist im 17. Jh. ungewöhnlich, in der Schriftsprache des 18. Jh. erst allmählich durch Beschäftigung mit dem mhd. wieder aufkommend
Kollokationen	königlicher aar; du junger aar
Dokumentation der Zeichenverwendung:	
Belege	Lessing Nath. 1, 3; alles moralische aas
Verwendungsbeispiele	ab und zu komme ich wohl dahin

Die hier bezeichneten mikrostrukturellen Sachverhalte bilden jeweils Klassen mit z. T. komplexen Subklassifikationsinventaren bzw. stark variablen Bezeichnungen für gleiche Sachverhalte. Bei den Wortarten unterscheidet M. Heyne weitgehend die üblichen morpho-syntaktischen Klassen. Mehrfachklassifikationen bei alternierender Wortart eines Etymons sind aber ebenso üblich wie das Aussparen der Wortartangabe bei Verben, z. T. auch bei Adjektiven. Während im Paradigma der Wortartangaben, abgesehen von den skizzierten Problemen, eine weitgehende Konsequenz in der Terminologie herrscht, werden z. B. stilistische oder diasystematische Sachverhalte vielfach frei diskursiv umschrieben. Der Sachverhalt der sprichwörtlichen Verwendung ist allein im Abschnitt A – Ab mit sechs verschiedenen Bezeichnungen angesprochen worden: *sprichwörtl.; sprichwörtlich; Sprichw.; Rechtssprichwort; im Sprichworte, in Sprichwörtern.*

Eine ähnliche Komplexität und Varianz der Substrukturen zeigt sich bei den Belegen. Idealtypisch bestehen die Belege aus einem objektsprachlichen Zitat, einer Autoren- und/oder Titelnennung und einem Stellenverweis. Der Belegbegriff wird jedoch sehr frei gehandhabt und führt so nicht nur zu einer

Fülle von Typvarianten, sondern auch zur vielfach diskursiv frei eingebetteten Belegform. Zur Veranschaulichung werden einige Beispiele zusammengestellt.

Belegbeispiel	Kommentar	Belegtypbez.
war doch der reiz der groszen arzneiflasche .. bald abgebraucht Immermann Münchh. 4, 121;	vollständiger Beleg mit Zitat, Autorennennung, Werknennung, Band- und Seitenangabe	Standardbeleg
flut abdämmen in einem Bilde Freiligrath 3, 123	Beleg mit Kurzzitat, diskursiver Einflechtung eines Interpretaments, Autorennennung, Band- und Seitenangabe	verkürzter Standardbeleg
als Kunstwort des Festungsbaus 1729 verzeichnet (abdachung, schiefe eines walles Hederich)	Wörterbuchbeleg mit seltener Nennung des Erscheinungsjahres und der sonst üblichen Beschränkung auf die Autorennennung	Wörterbuch-nachweis
die geschlachtete gans, das schwein, ferkel (Seume Spaz. 1, 96)	Verwendungstyp mit Autorennennung, Werknennung und Stellenangabe für einen elliptischen Zitattext	Vorkommens-nachweis nach Verwendungstyp (ohne Zitat)
des lebens mai .. mir hat er abgeblüht Schiller	Zitat mit Autorennennung ohne identifizierende Text- und Stellenangaben	Zitat mit Autorenidentifikation
unterschied zwischen dem letzten thaler, den man borgt, und zwischen dem ersten, den man abbezahlt Goethe Unterh. deutscher Ausgew.	Zitat mit Autorennennung und Textangabe, jedoch ohne identifizierende Stellenangabe	Zitat mit Werkidentifikation
allnächtlich, Adv. alle Nächte vorkommend oder wiederkehrend (Bürger Pfarrers Tocht. v. Taub.).	Autorennennung, Werknennung jedoch ohne Stellenangabe	zitatlose Werkidentifikation

Ein generelles Problem der Restrukturierung ergibt sich für die Segmentbildung überall dort, wo die diskursive Textform in einem syntagmatisch nicht trennbaren Zusammenhang mehrere Informationsklassen verschachtelt. Hier sind zahlreiche Mehrfachklassifikationen nötig, wenn später maschinell Textsegmente gebildet werden sollen, die die nötige Informationsautonomie besitzen. Dazu kommt, daß gerade die diskursive Verknüpfung immer wieder einen

Rückgriff auf den Gesamtartikel nötig macht, wenn die metasprachlichen Textteile narrativ zusammenhängend formuliert wurden und nur in Anmerkungsart durch Belege oder Verwendungsbeispiele unterbrochen sind.

Das komplexe Ergebnis der in Ausschnitten angedeuteten Restrukturierung von Artikeln aus M. Heynes Deutschem Wörterbuch wird im Anschluß in einer TEI-konformen Auszeichnung des Artikels AAL m. demonstriert. Die Richtlinien der Text Encoding Initiative (TEI 1994) bieten auch für die Kodierung von Strukturen im historischen Wörterbuch ein außerordentlich entwickeltes Strukturierungs- und Klassifikationsinventar. Sie erlauben so eine äußerst umfangreiche Anreicherung der Textdaten mit Informationen. In welchem Maße der TEI-Anwender davon Gebrauch macht, bleibt ihm allerdings selbst überlassen. Die Entscheidung über den Umfang der Auszeichnungen hängt in erster Linie davon ab, welche Informationen in die Textdaten eingebracht bzw. welche bereits im Text enthaltenen Informationen explizit gekennzeichnet werden sollen, damit später gezielt auf sie zugegriffen werden kann. Wörterbuchartikel sind hinsichtlich ihrer inhaltsstrukturellen und typographischen Gestaltung ohnehin äußerst komplex; wenn im Falle von hochgradig diskursiven Wörterbüchern wie dem von M. Heyne die Struktur der Artikel zudem keinem festen Schema folgt, so erhöht sich die Schwierigkeit, eine auf alle Artikel anwendbare Auszeichnungsmethode zu finden. Zudem ist eine automatisierte Auszeichnung vieler Informationseinheiten (z. B. etymologischer Erläuterungen) unmöglich. Sind, wie es in diskursiven Texten häufig der Fall ist, verschiedene Informationseinheiten zudem noch ineinander verschachtelt und/oder elliptisch aufeinander bezogen, stellt sich sogar die Frage, inwiefern die Auszeichnung überhaupt sinnvoll ist. Vielfach lassen sich nur Bruchstücke der gesamten Einheit markieren — zumindest, wenn man nicht durch großzügig übergreifende Markierungen hohe Ungenauigkeiten und Redundanzen in Kauf nehmen will und so z. B. als Ergebnis innerhalb einer Markierung "Etymologie" auch noch drei andere, für die Etymologie unerhebliche Informationseinheiten zu finden wären. Ein durch die Markierungen ermöglichter gezielter Zugriff auf solche Fragmente ist oft unbefriedigend, und ihre Extraktion z. B. in einer Suchabfrage "stelle mir alle Etymologie-Angaben zusammen" macht bei einem zu hohen Maß an Fragmentiertheit oder Redundanz überhaupt keinen Sinn.

Enthält der elektronische Text dieselben typographischen Merkmale wie der gedruckte, kann zumindest die Auszeichnung von Artikelanfang und Artikelende sowie einiger typographisch besonders gekennzeichneter artikelinterner Einheiten automatisiert erfolgen. Voraussetzung hierfür ist freilich eine weitgehend genaue und fehlerfreie Erfassung der typographischen Merkmale des gedruckten Textes bei der Digitalisierung; bei Texten, die nicht gescannt werden können und manuell erfaßt werden müssen, bedeutet dies einen erheblichen Mehraufwand. Die Möglichkeiten sind freilich eingeschränkt, da die typographischen Merkmale selten exakt an inhaltliche Merkmale gebunden sind. Zumeist sind zumindest die Stichwörter in einer eigenen Schriftart gedruckt und können damit problemlos automatisch markiert werden, sobald

aber eine Schriftart verschiedene Informationseinheiten auszeichnen kann, stößt die automatisierte Auszeichnung an Grenzen.

Ein großer Teil der wünschenswerten Markierungen könnte also nur manuell erfolgen. Im folgenden sei ein Beispiel für eine sehr elaborierte Auszeichnung gegeben; es sei darauf hingewiesen, daß hier noch nicht einmal alle Möglichkeiten, die die TEI-Guidelines bieten, ausgeschöpft sind. Der eigentliche Text ist zur besseren Übersicht fett gedruckt:

```
<!DOCTYPE TEI.2 system 'tei2.dtd' [
     <!ENTITY % TEI.dictionaries 'INCLUDE' >
     <!-- .... -->
]>

<tei.2>

<!-- ... -->

 <teiHeader>
     <fileDesc>
     <titleStmt>
       <title>Artikel AAL aus dem Heyneschen WB</title>
     </titleStmt>
     <publicationStmt>
       <p>bislang unver&ouml;ffentlicht</p>
     </publicationStmt>
     <sourceDesc>
      <bibl>
         <author>Moriz Heyne<author/>
         <title>Deutsches W&ouml;rterbuch.</title>
         <imprint>
              <pubPlace>Leipzig</pubPlace>
              <publisher>Hirzel Verlag</publisher>
           <biblScope type=volume>Bd. 1.</>
              <biblScope type=edition>2. Aufl.</>
              <date>1905</date>
         </imprint>
       </bibl>
     </sourceDesc>
     </fileDesc>

 </teiHeader>

 <text>
 <body>

 <!-- ... -->

 <entryFree n='1'>

  <form><orth>Aal</orth></form>,
```

```
<gramGrp>
      <gen>m.</gen>
   </gramGrp>

   <sense n='1'>
      <def>der bekannte Fisch</def>;
   </sense>

   <etym>
      altes <lang>gemeingerm.</lang> Wort,
      <lang>ahd.</lang> <lang>mhd,</lang>
      <mentioned>&acirc;1,</mentioned>
      dunkler Herkunft.
   </etym>

   <gramGrp>
      <number>
        Plur. die aale, <usg type=plev>wenig
       gebr&auml;uchlich</usg>
         <eg><cit>
           <q>die &auml;le:
           schleimecht fisch und ael</q>
           <bibl>
               <author>Garg.</author>
                <biblScope type=page>103;</biblScope>
           </bibl>
         </cit></eg>

<eg><cit>
            <q>(lasz sie sich wenden wie aele in einer
             reusze</q>
            <bibl><author>Goethe</author> im
           <title>G&ouml;tz</title></bibl>
         </cit></eg>
             , sp&auml;ter in <q>aale</q> ge&auml;ndert).
      </number>
   </gramGrp>

    <usg type=reg>Redensarten</usg>
       <q>glatt, schl&uuml;pfrig, schleimig wie ein
aal;</q>

    <usg type=reg>Sprichw.</usg>
        <q>wer den aal h&auml;lt bei dem schwanz,
           dem bleibt er weder halb noch ganz.</q>

   &shy;

   <sense n='2'>
```

```
      <form>
        <q>aal</q>
      </form>

        <def>auch von Aufgu&szlig;tierchen
              aal&auml;hnlicher Form,</def>

      <eg>
         <q>essig&auml;lchen, kleister&auml;lchen.</q>
      </eg>

   </sense>

   &shy;

<!-- ... -->

</body>
</text>

</tei.2>
```

Eine solche Anreicherung des Beispieltextes mit Mark-Up ist, wie angedeutet wurde, äußerst zeitaufwendig, und der nötige Aufwand steht ganz sicher in keinem angemessenen Verhältnis zu den dadurch ermöglichten Resultaten. Allein die Festlegung eines für alle Artikel anwendbaren Auszeichnungsschemas würde bei dieser Ausführlichkeit monatelange Planungsarbeiten durch Fachpersonal, das sowohl lexikographisch geschult als auch mit den TEI-Guidelines vertraut ist, erfordern. Es gilt, einen mit vertretbarem Aufwand zu realisierenden Mittelweg zwischen Maximal- und Minimalauszeichnung zu finden und zu evaluieren, ob das, was rechnerunterstützt erreichbar ist, zu brauchbaren Ergebnissen führt.

Es wurde daher in einem zweiten Experimentabschnitt ein Verfahren angewandt, das mit Hilfe einer u.a. auf die typographischen Merkmale gestützten, weitgehend automatischen Auswertung des Wörterbuchtexts Indizes erzeugt, mit denen sich wiederum TEI-konforme Auszeichnungen in den Text einbringen lassen. Abgesehen davon bieten die Indizes ein eigenes Arbeitsinstrument, entscheidend ist aber, daß sie eine TEI-konforme Aufbereitung des Textes ermöglichen, so daß die Vorteile von SGML/TEI zum Tragen kommen — nämlich die Abfassung der Textdaten in einem international standardisierten, plattform- und softwareunabhängigen und alterungsbeständigen Format, aus dem sich mit entsprechenden Hilfsmitteln bei Bedarf andere Formate z. B. für eine Druckvorlage oder die Publikation im WWW erzeugen lassen.

Das modifiziert fortgesetzte Experiment geht weiterhin von einer vorlagengetreuen Abschrift des Heyneschen Wörterbuchs als Grundlage aus. Ziel ist im weiteren eine Strukturierung anhand der typographischen Merkmale.

Dabei kann die Zeichen- von der Formatebene unterschieden werden. Auf der Zeichenebene eignen sich prominente Zeichen, wie zum Beispiel das Leerzeichen oder der Strichpunkt, als Merkmale, auf der Formatebene Absätze, Einrückungen sowie der Wechsel von Schrifttypen.

Anhand des typographischen Merkmals "Absatz" läßt sich der Wörterbuchtext in Segmente zerlegen, die z. T. als Einzelartikel, z. T. aber auch als Reihen von Nestartikeln zu bestimmen sind. Ferner lassen sich bestimmte Zeichen als "Wortende" interpretieren, so daß es möglich ist, den Text automatisch in Wortsegmente zu gliedern. Für jedes Wortsegment gilt nun, daß es in einer und nur in einer Schrifttype dargestellt wird. Deswegen eignet sich das Merkmal "Schrifttype", die Wortsegmente entsprechend der in den Schrifttypen enthaltenen impliziten lexikographischen Informationen zu klassifizieren. Ein Wort ist entweder ein "Stichwort", ein "Verfassername", "Beschreibungssprache" oder "Objektsprache". Es ergibt sich folgende Mengenverteilung der segmentierten Wörter:

Element	gesamt	unterscheidbare Zeichenfolgentypen
Stichwort	8 550	8 490
Beschreibungssprachlich	218 740	23 620
Objektsprachlich	239 500	45 380
Verfassername	18 830	540

Im ausgewerteten Wörterbuchabschnitt A – E können 18 830 Wortsegmente als Verfassernamen identifiziert werden. Davon entfallen z. B. 2 490 auf den Zeichenfolgetyp "Goethe", 4 auf den Zeichenfolgetyp "Goethes". Insgesamt lassen sich 540 verschiedene Zeichenfolgetypen im Bereich der Verfassernamen unterscheiden. Es handelt sich um gerundete Werte, da man je nach den Zeichen, die man als Wortende interpretiert, zu leicht divergierenden Ergebnissen kommt. Außerdem ist zu beachten, daß durch typographische Systemfehler im Drucktext sowie durch Fehlkodierungen bei der Erstellung der digitalen Fassung mit einer Fehlerquote von ca. 15% zu rechnen ist. Das bedeutet bei rein maschineller Ausführung der Auszeichnung des Wörterbuchtextes nach typographischen Merkmalen einen nicht unbeträchtlichen Fehlerfaktor, der hier wie auch bei den im weiteren vorgestellten Verfahren die Grenzen solcher automatisierten Auszeichnungsverfahren zu erkennen erlaubt.

Kombiniert man die bisher angewandten Verfahren, so lassen sich weitere Segmentbildungen erreichen. So ist durch Suchen, die sich auf die Kombination der absatzorientierten und der schriftartbezogenen Auszeichnung beziehen, eine Identifikation der Nestartikel als Einzelartikel möglich. Auch können die Wortartangaben, soweit im Drucktext vorhanden, umgebungs- und formatbezogen ermittelt werden. Die häufigen Lücken in den Wortartangaben zwingen jedoch zum systematischen Prüfdurchgang und zur Ergänzung. Die

aufgrund des skizzierten Verfahrens gewonnenen Artikelsegmente lassen sich etwa in folgender Form darstellen:

Stichwort	Wort-art	Text
A	n.	Ausruf des Ekels (ein ä-geschmack Goethe im Satyros 1); Nachahmung des Kindergeschreis (ders., Künstlers Erdenwallen); des Räusperns, Stockens: viel akzion! viel — ä! ä! — was ich sage! Wieland Abd. 3, 6. vgl. b
Aal	m.	der bekannte Fisch; altes gemeingerm. Wort, ahd. mhd. âl, dunkler Herkunft. Plur. die aale, wenig gebräuchlich die äle: schleimecht fisch und ael Garg. 103; (lasz sie sich wenden wie aele in einer reusze Goethe im Götz, später in aale geändert). Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz. — aal auch von Aufgußtierchen aalähnlicher Form, essigälchen, kleisterälchen.
Aalfang	m.	Fang der Aale.
aalglatt	adj.	ein aalglatter mensch.
Aalquabbe	f.	
Aalraupe	f.	aalähnlicher Fisch.
Aalreuse	f.	Reuse zum Aalfang.
Aalstecher	m.	Gabel zum Anspießen der Aale beim Fang.
Aaltierchen	n.	Aufgußtierchen.

Die automatische Segmentierung des Wörterbuchtextes in Artikel und Wörter und die Auszeichnung der Wörter gemäß ihrer Typologie erlaubt einfache systematische Zugriffe. Es können z. B. alle Artikel bzw. Artikelpositionen aufgesucht werden, die das Wort *Redensart* als beschreibungssprachliches Element enthalten. Diese beschreibungssprachlichen Vorkommen sind von Lemmavorkommen oder objektsprachlichen Vorkommen des Wortes *Redensart* zu unterscheiden. Die bislang weitgehend automatische Segmentierung des Wörterbuchtextes ermöglicht damit eine nicht unerhebliche Vorauswahl. Weitere Experimente der maschinell gestützten Strukturbildung stützen sich auf die Beobachtung systematischer Abfolgen typographischer Elemente.

So erweist es sich als möglich, im Heyneschen Wörterbuchtext das Element "Beleg" teilweise automatisch zu extrahieren, da sich die Standardbelegform durch eine relativ stabile Formatregelung auszeichnet. Dem Zitattext folgt eine Verfassernennung, dieser wiederum eine Stellenangabe jeweils in eigener Typographie. Diese typographische Sequenz ist zwar nicht monosem, aber bei einer automatischen Ausfilterung dieser Formatsequenz überwiegen jedoch die gewünschten Belege. Die manuelle Aussonderung der belegfremden Textfolgen bildet daher kein erhebliches Hindernis hinsichtlich des Arbeits-

aufwandes. Mit diesem maschinell vorbereiteten Segmentierungsgang für die Standardbelege können etwa 40% des Gesamttextes separat erfaßt und inhaltlich als Beleg ausgezeichnet werden. Das Ergebnis des Arbeitsganges ist ein Belegmodul, das sich zudem leicht weiter nach Objektsprache, Verfassern und Werk-/Stellenangabe maschinell gliedern läßt. Die tabellarische Übersicht im Anschluß zeigt das Belegmodul:

Stich-wort	Belegtext	Autor	Werk-angabe
A	ich lêre in daz â bê cê; des enhât er niht mê noch gelernet wan daz â		Pf. Amis 297;*
A	ich bin das a und das o, der anfang und das ende		Offenb. 1, 8;*
A	wolt nit A sagen, auf dasz er nicht müsz B sagen		Garg. 247. 2*
Aal	die äle: schleimecht fisch und ael		Garg. 103*
Aal	lasz sie sich wenden wie aele in einer reusze	Goethe	im Götz*
Aas	wo aber ein asz ist, da samlen sich die adler		Matth. 24, 28*
Aas	wenn fürsten geyer unter äsern sind	Lessing.	Nath. 1, 3*
Aas	alles moralische aas	JPaul	uns. Loge 3, 42
Ab	wenn einmal die wurzel ab sei	JGotthelf*	Schuldenb. 183
Ab	fuhr auf und ab	Freytag	Ahnen 1, 374*
Ab	ging .. im zimmer auf und ab	Freytag	Ahnen 4, 383*
Ab	die auf dem schlosse ab und zu ritten	Eichendorff*	Taugenichts 50
Ab	weiter ab	Lessing	Nath. 1, 4
Ab	kam ab der post ein kistlein	Hebel	2, 101*
Ab	freud und lust an allem ab und an, an und ab dem kleeblatt holder kinder*	Bürger	(60a)
Ä	ein ä-geschmack	Goethe	im Satyros 1
Aas			1. Mos. 15, 11
Ä	viel akzion! viel — ä! ä! — was ich sage!	Wieland	Abd. 3, 6.*

Der nach Ausfilterung der Standardbelege verbleibende Wörterbuchtext stellt als verkürzte Lesefassung ein eigenes Bearbeitungsprodukt dar.

Stichwort	Wort-art	reduzierter Artikeltext
Aal,	m.	der bekannte Fisch; altes gemeingerm. Wort, ahd. mhd. âl, dunkler Herkunft. Plur. die aale, wenig gebräuchlich später in aale geändert). Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz. — aal auch von Aufgußtierchen aalähnlicher Form, essigälchen, kleisterälchen. Zusammensetzungen:
Aalfang,	m.	Fang der Aale.

Aas,	*n.*	totes, faulendes Vieh. Altes westgerm. Wort (ahd. mhd. âs, ags. æs), Ableitung der Wurzel az essen, ursprüngl. als Speise der Raubtiere oder Fütterung für Hunde, Falken gedacht, dann verallgemeinert. Plur. in der alten Spr. wie Sing. später äser (z. B. bei Lessing, Kant), jetzt auch die aase. Obwohl biblisches Wort ist es doch in gewählter und bildlicher Sprache lieber vermieden als gebraucht, da es vielfach gemeines Schimpfwort: du aas Mephist. zur Hexe in Goethes Faust. vgl. raben-, schindaas. Zusammensetzungen

Einer Programm-Oberfläche, die später auf das digitalisierte Wörterbuch aufsetzt, ist es nun möglich, zwei verschiedene Ansichten ein und desselben Artikels zu bieten: einen Kurzartikel, ohne Belege, und einen Vollartikel, mit Belegen. Außerdem kann ein separater Zugriff auf das Belegarchiv ermöglicht werden. Dieses Verfahren, dem Benutzer verschiedene Ansichten eines Artikels anzubieten, sowie der Zugriff auf ein vom Wörtbuch getrenntes Belegmodul haben sich in der Praxis bewährt. Als Beispiel dafür kann wiederum der Robert Électronique gelten.

Das Verfahren der bislang durchgeführten maschinellen Textgliederung führt zu Modulbildungen, die sowohl paradigmainterne Suchen als auch über eine entsprechende Verknüpfung die Rückkopplung auf den Originalartikel gestatten. Das angewandte Verfahren erfordert kontrollierende Nacharbeiten. Diese bleiben aber in einem überschaubaren Umfang und setzen keine umfängliche konzeptionelle oder metalexikographische Kompetenz voraus.

Um einen Überblick über das verwendete Wortmaterial zu erhalten, ist es möglich, die nach ihrem Format ausgezeichneten Wortsegmente je für sich in einem Index zu gruppieren. Um im Hinblick auf die angestrebte Transferierbarkeit des Arbeitsverfahrens einen möglicherweise nur in M. Heynes Wörterbuch vorliegenden Spezialfall der Formatsyntax nicht zum Anlaß unrealistischer Allgemeineinschätzungen werden zu lassen, umfaßt die Indexerstellung in den anschließenden Beispielen den Gesamttext einschließlich der Standardbelege in der Differenzierung nach Objektsprache, Metasprache und Verfassernamen. Alle drei Indizes erschließen den gezielten Weg von den jeweiligen Wortformen zu bestimmten Stichwörtern bzw. zu Artikelpositionen. Als Problem bleiben generell die flektierten Wortformen bzw. historisch variable Wortformen. Ein weiteres generelles Problem liegt in der Belastung der Indizes durch Massenwörter wie Konjunktionen, Präpositionen, Artikel. Im objektsprachlichen Index sind Wörter aus Verwendungsbeispielen nicht von Belegwörtern unterscheidbar. Im metasprachlichen Index sind "echte" Definitionswörter von formulierungstechnisch bedingten Appellativen nicht unterscheidbar. Da die Werkbezeichnungen in den Zitaten typgleich mit der Metasprache sind, ist ein Appellativum *Ahnen* nicht ohne weiteres als Buchtitel zu erkennen. Ein Teil dieser Probleme läßt sich jedoch ausgrenzen, wenn man, wie vorge-

schlagen, die Standardbelege vorab segmentiert oder generell abweichend vom Originaldruck weitere funktionelle Formatunterscheidungen trifft. Eine einfache Möglichkeit zur Straffung der Indizes besteht darin, die Massenwörter durch Stopplisten oder manuelle Nacharbeit zu eliminieren. Zur Veranschaulichung werden jeweils kurze Ausschnitte der genannten Indizes abgebildet.

Objektsprache

Artikel-lemma	Index-wort
<Aal>	aal
<Aal>	aale
<Aal>	aale
<aalglatt>	aalglatter
<Aar>	aar
<Aar>	aar
<Aar>	aar
<Aar>	aar
<Aar>	aare

<Aar>	aaren
<Aar>	aares
<Aar>	aars
<Aas>	aas
<Aas>	aas
<Aas>	aase
<aasig>	aasicht
<aasig>	aasiger
<aasig>	aasiges
<ab>	ab

Metasprache

Artikel-lemma	Index-wort
<Aal>	aalähnlicher
<Aalstecher>	Aale
<Aalfang>	Aale
<Aalreuse>	Aalfang
<Aar>	Aars
<aasig>	Aas
<aashaft>	Aase
<Aasrabe>	Aase
<Aasvogel>	Aase
<aasig>	Aase
<ab>	abgetan
<aasen>	abschaben
<ab>	Ahnen
<aashaft>	ähnlich
<Aar>	allmählich
<A>	Alphabet

<A>	Alphabets
<A>	Alphabets
<Aas>	alten
<ab>	alten
<aasig>	älteren
<Aar>	älterer
<Aal>	altes
<Aas>	Altes
<ab>	Altes
<A>	anders
<A>	Anfang
<A>	anfangen
<Aalstecher>	Anspießen
<A>	aufgekommen
<Aal>	Aufgußtierchen
<Aar>	aufkommend

Autorennamen

Artikel-lemma	Autor
<Aar>	Schlegel
<Aar>	Bürger
<ab>	Bürger
<ab>	Eichendorff
<ab>	Freytag
<Aar>	Gleim
<A>	Goethe
<Aal>	Goethe
<ab>	Goethe
<Aas>	Goethes
<ab>	Hebel
<ab>	Gotthelf
<Aas>	JPaul
<Aas>	Kant
<Aas>	Lessing
<ab>	Lessing
<Aas>	Lessing
<A>	Gerhard
<Aar>	Platen
<ab>	Rückert
<ab>	Scheuchzer
<ab>	Schiller
<ab>	Stieler
<A>	Wieland

Im Rahmen des zweiten Teils des Digitalisierungsexperiments an M. Heynes Deutschem Wörterbuch sind mit den vorgestellten Zugriffen die Möglichkeiten einer automatischen Textauszeichnung weitgehend erschöpft. Das Ergebnis besteht in einer relativ groben inhaltlichen Strukturierung, die eine Reihe von systematischen Zugriffen erleichtert und unterstützt, jedoch vom erreichten Standard nicht an den des Robert Électronique heranreicht. Eine Steigerung des Strukturierungsniveaus ist zwar ausgehend von den typographischen Kennzeichnungen noch mit maschineller Unterstützung möglich, erfordert aber zwangsläufig in wachsendem Maß wieder manuelle Nacharbeiten. Um Möglichkeiten solcher halbautomatischen Ansätze zu demonstrieren, werden in einem dritten Abschnitt des Experiments einige Beispiele vorgestellt.

Der Mangel der Wortformenabhängigkeit im bisherigen Aufbereitungsstatus wirkt sich vor allem im metasprachlichen Index nachteilig aus, da die Mischung der im engeren Sinn definitionssprachlichen gegenüber den im diskursiven Stil formulierungstechnisch geforderten Wörtern, die typographische Gleichheit der historischen und flexivischen Variablen sowie Heynes terminologische Varianten ein gezieltes Arbeiten stören. Um zum Beispiel auf die Phraseologiekennzeichnung *sprichwörtlich* zugreifen zu können, müßte man nach allen variablen Bezeichnungen für diese Kategorie suchen lassen. Wie vorne gezeigt, ergeben sich dabei allein im Abschnitt A – Ab sechs verschiedene Varianten. Das gleiche Problem stellt sich mit den Varianten der flektierten Wörter, insbesondere bei Homographen wie *Macht* f. und *macht* vb. usw. Vor allem im Bereich unregelmäßiger Verben kann die Anzahl der Varianten erhebliche Ausmaße annehmen, so daß sie kaum vom Benutzer kontrolliert werden kann. Diese Sachverhalte lassen sich nur durch manuelle Lemmatisierung und eine weitergehende Klassifikation beheben. Im Rahmen des Experiments wurden dazu einige Weiterbearbeitungen durchgeführt.

Die im engeren Verständnis terminologischen Bezeichnungen für grammatische, semantische oder pragmatische Klassen umfassen ein relativ überschaubares, häufig verwendetes Inventar, das durch die paradigmatische Sortierung im Index überschaubar wird. Varianten lassen sich leicht erkennen und per Lemmatisierung vereinheitlichen. Große Teile der formalen und sprachgeschichtlichen Wortbeschreibung in den Artikeln könnten durch eine Aufbereitung dieser definitionssprachlichen Einheiten gezielt angesteuert werden. Weniger optimistisch ist eine analoge Bearbeitung der bedeutungsbeschreibenden Wörter zu beurteilen. M. Heynes z. T. eigenwilliger, vielfach elliptischer Stil dürfte eine zuverlässige Hyponym/Hyperonym-Suche kaum zulassen, weshalb man sich auf eine bloß kategoriale Kennzeichnung beschränken wird.

Man kann jedoch die Möglichkeiten der Textkodierung zur Erweiterung des Strukturmodells benutzen. Das Element "Beschreibungssprache" wird um die Attribute "Textwortlemma", "Artikellemma" und "Paradigma" erweitert. Das Attribut "Artikellemma" sichert den Rückbezug der beschreibungs- und objektsprachlichen Textwörter zum zugehörigen Artikel. Das Attribut "Textwortlemma" ermöglicht dem Textretrieval eine einheitliche Zugriffsebene. Es muß nicht mehr nach "adv." und "adverb" gesucht werden, sondern nur noch nach dem Lemma, unter dem beide zusammengefaßt sind. Das Attribut "Paradigma" zeigt exemplarisch inhaltliche oder kategoriale Paradigmenzuordnungen, die bei Abfragen als je eigene Inhaltsstrukturen erschlossen werden können. Es ist ausdrücklich darauf hinzuweisen, daß die vorgenommenen Klassifikationen weitestgehend nur manuell und unter Autopsie des Artikeltextes möglich sind. Die anschließende Übersicht zeigt einen entsprechenden Ausschnitt in der alphabetischen Sortierung nach den Einträgen der Spalte "Textwortlemma". (B = Bedeutungsbeschreibung, WB = Wortbildung, WA = Wortart, ST = Sprachstufe, SO = Sprachsoziologie, KA = Kasus, RE = Regionalsprachliche Bindung, SR = bestimmter Sprachraum, NU = Numerus, PH = Phraseologismus, KG = Kompositionsgruppe).

Artikel-lemma	Textwort	Textwortlemma	Paradigma
<Aas>	Ableitung	Ableitung	WB
<ab>	Adv.	Adverb	WA
<ab>	Adverbien	Adverb	WA
<Aal>	ahd.	althochdeutsch	ST
<Aar>	ahd.	althochdeutsch	ST
<Aas>	ahd.	althochdeutsch	ST
<Aas>	ags.	altsächsisch	ST
<ab>	Artikel	Artikel	WA
<Aas>	biblisches	biblisch	SO
<ab>	Dativ	Dativ	KA
<ab>	Dialekten	Dialekt	RE
<ab>	dichterischer	dichterisch	SO

<aasen>	Fischern	Fischer	SO
<Aal>	gemeingerm.	gemeingermanisch	ST
<Aar>	Gen.	Genitiv	KA
<Aar>	Gen.	Genitiv	KA
<Aasseite>	Gerbern	Gerber	SO
<aasen>	Gerberwort	Gerberwort	SO
<Aas>	gewählter	gewählt	SO
<Aar>	goth.	gotisch	ST
<A>	griech.	griechisch	ST
<Aar>	griech.	griechisch	ST
<ab>	griech.	griechisch	ST
<ab>	indogerm.	indogermanisch	ST
<Aasjäger>	Jagenden	Jagender	SO
<aasen>	Jägern	Jäger	SO
<ab>	kaufmänn.	kaufmännisch	SO
<A>	kirchl.	kirchlich	SO
<A>	lat.	lateinisch	ST
<A>	lateinische	lateinisch	ST
<ab>	lat.	lateinisch	ST
<A>	Lernens	Lernen	B
<Aal>	mhd.	mittelhochdeutsch	ST
<Aar>	mhd.	mittelhochdeutsch	ST
<Aas>	mhd.	mittelhochdeutsch	ST
<Aaß>	mhd.	mittelhochdeutsch	ST
<ab>	mhd.	mittelhochdeutsch	ST
<Aaß>	Müller	Müller	SO
<ab>	Mundart	Mundart	RE
<ab>	mundartlich	mundartlich	RE
<ab>	Nomen	Nomen	WA
<ab>	oberd.	oberdeutsch	SR
<ab>	Orts	Ort	B
<Aal>	Plur.	Plural	NU
<Aar>	Plur.	Plural	NU
<Aas>	Plur.	Plural	NU
<ab>	Präp.	Präposition	WA
<Aal>	Redensarten	Redensart	PH
<ab>	sanskr.	Sanskrit	ST
<Aas>	Schimpfwort	Schimpfwort	SO
<Aar>	Schriftsprache	Schriftsprache	SO
<ab>	schwäb.	schwäbisch	SR
<ab>	Schwankens	Schwanken	B
<ab>	schweiz.	schweizerisch	SR
<Aas>	Sing.	Singular	NU
<A>	sprichwörtl.	Sprichwort	PH
<A>	sprichwörtl.	Sprichwort	PH

<Aal>	Sprichw.	Sprichwort	PH
<Ä>	Stockens	Stocken	B
<ab>	Subst.	Substantiv	WA
<ab>	trennbaren	trennbar	WB
<ab>	getrennt	trennen	B
<Aar>	urgerm.	urgermanisch	ST
<Aar>	urverwandt	urverwandt	ST
<ab>	Verb.	Verb	WA
<ab>	Verben	Verb	WA
<ab>	Vorkommens	Vorkommen	B
<Aasjäger>	weidmännische	weidmännisch	SO
<Aas>	westgerm.	westgermanisch	ST
<Aal>	Zusammensetzungen	Zusammensetzung	KG
<Aas>	Zusammensetzungen	Zusammensetzung	KG

Die Textwortlemmatisierung schafft, wie in verschiedenen Fällen erkennbar ist, die Voraussetzung für eine Zusammenfassung terminologischer oder flexivischer Varianten. Es bietet sich an, diese Arbeiten am Index durchzuführen, was eine erhebliche Reduzierung des zu bearbeitenden Wortmaterials mit sich bringt. Die 218 740 beschreibungssprachlichen Elemente reduzieren sich auf 23 620 Elemente im Index, was einer Reduzierung auf rund 10% der ursprünglichen Menge entspricht. Legt man den Wert eines Attributes eines Elementes im Index fest, so ist dies so, als setzte man diesen Wert für alle Elemente, die demselben Zeichenfolgentyp entsprechen. Dieses sehr ökonomische Verfahren ist allerdings nicht unproblematisch. Im Falle von Homographen ist die Arbeit am Index nur bedingt möglich, da nicht wirklich allen Elementen mit dem gleichen Zeichenfolgentyp der gleiche Wert zugeordnet werden darf. Für die 49 beschreibungssprachlichen Elemente *macht* gilt für das Attribut "Textwortlemma", daß ihnen entweder der Wert *machen* vb. oder der Wert *Macht* f. zugeordnet werden muß. Genauso verhält es sich mit Zeichenfolgentypen, die beim Attribut "Paradigma" verschiedene Werte annehmen können. So zeigt sich für das beschreibungssprachliche Element *Müller*, daß diesem entweder der Wert "Definitionswort" oder "sprachsoziologische Markierung" zukommen kann. Für solche Einzelfälle ist es notwendig, jedes Element für sich zu sichten und zu entscheiden.

Deutlich über dem Aufwand für die bisher vorgestellten Kategorialklassifikationen lägen der manuelle Klassifikationsaufwand und die datentechnischen Aufbereitungen für Zugriffe wie den folgenden, in dem zum Phraseologiekennzeichen das objektsprachliche Textelement erfaßt werden soll. Eine automatische Trennung nach Formatabfolgen erweist sich als sehr fehleranfällig.

Artikelstichwort	kategoriale Bezeichnung	objektsprachliches Text
Anfang	sprichwörtlich:	aller anfang ist schwer; guter anfang, halbe arbeit;
Anfangen	sprichwörtlich:	das karnickel hat angefangen,
Anfrage	sprichwörtlich:	eine anfrage ist keine anklage.
Angel	sprichwörtlich und bildlich:	zwischen thür und angel stecken,
Angler	sprichwörtlich:	ein angler musz wissen wann er ziehen soll.
Angreifen	sprichwörtlich:	wer pech angreift, besudelt sich;
Antwort	sprichwörtlich:	keine antwort ist auch eine antwort; es gehört nicht auf alle fragen antwort.
April	sprichwörtlich	april thut was er will;
arg	Rechtssprichwort	kinder folgen der ärgeren hand;
Arm	sprichwörtlich:	grosze herren haben lange arme;
Arm	sprichwörtlich:	er grüszt gern, wo unser herrgott einen arm herausstreckt
armen	im Sprichworte	almosengeben armet nicht;
Armut	in Sprichwörtern:	armut lehrt viel böses; ist ein unwerter gast; der künste mutter; ist keine sünde, schändet nicht

Das Verfahren nähert sich an dieser Stelle freilich wieder den Bedingungen während der Ausgangssituation des Experiments, in der sich der Aufwand für die Restrukturierung einer lexikographischen Überarbeitung nähert.

Noch nicht berücksichtigt wurde bisher die Möglichkeit einer Erschließung der Verbindung von Quellenverzeichnis und Quellenangabe. Diese Verknüpfung erlaubt eine vom Zitiersystem des Wörterbuchs unabhängige, vollständige Zitatnachweisung. Sie erlaubt ferner bei entsprechender Aufbereitung Auskunft über Quellenkorpusstrukturen z. B. im Zusammenhang mit der Einschätzung von Suchzielen. Im speziellen Fall des Heyneschen Wörterbuchs schafft sie auch die Voraussetzung für die Standardisierung der Kurznachweise bei den Zitaten sowie eine Nachdatierung der Belege.

Die vollständige Erfassung des Quellenverzeichnisses zu M. Heynes Wörterbuch erfolgte als eigenes Restrukturierungsmodul. Die Datenstruktur berücksichtigt Vornamen und Familiennamen der Verfasser, Heynes bibliographische Fassung des Publikationstitels und den von ihm angegebenen Kurztitel für die Nachweise bei den Belegen im Wörterbuch. Zusätzlich werden die von Heyne in der Gliederung des Quellenverzeichnisses angelegten Periodenzuordnungen bei jedem Titel notiert. Darüber hinaus enthält der Datenbestand für die Einzeltitel mit Ausnahme der Sammeleditionen entweder ein bibliographisch bzw. aus den vorgefundenen Angaben ermitteltes Entstehungs- oder Ersterscheinungsdatum. Soweit diese Datierungen nicht aus M. Heynes eigenen Angaben stammen, sind sie anhand bibliographischer Hilfsmittel

rekonstruiert worden. Die rekonstruierten Datierungen erscheinen in Klammern. In der Rubrik "Texttyp" wird zwischen Editionen (Ed.), Originalausgaben (Oa.) und Wörterbücher (Wb.) unterschieden. Da M. Heyne in den Quellenangaben für die Belege bei literarischen Werkausgaben oft die Einzeltitel angibt, müssen diese in einer Substruktur separat erfaßt und datiert werden. Ein nicht unbeträchtliches Problem für den Arbeitsaufwand ergibt sich auch aus der z. T. wenig konsequenten Form der angegebenen Kurztitel.

	Vorname	Name	Titelansatz Heyne	Kurztitel Heyne	Texttyp	Datierung	Periodenzuordnung
		O.Vf.	Beowulf, herausgegeben von M. Heyne. 7. Auflage, besorgt von Adolf Socin, Paderborn und Münster 1903.	Beowulf	<Ed>	[8./9. Jh.]	ahd., as., ae.
		O.Vf.	Ezzos Leich s. Müllenhoff-Scherer	Ezzo	<Ed>	[11.Jh.]	ahd., as., ae.
		O.Vf.	Heliand, herausgegeben von Moritz Heyne. 4. Auflage. Paderborn 1905.	Heliand	<Ed>	[<822/40>]	ahd., as., ae.
4	E.	Wildenbruch, von *	Wildenbruch, E. v.*, Die Quitzows, Schauspiel in 4 Acten, 1888.	Wildenbruch	<Oa.>	1888	nhd. heutige Zeit
	...						
5	E.	Wildenbruch, von *	Wildenbruch, E. v.*, Der Generaloberst, Trauerspiel im deutschen Vers, 3. Aufl., 1890.	Wildenbruch	<Oa.>	1890	nhd. heutige Zeit

Eine Redatierung der Belege im Artikelkontext könnte innerhalb des markierten Materials durch Zeichenfolgenaustausch vorgenommen werden.

Die Digitalisierungsversuche mit M. Heynes Deutschem Wörterbuch zeigen, daß die digitale Erfassung vorliegender historischer Wörterbücher mit einer Reihe von spezifischen wörterbuchtypischen Gegebenheiten zu rechnen hat. Bereits eine erste Stufe der Digitalisierung mit Datenerfassung, Textformatierung und erforderlichen Korrekturen erfordert aufgrund fehlender Automatisierbarkeit eine beträchtliche Investition. Ein solches Produkt wäre nur unter dem Blickwinkel der Literaturversorgung nach dem für seine Herstellung erforderlichen Aufwand zu teuer. Für systematische Nutzungen ist es in dieser Aufbereitungsstufe aufgrund der Unstrukturiertheit weitgehend ungeeignet. Um Zugriffe zu ermöglichen, die mit elaborierten digitalen Wörterbuchversionen möglich sind, müßten erhebliche metalexikographische Strukturierungsarbeiten geleistet werden. Aufgrund der diskursiven Textstruktur und verschiedener Stilmerkmale stößt eine solche Strukturierung des Heyneschen Wörterbuchs auch sachlich an Grenzen. Vor allem ist eine konsequente Unter-

scheidung aller Inhaltsebenen kaum möglich, ohne verändernd in den bestehenden Text einzugreifen. Aber auch die sachlich vertretbaren Strukturierungen erfordern einen Arbeitsaufwand, der sich dem für eine inhaltliche Überarbeitung nähert. Die rechnergestützt möglichen Strukturierungen des Wörterbuchtextes erlauben eine Unterscheidung derjenigen Wörterbuchbestandteile, die in der Makrostruktur bzw. in der Mikrostruktur durch typographische Merkmale bestimmt sind. Außer den Artikelgrenzen sind dies nach der Originaltypographie des bearbeiteten Werks vor allem die Ebenen der Stichwörter, der Verfassernamen in Belegen sowie die Ebenen der Objekt- und Metasprache. Die Standardbelege sind zudem aufgrund einer bestimmten Formatsyntax identifizierbar. Eine automatische Kennzeichnung dieser Schichten ist möglich, verlangt aufgrund der verschiedenen Fehlerquellen jedoch nachträgliche manuelle Prüfgänge. Dieses Ergebnis automatischer Textauszeichnung bleibt deutlich hinter den beobachteten elaborierten Standards zurück. Sie kann jedoch schichtspezifische Suchen wirksam unterstützen und auf diese Weise einen Einstieg in systematische Wörterbuchbenutzungen eröffnen. Gleichzeitig bietet sie den Vorteil einer alterungsbeständigen, transferierbaren und plattformneutralen Datenkodierung im TEI-Format. Weitergehende Formen der Restrukturierung und maschinenlesbaren Auszeichnung des Wörterbuchtextes sind nur noch in sehr beschränktem Maß automatisierbar. Hier wäre als eine Möglichkeit der Rechnerunterstützung die temporäre Indexbildung zur weiteren Subklassifikation zu nennen. Generell ist bei solchen weiterführenden Arbeiten mit erheblichem Arbeitsaufwand zu rechnen, der vor allem im Hinblick auf die wissenschaftsgeschichtlichen Implikationen der vorliegenden historischen Wörterbücher eher im Rahmen lexikographischer Überarbeitungen zu diskutieren wäre. Insgesamt zeigen die Experimente am Heyneschen Wörterbuch, daß angesichts der "Goldrauschstimmung", die gegenwärtig Teile der Diskussion um die Erstellung retrospektiv digitalisierter Wörterbücher bestimmt, vor einer Überschätzung der Möglichkeiten gewarnt werden muß. Es war deutlich zu zeigen, daß eine retrospektive Digitalisierung historischer Wörterbücher kaum ohne sprachwissenschaftliche und lexikographische Bearbeitung und in ausschließlicher Stützung auf informatische Kompetenz zu einem befriedigenden Ergebnis führen kann. Es wäre bedauerlich, wenn die sicher zukunftsträchtigen Perspektiven digitaler Wörterbuchbearbeitungen und Wörterbuchnutzung durch negative Erfahrungen mit quick-and-dirty-Produkten dauerhaft beeinträchtigt würden.

Literaturhinweise

Deutsches Wörterbuch. 1854-1971. Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, I-XVI. Leipzig: S. Hirzel.

Deutsches Universalwörterbuch. o. J. Deutsches Universalwörterbuch A-Z. 3. Aufl. Version 1.1, PC-Bibliothek. Mannheim/Wien/Zürich: Duden.

Guidelines. 1994. Guidelines for Electronic Text Encoding and Interchange. Hg. v. C. M. Sperberg-McQueen und L. Burnard. Chicago/Oxford, 1994. Vgl. bes. Kap. 12: Print Dictionaries.

Heyne, M. 1890-1895. *Deutsches Wörterbuch*, I-III. Leipzig: S. Hirzel.

Milan, C. 1998. *Elektronische Lexikographie am Beispiel der Wörterbücher romanischer Sprachen*. Erscheint im Internet unter der Adresse http://www.uni-bamberg de.ba4hi99/milan.htm

Paul, H. 1992. *Deutsches Wörterbuch*. 9. neubearbeitete Aufl. v. H. Henne und G. Objartel unter Mitarbeit v. H. Kämper-Jensen. Tübingen: Niemeyer [CD-Version].

Retrospektive Digitalisierung. 1998. Retrospektive Digitalisierung von Bibliotheksbeständen. Berichte der von der Deutschen Forschungsgemeinschaft einberufenen Facharbeitsgruppen "Inhalt" und "Technik", dbi-materialien 166. Berlin: dbi.

Le Robert Électronique. 1994. Le Robert Électronique Dos-Macintosh-Windows (CD-ROM). Paris: Robert.

Sanders, D. 1860-1865. *Wörterbuch der deutschen Sprache*, I-II. 2. Aufl. Leipzig: O. Wigand.

Weigand, F.L.K. 1909-1910. *Deutsches Wörterbuch*, I-II. 5. Aufl. Gießen: A. Töpelmann.

Nachbemerkung

K. Casemir und M. Schulz danken wir für kritische Durchsicht und Diskussion des Manuskripts, F. M. Wohlers und R. Bohne für die Korrektur und Übersetzung.

The Corpus of the Danish Dictionary

Ole Norling-Christensen and Jørg Asmussen,
The Society for Danish Language and Literature,
Copenhagen, Denmark

Abstract: A Danish corpus, holding 40 million words of general language from the period 1983-92, was designed and compiled by DSL (The Society for Danish Language and Literature) in order to serve as a major source for a new six volume dictionary of contemporary Danish. The corpus includes written and spoken, private and professional, general and specialised language, and each of the 44 000 text samples is annotated with formalized information on these and other features of linguistic and sociological importance. The resulting multidimensional text type specification is useful for the extraction of (virtual or real) subcorpora and for statistical analyses. Specialized software has been developed for flexible interactive concordancing and analysis. The corpus is currently only accessible at the site of DSL; nevertheless, several scholars and students have been using it in their research. The experience gained by the staff of DSL is being reused in cooperative language engineering projects within the European Union, and in 1998 a publicly available corpus will be released as an outcome of the PAROLE project.

Keywords: CONCORDANCE, COPYRIGHT, CORPUS, DANISH, DICTIONARY, FREQUENCY, LANGUAGE ENGINEERING, MUTUAL INFORMATION, SGML, STATISTICS, SUBCORPUS, T-SCORE, TEXT TYPOLOGY, WORD DISTRIBUTION

Opsomming: Die korpus van die Deense Woordeboek. 'n Deense korpus wat 40 miljoen woorde uit die algemene taal van die periode 1983-92 bevat, is ontwerp en saamgestel deur die DSL (The Society for Danish Language and Literature) om te dien as 'n primêre bron vir die saamstel van 'n nuwe ses-volume woordeboek van hedendaagse Deens. Die korpus sluit geskrewe en gesproke, private en amptelike, algemene en gespesialiseerde taal in, en elk van die 44 000 teksvoorbeelde word voorsien van formele inligting oor hierdie en ander kenmerke van taalkundige en sosiologiese belang. Die geskepte multidimensionele tekstipe spesifikasie is nuttig vir die onttrekking van (virtuele of ware) subkorpora en vir statistiese ontledings. Gespesialiseerde programmatuur is ontwikkel vir veeldoelige interaktiewe konkordansiebou en ontleding. Alhoewel die korpus tans slegs toeganklik is by DSL, het verskeie leerlinge en studente dit al gebruik in hulle navorsing. Die ervaring wat opgedoen is deur die personeel van DSL word hergebruik in koöperatiewe taalmanipulasieprojekte binne die Europese Unie, en in 1998 sal 'n korpus wat beskikbaar sal wees aan die publiek, vrygestel word as 'n uitvloeisel van die PAROLE-projek.

Sleutelwoorde: KONKORDANSIE, KOPIEREG, KORPUS, DEENS, WOORDEBOEK, FREKWENSIE, TAALMANIPULERING, WEDERSYDSE INLIGTING, SGML, STATISTIEK, SUBKORPUS, T-TELLING, TEKSTIPOLOGIE, WOORDVERSPREIDING

Lexikos 8 (AFRILEX-reeks/series 8: 1998): 223-242

1. The Danish Dictionary

The DDO Corpus was built during the period 1991-93 in order to serve as a primary source for *The Danish Dictionary* (Den Danske Ordbog, DDO), a new dictionary of contemporary Danish being edited by The Society for Danish Language and Literature (Det Danske Sprog- og Litteraturselskab, DSL). This Society, which is a kind of academy, was founded in 1911 with the aim of providing scholarly editions of Danish works of linguistic or literary importance, as well as dictionaries of the Danish language. Legally it is a semipublic institution under the jurisdiction of the Danish Ministry of Culture, and its activities are financed in part by the Danish Government and in part by the Carlsberg Foundation[2] and various other public and private foundations.

DSL edited the 28 volume *Ordbog over det Danske Sprog*, which was published 1918-56. It is the authoritative dictionary of newer Danish (i.e. from after c. 1700). DSL is currently in the process of editing five supplementary volumes which extend the coverage of all the volumes to 1955. A dictionary of Old Danish (1100-1510) is also in progress, and among the recent text editions of the Society is *Dansk Nationallitterært Arkiv* (Archive of Danish National Literature) on CD-ROM (1992). During 1995-98 DSL took part in the European Union language engineering oriented project *PAROLE* (MLAP63-386/LRE-63368), the aim of which is the production of comparable, harmonized corpora and lexica for the languages of the Union.

The history of the Danish Dictionary project dates back to 1989, when the plans for changing the European Community into an Economic and Monetary Union were launched. A large minority of the Danish people was, and still is, sceptical of the Union. Among other things it is feared that Danish culture and language will slowly but surely disappear in the new Europe. In order to allay this fear, several initiatives were taken by the Government, and a think-tank set up by the prime minister advocated the idea of creating a Danish national encyclopaedia and a dictionary of modern Danish. Both projects were launched in 1991 with the support of private foundations and the Government. The dictionary work was entrusted to DSL, which had submitted a plan and a budget for it by the end of 1989. The funding is shared equally by the Government and the Carlsberg Foundation. An electronic manuscript ready for printing will be delivered to the publishing house Gyldendal in the course of 2002, and the six volumes will be published in 2002–03. The royalties are earmarked for future lexicographical work.

The dictionary will contain approximately 100 000 entries and provide information on spelling, word-class, inflection, valency, pronunciation, meaning, phraseology and etymology. Entries are supplemented with original quotations, illustrating the different usages. It aims to fulfil the needs of both professional and general users of Danish, whether native speakers or advanced learners. The dictionary is basically descriptive, but the description includes information on acceptability, i.e. the norm. In other words: it shows the language as it is, not as it should be, but at the same time it also guides the user. There was

therefore no doubt in the minds of the chief editors, Ebba Hjorth, Kjeld Kristensen and Ole Norling-Christensen, that the work should be largely corpus-based. Foreign experience in the field was eagerly studied, especially the English dictionary project *Collins Cobuild*, the implications of which gave much inspiration to the first phase of the work, the building of a corpus. Thanks to the authors, papers like Atkins et al. (1992) and Church et al. (1991) were available to the editors in manuscript during this period.

Some domestic experience was also available, including the theoretical considerations of the makers of the first Danish corpus *DANwORD*, 1,25 million words for frequency studies of five distinct text types from the period 1970-74 (Maegaard and Ruus 1987). Thanks to funding from the Danish Research Council for the Humanities, a few more corpora had been created around the end of the 1980s: a collection of Danish, English and French texts in the field of contract law, and a collection of Danish, Spanish and German texts about genetic engineering, each holding c. one million words for each language. The latter was of special interest to the dictionary project, as some of the texts were not technical language (LSP), but written by or for laymen. Furthermore, Prof. Henning Bergenholtz of the Aarhus Business School collected one million words of general language (newspapers, magazines and novels) for each of the years 1987-90. This corpus, *DK87-90*, is the reference corpus most widely distributed among researchers of Danish.

2. Design of the corpus

It is important to underline that DDO is a *dictionary* project having a fixed budget of around six million ECU and a fixed time frame of twelve years. The corpus was thus not an end in itself, but was primarily established in preparation for the dictionary, even though some thought was also given to other future needs. Consequently, time and costs had to be among the premises for many of the decisions made during the planning and compiling of the corpus, including the decision of limiting the corpus period to ten years with some overrepresentation of the most recent three years.

2.1 Size and structure

The corpus consists of samples of written and spoken Danish produced during the decade 1983-92. The samples were collected, standardized and annotated by the staff of the Danish Dictionary, with the assistance of several students and external typists.

The following three aspects were taken into consideration during the initial design of the corpus: how many running words should be included, what period should be covered, and what types of text should be included.

In view of the Cobuild experience, it was decided that the corpus should consist of 40 million running words and should cover the Danish general lan-

guage as comprehensively as possible. Setting the number of running words to be included was not a main criterion, as this number naturally depends on other important considerations, such as the breadth, variety and balance of the coverage. Even though the dictionary is meant to describe contemporary Danish from the 1950s until today, texts from before 1983 were not included in the corpus. The decade from 1983 onwards was mainly selected because most machine-readable texts available are from this period, and it was estimated to be too costly and time-consuming to extend the coverage with scanned and/or typed text dating back to the 1950s. Furthermore, supplementary sources would be available to cover the language from 1955 up to the start of the corpus. They include just under one million slips with excerpts made by the Board of the Danish Language (Dansk Sprognævn) since 1955, two newly updated comprehensive bilingual dictionaries (Danish-English Dict. 1990) and (Danish-French Dict. 1991), and a special dictionary of New Words in Danish (Riber Petersen 1984). For the time after 1992 no systematic investigations are made. However, observations made by the staff, as well as slips submitted by the *spORDhunde*, a group of c. 300 voluntary "word watchers" who collect original material for the project, are continuously considered for inclusion.

The decade 1983-92 was designated as the Dictionary's primary period, meaning i.a. that the quotations used to supplement the dictionary definitions are chosen mainly from this period. Furthermore, it was accepted that the later part of the primary period would receive special emphasis because the supplementary sources would partially cover the earlier part of the decade. However, the corpus is balanced in this respect to allow for diachronic studies. As can be seen from figure 1, subcorpora of up to 16 million words, equally distributed over the years in question, may be selected from the main corpus by taking up to 1,6 million words from each of the years 1983-92[3].

The aim for the broadest possible coverage meant that the corpus was designed to comprise of general and specialized language, written and spoken language, "public" and "private" language (technically a distinction is made between *reception* and *production*), "young" and "adult" language, as well as a variety of different media, genres and subject areas. Two kinds of text were intentionally excluded viz. translated text, which will notoriously be biased by the source language, and technical language, i.e. language produced by specialists for other specialists in the same field, which is outside the scope of a dictionary of general language. In this context, *specialized language* (which is included) therefore means nonfictional written (or spoken) language for non-specialists, for instance textbooks or magazines on specific topics. Only a single intensional exception was made to the exclusion of translated text: parts of a new translation of the Bible were included. However, even though news-agency stories and subtitles of foreign films and telecasts were avoided, the origin of, for instance, newspaper stories cannot always be known. Finally, in order to cover as much different text as possible, entire novels, textbooks etc. were not included, but only one or a few randomly selected chapters up to a maxi-

mum of 10 000 words from each.

Year	No. of samples	Pct.	No. of words	Pct.
1983	2 199	5,0	1 601 379	4,0
1984	2 069	4,7	1 978 855	4,9
1985	2 291	5,2	2 295 799	5,7
1986	2 234	5,1	2 812 292	7,0
1987	1 809	4,1	3 639 409	9,1
1988	1 442	3,3	2 918 484	7,3
1989	1 821	4,2	2 798 556	7,0
1990	7 160	16,3	5 734 530	14,3
1991	14 155	32,3	8 688 920	21,7
1992	8 611	19,7	7 309 353	18,2
1993	15	0,0	329 765	0,8
Total	43 806	99,9	40 107 342	100,0

Figure 1: Number of text samples and running words by the year they were produced/published

As it is difficult to use objective criteria to establish what makes a balanced corpus, a more common-sense approach was adopted. Three dichotomies were selected (written vs. spoken, reception vs. production, general vs. specialized), and on the basis of these the corpus was divided into eight distinct classes. For each class, the possible text sources were reviewed and a preliminary word-number target was set. In some cases, this was done very informally, such as for spoken language, where the target was "as much as possible, up to a maximum of 10 million words". The collection of text samples was thus an iterative process: after a part of the corpus had been collected, statistical information was used to investigate which classes were still underrepresented and the selectional criteria were adjusted accordingly. The statistics were calculated on the basis of the information contained in the annotations (the headers, see below) of each text sample.

2.2 Selection of the text samples

The main sources for data acquisition were (a) books, magazines and news-

papers (28 million running words), (b) radio and television broadcasts (3,8 million running words), and (c) leaflets, booklets, pamphlets etc. (2 million running words). Furthermore, the relevant parts of existing Danish corpora were included, viz. the 4 million words of *DK87-90* and those parts of the corpus of genetic engineering which were not technical language. Several publishers, as well as Danmarks Radio (the National Broadcasting Company), were extremely helpful in supplying us with machine-readable text, the biggest donation being three volumes of three (very different) newspapers from the newspaper publisher Berlingske, a total of c. 75 million words distributed over more than 200 000 separate pieces of text.

It should be noted that only a relatively small part of this newspaper text was included in the corpus. However, the large number of separate articles etc. was most useful for the final balancing and annotating of the corpus, as the text had been downloaded from an information retrieval system which also contained some information on the individual articles. Even though this information was rather informal and inconsistent, parts of it could be transformed by a computational analysis into the standard categories for genre and topic, after which a balancing selection could be made. The information on authors (mostly journalists) was collected in a database which meant that information on year of birth, sex, etc., only had to be looked up once for each language user. Moreover, the database counted the number of newspaper articles by each author, which helped to avoid overrepresentation of the most productive journalists.

One of the explicit aims of the Danish Dictionary is to account for the use of spoken as well as written language. However, while the Dictionary aims to *cover* written Danish, it settles for only *considering* spoken Danish. The reason for this is twofold: it is theoretically difficult to define and represent spoken language usage in a corpus, and it is not economically feasible to collect and transcribe a large body of spoken language samples. Special emphasis was still put on the inclusion of spoken language, and the corpus does in fact contain 7 million words from private interviews, political debates, radio and television broadcasts etc., which represent 17 pct. of the total corpus. Again, great willingness to help was encountered: transcribed sociolinguistic material and interviews made for sociological research were given by colleagues at universities, and the unedited transcriptions of several animated debates with improvised contributions from a large number of members were received from the parliament and the city council of Copenhagen.

Another explicit aim of the Danish Dictionary is to describe the Danish language as it is used "privately" by the majority of the population *(production)*, instead of concentrating solely on "public" language users, such as journalists, authors, and politicians *(reception)*. Great emphasis was therefore placed on incorporating such material as private letters, letters to the editor, diaries, and school essays, which represent a total of 11 pct. of the corpus.

3. Building of the corpus

During the early period of the dictionary project (September 1991 – December 1993) the text samples were scanned, typed in, or, if already in a machine-readable format, converted from various kinds of wordprocessing or typesetting formats. Information on author(s), text type etc. was attached manually or, to some extent, automatically to the respective text samples.

SGML, the international standard for generic description of textual structures and marking up texts, was used for annotating the corpus. An SGML document type called *CorpusEntry* was defined. It provides the means for registering extralinguistic information about the text and for unambiguously tagging some (socio)linguistic features of it. Each of the 43 806 text samples of the corpus is one CorpusEntry element which consists of a header followed by the text proper. In the language of an SGML document type definition this is formally expressed as:

```
<!DOCTYPE    CorpusEntry [
<!ELEMENT    CorpusEntry (Header, Text)>
-- followed by declarations of the Header and Text elements -- ]>
```

3.1 Coding of the header

The header is structured by means of SGML tags as shown in figure 2. It is made up of a number of fields which have been filled in with formalized information (attribute/value pairs) about the respective text samples during the compilation of the corpus. The fields typically specify the authors' age, sex and language variant (standard or regional), as well as medium, genre and subject area (topic) of the text. Some of the fields are of special importance in that only a value from a finite set can be assigned to them; they are marked by bars (‖) in the figure. These fields are used for corpus statistics, and they permit the use of special "filters" for creating virtual or real subcorpora according to a multidimensional text type specification; these can in turn be accessed separately or compared statistically, thus making the concept of "a balanced corpus" more flexible.

TextInfo	
TextID	Unambiguous identifier of the text sample — for citation purposes
Restrictions	
Anonymity	Proper names must be altered (A), or not (-), if cited
DD_Only	Text must only be used by The Danish Dictionary
TextTitle	Title of the text
VolTitle	Name of anthology, newspaper, magazine etc.
Publisher	Publishing house, broadcaster etc.

	PublTime	
	Day	{1, 2, ..., 30, 31}
	Month	{1, 2, ..., 11, 12}
‖	**Year**	{1983, 1984, ..., 1992, 1993}
	Certainty	The year of publishing is known exactly (-), or not (?)
	Location	E.g. book volume, newspaper section, page number
‖	**LangType**	{general, specialized}
‖	**Expression**	{written, spoken, and two intermediate types}
‖	**Aspect**	{reception, production}
‖	**AgeRelation**	{adult–adult, adult–juvenile, adult–child, ..., child–child}
‖	**Medium**	{book, journal, radio, diary, ...} — 13 possible values
	Genre	{novel, interview, essay, ...} — 131 possible values
‖	**GenreType**	A reduced classification for statistical use — 17 values
‖	**Topic**	{philosophy, geography, physics, ...} — 66 possible values
‖	**TopicType**	A reduced classification for statistical use — 12 values
	Group	Unambiguous identifier of a group of related text samples
	Number	Serial number within the text group
	Size	Number of tokens in the following text sample
	UserInfo+	(one or more language users: author(s)/speaker(s))
	UserID	Identifier referred to by speaker turns in the text
	Surname	Surname of the language user
	FirstName	First name of the language user
‖	**Sex**	{male, female, unknown}
‖	**Born**	{1880, 1881, ..., 1989, 1990}
	Certainty	The year of birth is known exactly (-), or not (?)
	BirthPl	Place of birth
	Residence	Place of residence
‖	**Region**	Dialectal region — 11 values
	Education	Education of the language user
	Occupation	Occupation of the language user
‖	**LangVar**	Language variant {standard, regional}
	Role	Communicative role of the language user, e.g. teacher, pupil

Figure 2: Structure of the header information which accompanies each text sample

3.2 Coding of the text

An important consideration when designing a corpus is how the printed and spoken text should be represented computationally. As a matter of course one specific character set must be used. Because work is done in a PC environment (operating system: OS/2), Code Page 850[4] was chosen. However, this is only the first decision to be made. One uniform and consistent annotation system is also needed. This must be suitable for future computational searches and ana-

lyses, and information of importance for these uses must be recorded. On the other hand, it may not prove feasible to spend resources (human as well as computational) on recording information which is regarded to be of less or no importance. Defining such a format is no trivial task. It implies a series of decisions on *which features* of the text one wants to depict in the corpus. Should there, for instance, be specific codes for the smell of the paper? — Probably not. The colour of the paper? — It might have some special meaning. The size of the letters? — Differences in size are likely to signify differences in text type, but the meaning of such differences will differ from one text to another. An obvious conclusion from these kinds of question is that the coding has to be generic and not just mirror how the printer chose to represent the different kinds of text: *business pages*, not pink paper; *headline*, not big bold type; *highlighted*, not italics, bold or small caps.

For the Danish corpus a very restricted set of textual features has been chosen to be marked up. The structure of the element Text depends on whether it consists of written language or of (transcribed) spoken language. Written language is divided up into paragraphs (the element *p*) which in turn are mostly nontagged strings of characters (the SGML category #PCDATA); these may, however, be interspersed with elements of special categories of text, like highlighted text or notes.

For spoken language the first level of subdivision normally is not paragraphs, but speaker turns. Most of the spoken text samples are conversations or interviews with more persons involved. Consequently, the header may contain two or more instances of the element *UserInfo*. Each of these contains a different three letter string in the subelement *UserID*, and each element *speaker_turn* contains an attribute *id* which refers to the *UserID*. The *speaker_turn* element consists of #PCDATA interspersed with entity references[5] like {hesitation} representing nonverbal sounds like "eh", "mmm"; {pause}; {uf} representing a passage that was incomprehensible to the transcriber; {laughter}; and with the elements *comment* (the transcriber's "stage directions" that are not part of the speech), and *uncertain* (a word or passage that the transcriber was not sure about). The full set of SGML tags used is defined and explained in figure 3.

```
<!ELEMENT Tekst   ( ekst.tekst ¦ ill.tekst ¦ lyd ¦ p ¦ regi ¦ replik ¦ skrift ¦ tanke ¦
                    vers )+ +(kommentar)>
   -- The tag for the Text element --
<!ELEMENT ekst.tekst   ( ill.tekst ¦ p )+ >
   -- external text: part of text which was typographically placed in e.g. margins or
      boxes and thus not part of the running text --
<!ELEMENT f   ( #PCDATA ) >
   -- highlighted: enhanced part of the running text (e.g. italics, bold, or small caps in
      the printed original) --
<!ELEMENT ill.tekst   ( p ¦ vers )+ >
   -- caption: underline of illustration, table etc.; text inside an illustration --
<!ELEMENT kommentar   ( ( #PCDATA ¦ usikker)+ ¦ ( p ¦ vers )+ ) -(kommentar) >
```

```
    -- comment: transcriber's or editor's comment; not part of the text --
<!ELEMENT lyd   ( p+ ) >
    -- sound: in comic strips etc: rendering of sounds like Riiinnggg, Bam! --
<!ELEMENT note   ( #PCDATA | f )+ >
    -- foot- and end-notes; this element which contains the text of the note, is placed at
       the point of the text where the note-reference of the original was written --
<!ELEMENT p   ( #PCDATA | f | note | usikker )+ >
    -- paragraph: One or more empty lines between paragraphs in the original are
       represented by one instance of the newline-entity {NL} --
<!ELEMENT regi   ( p | skrift )+ >
    -- stage direction, in comics: the narrative text, like "Copenhagen, Townhall Square.
       An evening in September. It is six o'clock" --
<!ELEMENT replik   ( p | vers )+ >
    -- speaker turn in spoken language; speech balloon in comics --
<!ELEMENT skrift   ( p+ ) >
    -- writing, in comics etc.: text in the picture which is neither sound, stage direction,
       speaker turn, nor thought, but for e.g. posters, notes, letters, documents, graf-
       fiti --
<!ELEMENT tanke   ( p+ ) >
    -- thought, in comics etc.: the thoughts of the characters, rendered in thought
       balloons --
<!ELEMENT usikker   ( #PCDATA ) >
    -- uncertain: part of spoken language which was not identified with certainty by the
       transcriber --
<!ELEMENT vers   ( p+ ) >
    -- verse: metrical text. Each stanza under this element is marked up as a paragraph;
       its verses are separated by slashes --
```

Figure 3: Structure of the text element as defined in the DTD

4. The lexical database

In parallel to the corpus building, methods for reuse of existing lexical sources were developed, and a database of 340 000 words (i.e. lemmas) was extracted/constructed from the machine-readable versions of some standard printed dictionaries, *viz.* the official spelling dictionary (Retskrivningsordbogen 1986), Danish-English Dict. (1990), Danish-French Dict. (1991), supplemented with word-lists from the Board of the Danish Language. The database holds formalized morphological information, as well as unformalized (except for subject field) semantic and contextual information extracted from the source dictionaries. Using the inflectional information given in the database, all possible inflected forms of the lemmas were generated and compared to the stock of word forms that are present in the corpus. The remaining forms, which were not identified during this run, have been further investigated and gradually added to the database as new words or as unofficial spelling variants of exist-

ing ones. The selection of lemmas for the Danish Dictionary, as well as a tentative assignment of dictionary entry size, was made by the help of the information kept in the lexical database, including the word frequencies found in the corpus and the relative size of entries in the printed dictionaries.

5. Using header information for making a subcorpus

As a simple example of the use of the feature/value pairs of the headers for the design and extraction of subcorpora, as well as for the evaluation and further balancing of the resulting subcorpus, brief consideration will be given to the case of a Danish research institute, active in the field of machine translation, which needed a specialized corpus of text covering a range of 10 different, but somehow related, subject fields. Summing up the numbers of text samples and running words of the entire corpus for the specified values of the text type feature *topic* rendered the result shown in figure 4, which is at the same time the composition of the largest possible subcorpus that will fit the demand.

Topic code and name		No. of samples	Pct.	No. of words	Pct.
191	science (general)	76	1,6	115 407	3,4
195	communication	178	3,7	108 491	3,2
30	society	1 122	23,1	925 329	27,4
331	business	1 428	29,4	815 653	24,1
38	social services	635	13,1	453 074	13,4
50	natural sciences (in general)	158	3,3	127 143	3,8
60	technology	144	3,0	147 253	4,4
627	transportation	731	15,1	389 582	11,5
66	industry	147	3,0	68 052	2,0
68	crafts	234	4,8	230 535	6,8
Total		**4 853**	100	**3 380 519**	100

Figure 4: The number of samples and running words (tokens) for 10 of the 66 topics (subject fields) which are distinguished in the corpus

The composition of this subcorpus according to other text type features can now be investigated and compared to that of the general (reference) corpus and

be used for further balancing, if needed. Figure 5 is just one short example to serve as demonstration.

	Entire corpus		Selected subcorpus	
Authors' sex	No. of words	Pct.	No. of words	Pct.
unknown	14 539 129	36,3	1 164 447	34,4
female	7 648 688	19,1	532 964	15,8
male	17 919 525	44,7	1 683 108	49,8

Figure 5: Composition of the subcorpus summed up by the text type feature *sex of the author*, compared to that of the entire corpus

It can be seen from the table that an author has been identified for a greater proportion of the selected text than for the source corpus, and that the over-representation of male authors is even more marked. If another balance is desired, the user must discard some text samples, thus making a smaller, but more balanced subcorpus.

6. Exploitation of the corpus

There are two versions of the corpus, a master copy, which is a collection of SGML-coded text files, and a compiled and indexed version which is available on-line; the latter version is used every day by the editorial staff for making concordances and statistical analyses as part of the work on the dictionary. The master copy is used for special examinations which cannot be made by the interactive tool. It is continually refined, and at intervals a new compiled version is made from it. The refinement of the corpus includes correction of (technical) errors, the disambiguation of certain characters, and the making of some additional annotations. Among the errors that were corrected, were multiple instances of the same text sample, wrong dating (the machine-readable version supplied by a publisher proved to be a later version than the known printed book) and a few data conversion errors.

As to disambiguation, a clear-cut definition of which characters are part of a word and which are not, is necessary for simple and efficient computational processing of text. Apostrophes may be part of (contracted) words, but they are also frequently used as quotation marks; a hyphen is part of a word, whereas a dash is a punctuation mark, but quite often the same character is used for both. These ambiguities were resolved automatically with a high degree of certainty.

6.1 Two problems: abundance and scarcity

The lexicographer working with corpora runs into two basic problems: the theoretical problem of the significance of infrequent or missing occurrences of

some linguistic phenomena, and the practical problem of being flooded with too many instances of others. The former problem can only be solved by making the corpus even larger, or by relying on sources external to the corpus. To cope with the latter, computational tools are needed in order to structure the flood; without such tools, large corpora will not be of much use.

Finding the sense in a large corpus can be seen as the repetitive process of making ever more specific queries. The first basic query is that all the instances of a certain lemma be given. The following queries include contextual restrictions which can be made more precisely the more annotated the corpus becomes. The querying is repeated until some characteristic behaviour of the lemma crystallizes. Once such behaviour (e.g. one meaning, one valency frame) has been recognized and described by the lexicographer, the instances of it may be discarded and the procedure repeated for the remaining instances.

There is, however, one class of important questions that cannot meaningfully be answered solely on the basis of the immediate context of the instances of a lemma. Computational exploration of the collocational behaviour of a word is not possible without some knowledge of the corpus as a whole. The mere observation that one word seems to be occurring frequently in the neighbourhood of another word does not in itself indicate an affinity between the two, neither does a seemingly infrequent occurrence indicate the absence of such an affinity. Only a statistical calculation that takes into account the total number of occurrences of the words in question can give a reliable indication. A useful survey of methods and tools for identifying collocations in corpora is given in Fontenelle et al. (1994).

Since the work of Church et al. (1991) three statistical methods for collocational studies have become more or less standard. These or similar methods should be part of any toolbox for the analysis of large corpora. *Mutual information* (or the cognate *Z-score statistics*) reveals positional interdependence between two words by comparing the observed frequency of a co-occurrence to the calculated frequency for co-occurrence by chance. *Scale statistics* calculates the mean and the standard deviation of the distance between such pairs, thus giving a measure of the fixedness of the collocation in question. The more sophisticated *T-score test* looks for significant differences between the immediate neighbourhoods of two different words, typically pairs of near synonyms like "strong"/"powerful" or "his"/her". The observed neighbouring words, e.g. words in the position immediately to the right of the two, are ranged on a scale spanning from those having greatest affinity to one of the synonyms, through those which are neutral, to those with greatest affinity to the other synonym.

6.2 An interactive corpus tool

For corpus search and interactive analysis, a tool called Corpus·Bench was developed by the Danish software house TEXTware A/S according to specifications made jointly by Longman Publishers (UK) and the Danish Dictionary. It is

commercially available and is also being used by a few other publishing houses and academic institutions.

Concordances can be built in real time according to complex search criteria. The concordance lines can be interactively tagged according to several user-defined criteria, and they can be sorted by almost any combination of criteria. Moreover, the statistically-based methods for collocational analysis mentioned above are available, and frequency information, including frequency distribution over e.g. text types, can be obtained.

For the use by Corpus·Bench, the corpus must be compiled and indexed by a separate software package called Corpus·Build. It allows the user to design the overall structure of the corpus database, such as the definition of the alphabet, character mappings and separators. It also provides a tool for building and maintaining an optional inflectional dictionary that can be accessed by the retrieval system and facilitate searching for lemmas rather than individual word forms. Corpus·Build can handle the indexing of large SGML-annotated corpora (at least 100 million words). The annotations may reflect any kind of information on the text document, e.g. headers, morphosyntactic tags etc.

6.3 Working with Corpus·Bench

Almost any search criterion can be used to create concordances from the corpus. As Danish has a more complex inflectional structure than e.g. English, a concordance normally should be based on a lemma rather than a single word form. An inflectional dictionary, based on the above-mentioned lexical database, was therefore added to the retrieval system.

One can scroll through a concordance listing, view the contents of header fields together with the corresponding lines in the concordance, jump into the corresponding document by clicking the mouse on a concordance line, mark up lines with one's own annotations, and sort the lines according to any combination of keyword, left context, right context, user-defined tags, and text type information. Concordances or parts of them can be printed out or copied either to a file or to the Windows-OS/2 clipboard in order to paste them into another document, such as a dictionary entry in the dictionary compilation system.

Search criteria based on keywords can be combined with two types of filters: word filters and/or text type filters. Word filters specify the absence or presence of additional words or lemmas in a given contextual position or range. Text type filters specify the contents of certain header fields (cf. 3.1 above). Any logical combination of up to eight word filters and text filters can be applied to a query, which allows the user to specify queries such as

> "display a concordance listing with the keyword 'typisk' *typical(ly)* AND the word 'dansk' *Danish* OR 'engelsk' *English* OR 'fransk' *French* OR 'tysk'

German in context position +1 in text by persons born outside Denmark (Region=X)".

What came out of this example query were two statements: *Hiding one's light under a bushel may be a typical Danish expression, but doing so is not a salient Danish feature* and *Something being typically French implies that the opposite, too, is typically French.* Both authors happen to be born in the former Soviet Union.

Filters can be defined for all types of queries, including word-lists and statistics.

Word-lists show words according to specified search patterns (which will normally contain wild cards). As compound words are very common in Danish, a word-list can be used to investigate the productivity of a given word. For example, Corpus-Bench can list all words with the string "engelsk" in them (search pattern: **engelsk**), and the resulting list can be sorted alphabetically or, like here, by frequency:

Word	**Abs. frequency**
engelsk *(English)*	3 081
engelske *(English)*	2 630
engelsksprogede *(English-language [adj.])*	53
engelsktalende *(English-speaking)*	38
engelsksproget *(English-language [adj.])*	35
engelsklærer *(English teacher)*	24
engelskundervisning *(teaching of English)*	22
engelskundervisningen *(the teaching of English)*	10
engelsktime *(English lesson)*	10
engelskfødte *(English born)*	9
engelskkundskaber *(knowledge of English)*	8
engelskgræs *(thrift [a plant])*	7
dansk-engelsk *(Danish-English)*	7
engelsklæreren *(the teacher of English)*	6
oldengelske *(Old English)*	5
engelsk-amerikanske *(Anglo-American)*	5

Frequency lists give the absolute and relative frequency of the word forms belonging to a given lemma. By defining filters, one can investigate the use of a given word in different subcorpora. It is also possible to compare the frequencies of words that are related to each other. For the word "virus", two genders and several inflectional variants are permitted; the frequency list, giving the number of instances and the number per million running words, shows that inflection is normally avoided, and that far from all of the inflected forms are used:

	Abs.no.	Per mil.		Abs.no.	Per mil.		Abs.no.	Per mil.
virus	813	20,35	virussets	1	0,03	viruserne	0	0,00
virus's	0	0,00	virusets	4	0,10	virussene	0	0,00
viruses	0	0,00	virusser	5	0,13	virusserne	0	0,00
virussen	11	0,28	viruser	6	0,15	viraene	1	0,03
virusen	20	0,50	vira	43	1,08	virusenes	0	0,00
virusset	5	0,13	virussers	0	0,00	virusernes	0	0,00
viruset	25	0,63	virusers	0	0,00	virussenes	0	0,00
virussens	1	0,03	viras	3	0,08	virussernes	0	0,00
virusens	4	0,10	virusene	0	0,00	viraenes	0	0,00
						Total	942	23,58

A **word distribution report** shows the use of words which are distributed according to the contents of a header element, e.g. the year of birth, subject area, or time of publication. The verb "start", borrowed from English, was originally only used in connection with motors, cars and the like. However, it is gradually also taking over the more general meaning of "begynde" (*begin*); this is mirrored by the word distribution report by age:

Birth	Abs.no.	Total	Per mil.	Dev.pct.
?	6 808	18 921 566	359,80	+19%
1910s	146	1 129 424	129,27	–57%
1920s	431	2 322 307	185,59	–39%
1930s	734	3 842 267	191,03	–37%
1940s	1 735	6 570 818	264,05	–13%
1950s	1 612	5 196 341	310,22	+2%
1960s	643	1 916 274	335,55	+11%
1970s	20	48 836	409,53	+35%
Total	12 129	39 947 833	303,62	

A **mutual information report** displays a list of words that occur with a significantly high probability together with the keyword in a given contextual position or range. The report thereby identifies typical collocations. Most of the following left-side collocators of "interesse" (*interest*) represent expressions of the type *in the interest of ....* The factor *mut.inf.* measures how many times more frequent than chance the co-occurrence is[6], and *coocc.* is the actual frequency of each co-occurrence:

	mut.inf.	coocc.
nyhedens (*of novelty*)	6 582,28	[15]
sandhedens (*of truth*)	1 513,92	[23]
alles (*of all, common*)	730,34	[34]
medlemmernes (*the members'*)	677,59	[10]
offentlighedens (*of the public*)	639,94	[10]
almen (*common*)	440,22	[15]
fornyet (*renewed*)	288,62	[14]
stigende (*growing*)	270,42	[78]

befolkningens *(of the population)*	197,33	[10]
størst *(greatest)*	187,52	[21]
manglende *(missing)*	181,31	[40]
voksende *(growing)*	167,07	[19]
speciel *(special)*	147,33	[20]
øget *(additional)*	111,03	[22]
stor *(great)*	107,82	[280]
betydelig *(considerable)*	98,78	[14]
offentlig *(public)*	96,42	[14]
historisk *(historical)*	95,49	[14]
samfundets *(of society)*	93,75	[10]
fælles *(common)*	84,83	[59]
særlig *(special)*	82,82	[61]

Finally, **T-score reports** are used for investigating differences in the use of words that are related to each other in some aspect. A T-score report can be thought of as two mutual information reports compared to each other. The report given below shows what is — to a Dane — typically German but at the same time untypically French and vice versa. While T-score reports normally do not show unexpected results when based on adjectives of nationality, they are very useful in lexicography for the investigation of slight differences in the use of almost synonymous adjectives, e.g. "strong" *vs.* "powerful" or "big" *vs.* "large".

German	T-score	French	T-score
genforening *(reunification)*	9,78	revolution *(revolution)*	–10,55
2 *(television channel)*	7,14	præsident *(president)*	–6,95
besættelse *(occupation)*	6,14	franc *(= the currency)*	–6,80
soldater *(soldiers)*	5,87	skole *(school)*	–5,70
forbundsbank *(federal bank)*	5,72	francs *(= the currency)*	–5,59
rente *((rate of) interest)*	5,27	kartofler (franske kartofler) *(potatoes (potato crisps))*	–4,50
1 *(television channel)*	5,03	koloni *(colony)*	–4,23
stater *(lands)*	4,98	visit *(f.v. = flying visit)*	–4,03
soldat *(soldier)*	4,82	konge *(king)*	–3,78
forbundskansler *(Federal Chancellor)*	4,67	polynesien *(Polynesia)*	–3,64
enhed *(unity)*	4,56	køkken *(kitchen, cooking)*	–3,62
fransk *(French)*	4,43	kunst *(art)*	–3,62
værnemagt *(Wehrmacht)*	4,10	off. *(= official (language))*	–3,48
rige *(State, Reich)*	4,10	ord *(word)*	–3,42
kansler *(Chancellor)*	4,10	revolutions *(revolution's)*	–3,36
besættelsesmagt *(occupying power)*	4,10	alper *(Alps)*	–3,36

What makes Corpus·Bench different compared to most other commonly-used corpus retrieval systems is its capability of handling extra-textual information. Queries are not limited to the raw text of the corpus, but may be modified by the information supplied in the headers, as well as by part-of-speech tags, if available.

7. Third parties' use of the corpus

The linguistic resources developed for the dictionary project, the corpus, as well as the lexical database have already been widely used by researchers and students of Danish. Among the topics for corpus based term papers and theses written by university students are "The concept *sand* (true)", "Topology and interpretations of the adverb *kun* (only, just)", and "Topology of some adverbials in spoken language". For a term paper on automatical identification of technical terms in professional text, a lemmatized list of frequent words in general language was produced. PhD theses and studies by senior researchers include work on prototypical sensory and speech act verbs; onomatopoeic words in written and spoken Danish; valency patterns of adjectives; the concept *politician*; stylistics; lexical semantics; and some derivational affixes. A corpus-based study of types of language errors was made as part of preparatory work on a syntax checker for Danish.

7.1 Criteria for access

The access to the corpus for external users is regulated by three kinds of considerations: copyright, resources available, and a wish for survival.

During the compilation of the corpus no formal copyright agreements were made; and it would in fact have been a major job to find the authors of 44 000 distinct pieces of text and get their permission. The publishers and others who supplied text, were promised that it would only be used for dictionary work and other research; furthermore, as far is known most of them did not ask permission from the actual copyright holders, namely the authors. Consequently, the corpus had to be handled like photocopies: it is permissible to make one copy for personal use, but illegal to duplicate and distribute copies. External users, therefore, normally do not receive (sub)corpora, but rather concordances or word-lists, or they are invited to query the corpus on the premises of the Dictionary, where a special subcorpus for guests is available. The "guest corpus" excludes a few million words on which special restrictions were laid by the suppliers. However, making concordances or word-lists, and instructing guests in the use of the corpus tools, encroaches upon time for working on the dictionary, and given the sparse resources available, help has to be somewhat limited. On the other hand: widespread use of the corpus for many different purposes would prove the need for its continuation after the end of the dictionary project. That is where the wish for survival comes in, and that is one reason why every instance of external use is carefully recorded. No charges have been made so far, partly because quite a few of the users of the corpus — or their institutions — were in fact also providers of textual material to the corpus.

8. The PAROLE project

A new dimension, and a new approach to the question of availability, was added to DSL's corpus work when the Society became a partner of the PAROLE project in 1994, the aim of which was to provide publicly accessible harmonized comparable corpora and lexica (i.e. dictionaries which can be accessed and used by computer programs) for the official languages of the European Union and for Catalan and Irish — a total of 14 languages. The corpora focus on written language, and their primary target group is the language industry. Consequently, the design criteria are not the same as for the corpus of the Danish Dictionary; among other things, childrens' language and other nonstandard variants have been left out. Three kinds of corpora should be made, viz. a 20 million word publicly accessible corpus, a 3 million word distributable corpus, and a 250 000 word morphosyntactically tagged, and manually checked, corpus. Producing the tagged corpus was by far the most labour-intensive part of the job, as no experience in this field, let alone an automatic tagger, was available for Danish. The next step will now be to use the 250 000 words for training some taggers which are known to have been successfully used with other languages.

9. Future development

As already mentioned, the immediate goal of the project is a manuscript for a six volume dictionary of contemporary Danish, which will be completed by 2002-03. Further objectives for the future of the corpus include a strengthening of the diachronic dimension of the corpus, as well as the integration of computational methods in the philological editorial work of the Society. Techniques used for corpus building and analysis may also prove useful for the preparation of scholarly text editions, as well as for the use of such editions, which are likely to be published electronically in the not too distant future. As to the future of the dictionary, an electronic version is likely to be the next step. It will be accessible not only by headwords (semasiologically) but also by concepts (onomasiologically). Preparations for such access are part of the ongoing work. Furthermore, it may provide far more examples of real language than the printed version.

Notes

1. The authors want to thank their colleagues at *The Danish Dictionary*, Henrik Andersson and Ebba Hjorth, for input to and comments on the manuscript.

2. The Carlsberg Foundation, the owner of most Danish and several foreign breweries, is among the most important sponsors of Danish science and scholarship.

3. The 15 texts from 1993 were a series of transcribed interviews planned for 1992. They happened to be delayed for a couple of months. It was, nevertheless, decided to include them.

4. The IBM specific character set "Code Page 850 (Latin 1)" holds an inventory of Western European letters which is close to that of the ISO character set 8859-1. Conversion between the two character sets is not a major problem.
5. In order to make the text more readable to humans braces {...} have been chosen for the delimiting of SGML-entity references, instead of the standard SGML-delimiters &...; (ampersand ... semicolon) which can therefore be used with their original meaning. The braces are reserved characters that are not used for other purposes. A newline-entity {NL} is inserted as section delimiter, i.e. in places where the original text had one or more empty lines between paragraphs.
6. For instance, the word "stor" *(big, great)* appears 107 times more frequently to the left of "interesse" than would be expected if the words were randomly distributed. Strictly speaking, in information theory *mutual information* is defined as the logarithm to the base 2 of the figures which are here called *mut.inf.*

References

Atkins, Sue, Jeremy Clear and Nicholas Ostler. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7 (1): 1-16.

Church, Kenneth, William Gale, Patrick Hanks and Donald Hindle. 1991. Using Statistics in Lexical Analysis. Zernik, Uri (Ed.). *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon.* New Jersey: Erlbaum.

Danish-English Dict. 1990. Vinterberg, Hermann and C.A. Bodelsen: *Dansk-engelsk Ordbog*. Third edition edited by Viggo Hjørnager Pedersen. Copenhagen: Gyldendal.

Danish-French Dict. 1-2. 1991. Blinkenberg, Andreas and Poul Høybye. *Dictionnaire Danois-Français/Dansk-fransk Ordbog*. Fourth edition edited by Jens Rasmussen et al. Copenhagen: Arnold Busck.

Fontenelle, Th., W. Bruls, L. Thomas, T. Vanallemeersch and J. Jansen. 1994. *Designing and Evaluating Extraction Tools for Collocations in Dictionaries and Corpora.* Document D-1a of DECIDE (MLAP-Project 93-19), Luxembourg.

Maegaard, Bente and Hanne Ruus. 1987. The Composition and Use of a Text Corpus. Cappelli, A., L. Cignoni and C. Peters (Eds.). 1987. *Studies in Honour of Roberto Busa S.J.* Linguistica Computazionale IV-V: 103–21. Pisa: Giardini Editori.

Retskrivningsordbogen. 1986. Copenhagen: Dansk Sprognævn/Gyldendal.

Riber Petersen, Pia. 1984. *Nye Ord i Dansk 1955-75.* With contributions by Jørgen Eriksen. Copenhagen. Dansk Sprognævns skrifter 11. Copenhagen: Gyldendal.

PEDANT: Parallel Texts in Göteborg

Daniel Ridings, *Språkbanken,*
Institutionen för svenska språket, Göteborgs universitet,
S-412 98, Göteborg, Sweden

Abstract: The article presents the status of the PEDANT project with parallel corpora at the Language Bank at Göteborg University. The solutions for access to the corpus data are presented. Access is provided by way of the internet and standard applications and SGML-aware programming tools. The SGML format for encoding translation pairs is outlined together. The methods allow working with everything from plain text to texts densely encoded with linguistic information.

Keywords: SGML, PARALLEL CORPORA, MORPHOSYNTACTIC ENCODING, LEMMATIZATION, MULTIWORD UNITS, COMPOUND WORDS, INTERNET ACCESS

Opsomming: In hierdie artikel word 'n beskrywing gegee van die stand van die PEDANT-projek met parallelle korpora by die Taalbank by die Universiteit van Göteborg. Oplossings vir die verkryging van toegang tot die korpusdata word aangedui. Toegang word verskaf deur middel van die Internet en standaardtoepassings en SGML-sensitiewe programmeringshulpmiddels. Die SGML-formaat vir die enkodering van vertaalpare word gesamentlik geskets. Hierdie metodes laat toe dat gewerk kan word met enigiets vanaf suiwer teks tot tekste wat taalkundig dig geëtiketteer is.

Sleutelwoorde: SGML, PARALLELLE KORPORA, MORFOSINTAKTIESE ENKODERING, LEMMATISERING, MEERWOORDIGE EENHEDE, SAAMGESTELDE WOORDE, INTERNETTOEGANG

Introduction

This is a second progress report dealing with the work being done in Göteborg on parallel texts. It will cast light on new developments and adjust or correct some of the plans expressed in the first report (Danielsson and Ridings 1996b).

Goals

The first of many goals is nearly achieved, namely the creation of a text collection substantial enough to provide raw material for further research. The original ambition to include as many languages as possible has been reduced to concentrating on Swedish, English, German, French and Italian. This language combinations play an important role in the university's translator training programme and in the research interests of graduate students who participate actively in the expansion of the text collection.

A second goal that was deemed to be within realistic reach was to create a foundation for multilingual lexical tools for the translator training program that began at the Faculty of Humanities in January 1997.

These two priorities together with their implications have been the guiding principles to which competing considerations such as target languages, genres, public access, and any attempt at balancing have been subordinated.

Language coverage

So far Swedish has been considered the common denominator for all possible language pairs. This does not necessarily imply that Swedish is always the source language (L1), merely that a text is considered interesting when there is a Swedish version. It is fairly easy to build a collection of texts in German, French and English since those languages are nearly always represented in various documents from the European Union even when the document in question is not available in all of the eleven official languages. Such documents, that is documents without a Swedish counterpart, are not found in the collection.

Despite the fact that Swedish is the pivotal language, this does not exclude the possibility of working with combinations such as French-German, French-Italian, English-French, etc. If a text in Swedish has been aligned with versions in English, French, German, etc., then it is just as possible to align English with French or German for that same text. There is in fact a sizeable amount of material that would permit such investigations. To provide definite numbers to describe the size of a parallel text collection is prone to create misunderstandings. Do the numbers represent the number of words in the whole collection in all the languages or is the collection to be measured by the number of words in only one language for which there are counterparts in another language? In the latter case, is one to count only the words in Swedish texts for which there are always counterparts in one language, or in several languages? Since Swedish is the pivotal language, there will always be a Swedish version of any given text, but not necessarily a German or a French version. After this cautionary note, it can be safely said that there are well over half a million words in Swedish versions for which there exist counterparts in all of the following languages: English, French, German, Italian and Spanish. This means that investigations can be performed on any combination of these languages on a substantial amount of material. If the size is related to language pairs, then there are about a million words of Swedish-English, Swedish-French and Swedish-Italian collections, that is, a million words in Swedish aligned with a million words in English, etc.

The format of PEDANT

The administration of a collection of parallel texts involves all the anticipated

problems and pitfalls associated with monolingual corpora and more. Mistakes were made early on involving the construction of the textual database that need to be rectified. Such mistakes can have far-reaching consequences for the whole system.

PEDANT is represented in two basic formats: a traditional corpus format and a textual database. Both formats have their strengths and weaknesses.

The textual database simplifies the process of providing access to an end user since it is a fairly straightforward process to create a usable graphical interface. The original database manager was Microsoft Access and is still being used. The disadvantage is that this database manager does not lend itself to making material available on the network. This creates a situation where the material is only accessible on one machine and all users must go through the owner of the machine to make queries or printouts. This is often desirable, since not all material is intended to be publicly available, but it is less than satisfactory for the owner of that machine.

The corpus format permits one to use the various tools that have already been developed for monolingual text-processing. One has access to all the standard Unix programs and filters, part-of-speech taggers, lemmatizers, etc.

Both methods of working are manageable as long as only one language pair is involved. The alignment of one unit in one language with another in any given context will always be the same. Once it has been established, it is stable. The parallel text collection in Göteborg, as was mentioned above, contains a substantial amount of material in six languages and the aim is to keep the collection as flexible as possible in order to allow for a diversity of language pairs. Even if the primary concern is with Swedish-French, the possibility of querying German-French for the same text should not be ruled out. This is where the problems arise and where the mistakes were made when the first textual database was created. The following examples can serve to illustrate the problem.

> *** Link: 1-1 ***
>
> I dag genomgår den mänskliga kommunikationen djupgående förändringar genom uppkomsten av globala nätverk för information och kommunikationstjänster, för text, ljud och bilder.
>
> Today, we are experiencing profound changes in human communications with the arrival of global networked information and communication services for text, sound and images.

The point in question here is the alignment. The information content is found in one sentence in Swedish and one sentence in English. Assuming that the information is saved as records or text lines on disk one could easily work out a simple script in any standard Unix language that could search for a word in Swedish: if it is found, then read the next line and the English equivalent will be located there. In such a case each record holds a certain unit of information and is directly related to the following record. So there is a 1-1 relationship between the bits of information and a 1-1 relationship between the orthographic sentences.

When a third language is involved the inappropriateness of this format becomes apparent.

> *** Link: 3-1 ***
> I dag genomgår den mänskliga kommunikationen djupgående förändringar genom uppkomsten av globala nätverk för information och kommunikationstjänster, för text, ljud och bilder. Dessa nätverk innebär spännande utmaningar. De kan också märkbart påverka det språk vi använder.
> Das Auftauchen globaler vernetzter Informations- und Kommunikationsdienste für Text, Ton und Bilder bewirkt heute tiefgreifende Veränderungen in der interpersonellen Kommunikation, die faszinierende Herausforderungen mit sich bringen, gleichzeitig aber auch die Sprachen, die wir benutzen, stark beeinflussen.

In a sense we have the same as the above: one unit of information juxtaposed with the same information in another language. But it now requires three orthographic sentences in the first language and only one sentence in the second. One can still use fairly simple tools but this requires repeating the one version, Swedish in this case, every time that it is aligned with a new language.

If we move our focus from text files on disk to rows and tables in a database the inappropriateness of this format becomes all the more apparent. The various language versions will have to be repeated for every combination with a new language. We will need tables for both the Swedish-English and Swedish-German versions, despite the fact that the Swedish element in both cases is identical. If this is repeated for the other major languages of PEDANT, French, Italian and Spanish, and if to this are added all the combinations of the other languages, French-German, French-Italian, etc., this format becomes quite unwieldy.

This is however easily alleviated by keeping the one version in a separate file from the other, so that instead of two languages in one file, we have two files, each containing one language version. There is still a 1-1 relationship between every line in the two files with regard to content but there is also the possible discrepancy between the orthographic sentences. The easiest way to efficiently access the two pairs is to index the files by word. Each word is given a pointer to the line where it occurs and each line is given a pointer to the equivalent line in the second language version.

This works well for two languages, but when a third one is involved, problems arise. Since the German and English versions quite naturally do not have the same relationship to Swedish with regard to alignment of orthographic sentences they cannot be used against each other. If we want to align them with each other, we will have to create new versions that are correctly aligned with each other, resulting in copy upon copy every time other languages are involved. If we move over from disk files to a database the situation becomes complicated. Even if we were only interested in comparing Swedish to five other languages, the Swedish version will have to be repeated five times.

What is called for is a format that keeps the orthographic sentences stored

in one unit, be it text record or table row, and a separate mechanism that keeps track of which sentence or sentences are aligned with which sentence or sentences in another language. The various text versions can then remain stable with only a small amount of indexing information changing for each language pair. The next two sections will explain how this was done.

Text representation

Independent of the problems mentioned in the previous section, it was obvious that a standard format for the texts would be required, if for no other reason than to provide specifications for software development. At this point "standard" need not refer to anything other than accepted practice within the project.

The PEDANT project decided to use the TEI as the basis for its corpus encoding. The deciding factor was the software package developed by the MULTEXT project consisting of a straightforward API between C programs and SGML encoded files (Thompson et al. 1995). The package is called "Normalised SGML Library" (NSL). This led to a fairly simple solution of the problems described above. PEDANT's first experiences with this package have been described in an earlier report (Danielsson and Ridings 1996a) but NSL has improved since then, partly due to our requests, and our working methods have changed accordingly.

External structure

The lack of information in the literature on parallel texts concerning the issues of representation, access and storage has been pointed out by Armstrong (1996: 17). The present section will be PEDANT's contribution to filling some of these gaps. For this reason it will contain a greater amount of detail than would be expected.

Every language in the PEDANT collection is contained in its own corpus. There is *pedant-se*, *pedant-en*, *pedant-de*, etc.[1] Each corpus is technically a monolingual corpus with no explicit information relating the one language to the other. Each corpus consists of a corpus header followed by the individual texts of that language, each with its own text header. In this respect each corpus is a straightforward implementation of the TEI specifications for corpora. There is no higher element, no project header, encompassing all the corpora.

The SGML elements in the individual texts are very limited: <p> and <s> for the most part, that is, a division into paragraphs and sentences, which is exactly what is required by our aligner. It is technically possible to define other elements as synonyms for <p> and <s> for the aligner, thus allowing for a richer level of annotation, but this is not done to any great extent.[2]

The following two extracts, one from the Swedish corpus and one from the English corpus, are typical.

<P id=se-000001.13><S id=se-000001.13.1 LANG=se>Europeiska rådet noterar med tillfredsställelse några anmärkningsvärda framgångar på området för yttre förbindelser som har uppnåtts sedan dess senaste möte och i vilka Europeiska unionen har spelat en avgörande roll:</S></P>
<P id=se-000001.14><S id=se-000001.14.1 LANG=se>- Undertecknandet av Dayton-avtalet i Paris som sätter punkt för det förödande kriget i det f.d. Jugoslavien och som grundar sig på avsevärda europeiska ansträngningar under de gångna månaderna på det militära och humanitära området samt inom ramen för de förhandlingar som förts; Europeiska rådet erkänner Förenta staternas avgörande bidra vid en ytterst viktig tidpunkt.</S></P>

The equivalent text from the English corpus displays an equivalent structure.

<P id=en-000001.13><S id=en-000001.13.1 LANG=en>The European Council notes with satisfaction some significant achievements in the area of external relations which have occurred since its last meeting and in which the European Union has played a decisive role:</S></P>
<P id=en-000001.14><S id=en-000001.14.1 LANG=en>- the signing in Paris of the Dayton Agreement, which puts an end to the terrible war in former Yugoslavia and builds on considerable European efforts over the preceding months in military, humanitarian and negotiating terms.</S><S id=en-000001.14.2 LANG=en>The European Council recognizes the decisive contribution made by the United States at a crucial moment;</S></P>

Each text, paragraph and sentence is assigned a unique identification number, `id=xx`. The text-id is simply an incrementing number. The sequence is based on Swedish texts. So the first Swedish text is se-000001, the second se-000002, the third se-000003 and so on. It should be recalled that Swedish is the pivotal language so there will be no gaps in the numbering. The corresponding text in the other languages receive the same numbers with prefixes according to the language: de, en, fr, it and es. The numbering for other languages will display gaps, since not every text in Swedish is represented in all the other languages. Leaving gaps in the numbering allows us to integrate such missing texts at a later date. At the same time it is easy to identify which texts correspond to each other since it is seen in the file naming conventions.

The important point to note here is that no alignment information is recorded in these collections. From the technical point of view every language is stored as if it were a monolingual corpus; the SGML *document type* of every collection is `TEICORPUS.2` and can be worked with independently as such. An extract of the corpus header for Swedish can be seen below:

```
<!DOCTYPE teiCorpus.2 SYSTEM "tei2.dtd" [
<!ENTITY % TEI.extensions.ent SYSTEM "pedant.ent">
<!ENTITY % TEI.extensions.dtd SYSTEM "pedant.dtd">
]>
<teiCorpus.2>
```

```
<teiHeader  type=corpus>
 <fileDesc>
  <titleStmt>
   <title>PEDANT : Swedish component</title>
   <respStmt>
    <name>Authors</name>
    <resp>Collection and Alignment</resp>
   </respStmt>
   ...
```

The first line declares the document to be a TEI corpus document.[3] The second and third lines contain our customizations of the TEI dtds that we use. We do not use the TEI unaltered but have introduced numerous customizations using the recommended mechanisms for doing so (Sperberg-McQueen and Burnard 1994: 737–744). They are collected in two files that are unique for the project, *pedant.ent* and *pedant.dtd*.

One of the changes we make has to do with the TEI's element <W>. The first thing we do, in the file *pedant.ent*, is to block the TEI's own declaration of this element:

```
<!ENTITY % w 'IGNORE' >
```

There are two things to remember here: (a) Entities get their values on a first come first serve basis and (b) the file *parole.ent* is read and dealt with *before* the relevant dtd in the TEI system. So when the parser arrives at the following declaration in *teiana2.dtd*, the declaration for <W> is ignored:

```
<!ENTITY % w 'INCLUDE' >
<![ %w; [
<!ELEMENT %n.w;   - -   ((#PCDATA | %n.seg; | %n.w; |
                        %n.m; | %n.c;)*)                >
<!ATTLIST %n.w;         %a.global;
                        %a.seg;
          lemma         CDATA                 #IMPLIED
          TEIform       CDATA                 'w'       >
]]>
```

The %w entity has already received a value of "IGNORE" so the whole section is skipped. In the file *pedant.dtd* we have our own project variant of the element as follows:

```
<!ELEMENT %n.w;    - -  ((#PCDATA | %n.seg; | %n.w; |
                        %n.m; | %n.c;)*)                >
<!ATTLIST %n.w;         %a.global;
                        %a.seg;
                        %a.xPointer;
```

```
lemma          CDATA          #IMPLIED
msd            CDATA          #IMPLIED
TEIform        CDATA          'w'          >
```

We added some of our own attributes. We do not use the ANA attribute to provide morphosyntactic tags, but MSD, "morphosyntactic description." There are two reasons for this. In the first place, we do not want to get involved in a complicated mechanism involving IDREFS, which is what ANA expects to be assigned. It is attractive, but complicated, and all the more so since our tags can contain the character "@", which is invalid in those contexts. In the second place, the Parole project (LE2-4017) in which we are involved, uses the MSD attribute, inherited from EAGLES.

In addition to the new attribute for morphosyntactic tags, we also wanted a LEMMA attribute in order to simplify searches and other actions that function best when one has access to lemmata rather than only-word types.

Up to this point our methods of working with the material differ little from those associated with standard monolingual text collections, but the situation changes when our method of providing explicit links between translation equivalents is considered.

Alignment information

For the reasons explained above we decided to keep all details about alignments in separate documents. One document contains all alignments between two languages. There is a document for Swedish-English and another document for Swedish-German and if we ever decide to align English with German we would have yet another document for English-German. Each one of these documents contain all of the alignment information for all the texts for the language pair it describes.

The way we do this is by creating a new "corpus", technically speaking, but the individual "texts" in the corpus document (the SGML corpus document) do not contain any natural language, only alignment links. For any given document from the monolingual collection that has been aligned with another language there will be the same number of "paragraphs" and the same number of "sentences", that is, the same number of <P> and <S> elements, but their content will be minimal.

We have introduced two new elements to the TEI system, <SSEG> and <TSEG>, "source segments" and "target segments." We see the results of alignment in pairs; one segment of one text is aligned with one segment of another text. These segments can contain a combination of "sentences". The source segment might consist of two sentences while the target segment consists of only one. The possible combinations are: 1-1, 2-1, 2-2, 1-0 and vice versa. An alignment "pair", consisting of exactly one <SSEG> and one <TSEG>, is itself contained in one single <SEG>, which should possibly be renamed to <ASEG>, alignment segment, by analogy.

So a paragraph can contain one or more <SEG> elements, which each contain an alignment pair, <SSEG> and <TSEG>, and each of these latter elements can contain one or more sentences which in turn can contain one or more words and so on down the branches of the SGML document tree, i.e.:

```
<P>
   <SEG>
        <SSEG></SSEG>
        <TSEG></TSEG>
   </SEG>
</P>
```

Information is attached to the <SSEG> and <TSEG> elements that points back into the relevant monolingual corpora where the actual sentences are found. For example, an alignment in the Swedish-English collection appears as follows:

```
<P ID=SE-000001.13>
<SEG ID=SE-000001.S13>
<SSEG DOC=sedoc FROM='id (SE-000001.13.1)' ID=SE-000001.SS13></SSEG>
<TSEG DOC=endoc FROM='id (EN-000001.13.1)' ID=EN-000001.TS13></TSEG>
</SEG>
</P>
<P ID=SE-000001.14>
<SEG ID=SE-000001.S14>
<SSEG DOC=sedoc FROM='id (SE-000001.14.1)' ID=SE-000001.SS14></SSEG>
<TSEG DOC=endoc FROM='id (EN-000001.14.1)'
                TO='id (EN-000001.14.2)' ID=EN-000001.TS14></TSEG>
</SEG>
</P>
```

The various elements for "segments" receive their own ID values, since they do not occur anywhere else than in this document. The <SSEG> and <TSEG> elements have additional attributes, DOC, FROM and TO. Alignment information is assigned to these attributes.

The DOC attribute provides information on the corpus in which the sentences can be found, that is, it points to the relevant monolingual corpus.

The FROM and TO attributes contain the extent, measured in sentences, in the monolingual corpus. The values they are assigned, are the values that have been assigned to the ID attribute of the sentences in question. A missing TO attribute defaults to the same as the FROM attribute, that is, one sentence. In the first paragraph in the example above, we have a 1-1 alignment. In the second paragraph we have a 1-2, that is, one Swedish sentence corresponds to two English sentences.

Similar information is recorded for every aligned text in PEDANT. The information is stored as its own corpus, that is, a Swedish-English corpus, a Swedish-German corpus, etc. The corpus header is as follows:

```
<!DOCTYPE teiCorpus.2 SYSTEM "tei2.dtd" [
<!ENTITY % TEI.extensions.ent SYSTEM "pedant.ent">
<!ENTITY % TEI.extensions.dtd SYSTEM "pedant.dtd">
<!ENTITY sedoc SYSTEM "/Pedant/Corpus/Swedish/Swedish.nsg" CDATA SGML>
<!ENTITY endoc SYSTEM "/Pedant/Corpus/English/English.nsg" CDATA SGML>
]>
<?NSL LINKS S SSEG S TSEG>
<teiCorpus.2>
 <teiHeader  type=corpus>
  <fileDesc>
   <titleStmt>
    <title>PEDANT : Swedish-English component </title>
    <respStmt>
     <name>Authors </name>
     <resp>Collection and Alignment</resp>
    </respStmt>
    ...
```

The beginning of the header contains three significant lines: 4 and 5, and 7. Recall the value assigned to the DOC attributes of the <SSEG> and <TSEG> elements: sedoc and endoc. Lines 4 and 5 map these to the relevant monolingual corpora. Line 7 is a *processing instruction* and is unique for the LT NSL package we are using as mentioned above. It provides the information that <S> elements can be linked to <SSEG> and <TSEG> elements.

All of this is tied together via the LT NSL utility called *mknsg*. We run the command as follows:

```
mknsg -D sgml -D sgml/pedant Swedish-English.sgm | <further processing>
```

This command ties all the information together and prints to standard output where it can be piped into other programs for further processing or redirected to disk. The output is as follows:

```
<P ID=SE-000001.13>
<SEG ID=SE-000001.S13>
<SSEG ID=SE-000001.SS13>
<S ID=SE-000001.13.1 LANG=SE>Europeiska rådet noterar med tillfredsställelse
några anmärkningsvärda framgångar på området för yttre förbindelser som har uppnåtts
sedan dess senaste möte och I vilka Europeiska unionen har spelat en avgörande
roll:</S></SSEG>
<TSEG ID=EN-000001.TS13>
<S ID=EN-000001.13.1 LANG=EN>The European Council notes with satisfaction
```

```
some significant achievements in the area of external relations which have occurred since
its last meeting and in which the European Union has played a decisive role:</S>
</TSEG>
</SEG>
</P>
<P ID=SE-000001.14>
<SEG ID=SE-000001.S14>
<SSEG ID=SE-000001.SS14>
<S ID=SE-000001.14.1 LANG=SE>- Undertecknandet av Dayton-avtalet i Paris
som sätter punkt för det förödande kriget i det f.d. Jugoslavien och som grundar sig på
avsevärda europeiska ansträngningar under de gångna månaderna på det militära och hu-
manitära området samt inom ramen för de förhandlingar som förts; Europeiska rådet er-
känner Förenta staternas avgörande bidrag vid en ytterst viktig tidpunkt.</S>
</SSEG>
<TSEG ID=EN-000001.TS14>
<S ID=EN-000001.14.1 LANG=EN>- the signing in Paris of the Dayton Agree-
ment, which puts an end to the terrible war in former Yugoslavia and builds on consid-
erable European efforts over the preceding months in military, humanitarian and nego-
tiating terms.</S>
<S ID=EN-000001.14.2 LANG=EN>The European Council recognizes the deci-
sive contribution made by the United States at a crucial moment;</S></TSEG>
</SEG></P>
```

This might seem to be a complicated procedure at first sight, but most of the details are fully automated and there are some further advantages in that all software can build upon the same library. This will be illustrated by our lemmatizer, PEDAL.[4]

The expansion of the links, as mentioned above, can be saved to disk or piped into other SGML tools. The LT NSL package has introduced a way of working with "semivalid sgml". A valid SGML file is run through the basic tool *mknsg*, which parses the DTD permanently and caches a binary version on disk. This cached version is the one used by all subsequent programs. One of the characteristics of "semivalid sgml" is that the segment piped into a program must be valid when compared to the cached DTD, but the segment piped in need not itself contain the whole document tree. The excerpt above, for instance, would be valid input to other NSL tools, since the elements and their contents are valid. There is no header, no <TEXT> elements, but the <P> elements are syntactically correct with regard to that part of the DTD that deals with them.

One of the tools, *sggrep*, that comes along with the library can help to exemplify this. Let us assume that we want to identify all translation pairs in which the Swedish half contains the expression "sätter punkt för". To do so, we would replace the <further processing> in the above command with:

```
sggrep ".*/SEG" "SEG/SSEG" "sätter punkt för"
```

The first parameter provides the program with the depth of our query in the SGML document tree, namely, down to the level of <SEG>. The second parameter provides the subquery, the space in which the search will be done. In effect, it defines the segment of the SGML document we want reported from the query. Recall that alignment pairs are contained in the <SEG> element, exactly one pair per element. So the result returned will be the whole <SEG> element. The actual query, however, will be limited to the <SSEG> element. The program will not look for "sätter punkt för" in the English element, <TSEG>, but it will be returned together with the Swedish segment, since it too is contained by the <SEG> element. The result is similar to what appears in the excerpt above and is, in fact, the way it was produced for this report. It could have been piped into yet other tools performing statistical analysis, lemmatization, etc. The possibilities are numerous when all tools work with the same API. Further examples can be found in an earlier report (Danielsson and Ridings 1996a: 7–12).

Database

The original database manager was Microsoft Access and it is still being used for quite a few tasks successfully. Since we wanted to provide network access as well, we began experimenting with some of the freely available relational databases for Unix: miniSQL and mySQL.

Network access

Network access is provided by way of (a) miniSQL and more recently mySQL, (b) cgi-scripts and (c) any browser on any platform that supports tables and forms. Those interested can turn to *http://svenska.gu.se/PEDANT* for a demonstration.

Some of our first attempts at identifying equivalents on the basis of one Swedish word can be seen in figure 1 in the Appendix. It is a web-based system working against the mySQL database.

Linguistic tools

The tagset

PEDANT uses a tagset that is mappable on a 1-1 basis with the SUC tagset. The SUC tagset was designed by Eva Ejerhed. In February 1997 Ejerhed and Ridings adjusted the PAROLE tagset in such a way that the PAROLE tagset and the SUC tagset are interchangeable. This is evidenced in the SGML version of the SUC corpus. A table comparing the two sets together with example words can be found at *http://ldb20.svenska.gu.se*.

Unlike our lemmatizer, Brill's tagger has not yet been made SGML aware. The tagger and the alignment program are the only two programs left in our repertoire that have not been written for the NSL API. In the case of Brill's tag-

ger we feel that there are other adjustments we want to make, particularly in the lexical rule component, but the time or human resources are not yet available.

This is not so problematic since Brill's tagger requires two properties of a text that is to be tagged: (a) it must be segmented into sentences and (b) it must be tokenized.

Our texts must be segmented into sentences before they can be aligned so the first requirement is met. With regard to the second requirement, we have a tokenizer that works directly with the SGML files through the NSL API. This means that we can export our texts from SGML in such a way that there is a 1-1 relationship between the exported file's sentences and tokens and the original texts. This being the case it is a simple matter to take the tagged results from Brill's tagger and map the morphosyntactic tags back onto the proper attributes of each token.

The results of tagging are as follows:

```
<P ID=SE-000001.13>
<S ID=SE-000001.13.1 LANG=SE>
<W MSD='AQP0SNDS'>Europeiska</W>
<W MSD='NCNSN@DS'>rådet</W>
<W MSD='V@IPAS'>noterar</W>
<W MSD='SPS'>med</W>
<W MSD='NCUSN@IS'>tillfredsställelse</W>
<W MSD='DI@0P@S'>några</W>
<W MSD='AQP0SNDS'>anmärkningsvärda</W>
<W MSD='NCUPN@IS'>framgångar</W>
<W MSD='SPS'>på</W>
<W MSD='NCNSN@DS'>området</W>
<W MSD='SPS'>för</W>
<W MSD='AQC00N0S'>yttre</W>
<W MSD='NCUPN@IS'>förbindelser</W>
<W MSD='PH@000@S'>som</W>
<W MSD='V@IPAS'>har</W>
<W MSD='V@IUPS'>uppnåtts</W>
<W MSD='RG0S'>sedan</W>
```

A file that is morphosyntactically tagged is then run through PEDAL, the lemmatizer, assigning one of three alternatives to the LEMMA attribute of the <W> element: (a) a lemma if a lemma is found with a matching morphosyntactic tag, (b) "not-found" if it was not possible to resolve the word type to its base form, or (c) "no-msd-match" in the event that the word type could be resolved to a base form but the morphosyntactic description associated with the word type does not match the description provided by PEDAL.

PEDAL works directly with SGML files and provides an opportunity to illustrate how simple it is to integrate an SGML parser with one's own code. The core of the lemmatizer is made up of the following lines of code:

```
strcpy(qustr, ".*/W");
qu=ParseQuery(dct, qustr);

while( ( item=GetNextQueryItem(inf, qu, outf ) ) ) {
    msdVal = GetAttrStringVal(item, AttrName);
    strcpy(wordtype, item->data->first);
    if (msdVal != NULL) {
        if (*msdVal == 'N'
            || *msdVal == 'V'
            || *msdVal == 'A' || *msdVal == 'D'
            || *msdVal == 'P') {
            strcpy(lemma,lemmatize(wordtype, msdVal));
        } else {
            strcpy(lemma, item->data->first);
        }
        if (!strcmp("not-found", lemma)) {
            strcpy(lemma,guess_lemmatize(wordtype, msdVal));
        }
        PutAttrVal(item, LemmaAttr, lemma);
    }
    PrintItem(outf, item);
    FreeItem(item);
}
```

The first two lines set up the SGML query, `".*/W"`. The dot is a wildcard meaning "any" and the star is the standard kleene star meaning "zero-or-more". What is being referred to here are elements in the SGML document tree. On the fourth line the query can be read as "search down the SGML tree, traversing all elements until we descend down to the level of <W>". That is the base element of our documents. There are no other elements below the <W> element. The API passes all elements down to that level to the output, but turns over <W> elements to the program for processing. After processing this base element is printed to output by the third line from the end, `PrintItem(outf, item)`, thus completing the whole document.

The fifth line:

```
msdVal = GetAttrStringVal(item, AttrName);
```

reads the MSD attribute, the morphosyntactic tag, of each <W> element. This is passed on to the lemmatizing routine. At this point we only lemmatize nouns, verbs, adjectives, determiners and pronouns. All other classes of words simply get their word type copied to the LEMMA attribute (line 14).

The word type and the morphosyntactic tag are then sent to the lemmatizing routine (line 12). The core of the lemmatizing routine is:

```
resp = recognizer(wordtype,  Lang, 0, 0, (FILE *)NULL);
strcpy(lemma, "not-found");
if (resp != (RESULT *)NULL) {
    strcpy(lemma, "no-msd-match");
    for (rp=resp; rp; rp=rp->link) {
        sscanf((char *)rp->feat, "\[%s %s", tmp_l, tmp_a);
        eos = (char *)rindex(tmp_a, '\0');
        --eos; *eos = '\0';
        if (!strcmp(tmp_a, msdtag)) {
            strcpy(lemma, tmp_l);
        }
    }
    free_result(resp);
}
return(lemma);
```

The first line searches for all the possible base forms of the word type. Before anything else is done, the lemma is set to "not-found". If it is found, then this will be overwritten, if it is not found, then this will be returned. In line 3 a check is made for results: if there were results, then the lemma is set to "no-msd-match", that is, a lemma was identified, but its morphosyntactic description did not match the word type's. This will be overwritten if it proves not to be the case.

The `for`-loop walks through all of the possible interpretations of the word type. The `if`-statement in the loop compares each interpretation's morphosyntactic description with that of the original word type. If they match, then the proposed lemma is copied to the lemma string and will eventually be returned. If no matches are found, then the lemma string retains "no-msd-match" and this is returned. This signals all the places where the results of the tagger deserve manual control. There can be other places as well, but this catches a lot of the mistakes, though in practice, they are not that many. The resulting output is as follows:

```
<P ID=SE-000001.13>
<S ID=SE-000001.13.1 LANG=SE>
<W LEMMA='europeisk' MSD='AQP0SNDS'>Europeiska</W>
<W LEMMA='råd' MSD='NCNSN@DS'>rådet</W>
<W LEMMA='notera' MSD='V@IPAS'>noterar</W>
<W LEMMA='med' MSD='SPS'>med</W>
<W LEMMA='tillfredsställelse' MSD='NCUSN@IS'>tillfredsställelse</W>
<W LEMMA='någon' MSD='DI@0P@S'>några</W>
<W LEMMA='anmärkningsvärd' MSD='AQP0SNDS'>anmärkningsvärda</W>
<W LEMMA='framgång' MSD='NCUPN@IS'>framgångar</W>
<W LEMMA='på' MSD='SPS'>på</W>
<W LEMMA='område' MSD='NCNSN@DS'>området</W>
<W LEMMA='för' MSD='SPS'>för</W>
```

```
<W LEMMA='yttre' MSD='AQC00N0S'>yttre</W>
<W LEMMA='förbindelse' MSD='NCUPN@IS'>förbindelser</W>
<W LEMMA='som' MSD='PH@000@S'>som</W>
<W LEMMA='ha' MSD='V@IPAS'>har</W>
<W LEMMA='uppnå' MSD='V@IUPS'>uppnåtts</W>
```

Three lines in the code have not been discussed yet:

```
if (!strcmp("not-found", lemma)) {
    strcpy(lemma,guess_lemmatize(wordtype, msdVal));
}
```

As mentioned above, if the lemmatizing routine does not find a match, it returns the string "not-found". It turns out that almost all of these cases are compound words. This is a familiar problem for all of those dealing with Germanic languages (Hellberg 1978: 21–28; Karlsson 1992: 15–17) and renders otherwise attractive publicly available packages more or less useless.

Our approach to this is simple but works satisfactorily. We have a limited lexicon section that allows certain forms to lead back into the main lexica (cf. Karlsson 1992: 15). In general, however, we are working with the assumption that (a) "not-found" words are compounds and (b) that the longest segment on the right-hand side of the compound for which a base form is identified with the correct morphosyntactic description, provides us with the best compound boundary. In other words, the routine `guess_lemmatize` walks through the word type backwards and returns the longest segment.

The words that have been identified by the second form of the lemmatizing routine can be easily identified, i.e.:

```
pedal Swedish.msd.nsg | sggrep -r ".*/W" "W[LEMMA='.*_.*']" ""
```

The above command pipes the result of the lemmatizer into one of the tools that is included in the NSL package, *sggrep*, a version of *grep* that understands the structure of SGML documents.

The first parameter of the command, `".*/W"`, provides the depth of the query, that is, all the way down the document tree to the level of the `<W>` element. The second parameter, `"W[LEMMA='.*_.*']"`, provides the subquery, that is, the scope of the document that will be queried and returned. A search on attributes to an element is provided within square brackets, the `LEMMA`, in this case. The third parameter is empty because we are not searching for specific words (element content), but for words with certain attributes. Attribute values can be expressed with regular expressions if the `-r` flag is provided. So we are searching for all lemmata which contain an underscore, put there to mark the suggested compound boundary. The result is:

```
<W LEMMA='regerings_konferens' MSD='NCUSN@DS'>regeringskonferensen</W>
<W LEMMA='reflexions_grupp' MSD='NCUSG@DS'>reflexionsgruppens</W>
```

```
<W LEMMA='morgon_dag' MSD='NCUSG@DS'>morgondagens</W>
<W LEMMA='Dayton-_avtal' MSD='NCNSN@DS'>Dayton-avtalet</W>
<W LEMMA='åtgärds_plan' MSD='NCUSN@DS'>åtgärdsplanen</W>
<W LEMMA='Barcelona_förklaring' MSD='NCUSN@DS'>Barcelonaförklaringen</W>
<W LEMMA='medelhavs_område' MSD='NCNSN@DS'>medelhavsområdet</W>
<W LEMMA='medelhavs_område' MSD='NCNSN@DS'>Medelhavsområdet</W>
<W LEMMA='åsikts_utbyte' MSD='NCNSN@IS'>åsiktsutbyte</W>
<W LEMMA='Europa_parlament' MSD='NCNSG@DS'>Europaparlamentets</W>
<W LEMMA='diskussions_fråga' MSD='NCUPN@DS'>diskussionsfrågorna</W>
<W LEMMA='valuta_enhet' MSD='NCUSN@DS'>valutaenheten</W>
<W LEMMA='anpassnings_kostnad' MSD='NCUPN@DS'>anpassningskostnaderna</W>
<W LEMMA='råds_förordning' MSD='NCUSN@IS'>rådsförordning</W>
<W LEMMA='ecu-_korg' MSD='NCUSN@DS'>ecu-korgen</W>
<W LEMMA='valuta_enhet' MSD='NCUPN@IS'>valutaenheter</W>
<W LEMMA='Ekofin-_råd' MSD='NCNSN@DS'>Ekofin-rådet</W>
<W LEMMA='euro-_sedel' MSD='NCUPN@DS'>euro-sedlarna</W>
<W LEMMA='Budget_disciplin' MSD='NCUSN@DS'>Budgetdisciplinen</W>
<W LEMMA='budget_ordning' MSD='NCUSN@DS'>budgetordningen</W>
<W LEMMA='euro-_område' MSD='NCNSN@DS'>euro-området</W>
```

One of the above analyses is the result of a previous correction of PEDAL's lexicon, namely:

```
<W LEMMA='medelhavs_område' MSD='NCNSN@DS'>medelhavsområdet</W>
```

The original analysis put the compound boundary after "medel" (middle), since "havsområde" (sea-area) is listed in the lexicon making it the longest right-hand segment. We added "medelhavs" to the lexical listings with a continuation into the noun lexicon. This results in the compound being recognized and returned instead of "not-found" and the guessing routine never gets called. All other forms in the above list have been produced by the principle of longest segment to the right with matching MSD values of the original word type.

Excursus

The method of storing alignment information was outlined above. At this point one of the benefits of our architecture can be indicated by drawing attention to the following lines:

```
<!ENTITY sedoc SYSTEM "/Pedant/Corpus/Swedish/Swedish.nsg" CDATA SGML>
<!ENTITY endoc SYSTEM "/Pedant/Corpus/English/English.nsg" CDATA SGML>
```

These lines linked back into the individual monolingual corpora for the language pair that was aligned. The versions of *Swedish.nsg* and *English.nsg* in these lines are the versions with a minimum of mark-up: paragraphs and sentences.

Let us assume that the lemmatization described in the previous section was saved to disk, rather than just piped into other tools, for example to *Swedish.lemma.msd.nsg*. We can then change the two lines above to read:

```
<!ENTITY sedoc SYSTEM
            "/Pedant/Corpus/Swedish/Swedish.lemma.msd.nsg" CDATA SGML>
<!ENTITY endoc SYSTEM "/Pedant/Corpus/English/English.nsg" CDATA SGML>
```

This provides us with the same information about alignment pairs, but this time we will have access to a richer array of information when it comes to the Swedish component, namely lemmata and morphosyntactic descriptions. The paragraph and sentence IDs are the same in both versions, the base version and the morphosyntactic tagged version, so the alignment information in the Swedish-English component will point back to the same sentences.

This method of working allows us to experiment with various levels of annotation without cluttering up the base version. One level of annotation that interests us at the moment is one that marks phrases below the sentence level. This will be dealt with below.

Current directions

This section is only a preliminary sketch of some of the directions that are being pursued at the moment. It is by no means exhaustive since there are now many others involved with various investigations and they will be reported on in their own time by the individual researchers involved.

Equivalents below the sentence level

Once a collection of orthographical sentences has been correctly aligned with translation equivalents the most natural next step is to identify smaller chunks. Word-to-word alignment will always be difficult since most translations do not display such a structure.[5]

The department's background is lexicographical and the work being done in PEDANT reflects that fact. One of our distant goals is to create a lexicographical workbench for bilingual lexicography. A major aspect of this goal is to introduce the methods from corpus-based monolingual lexicography to the multilingual sphere.

Word tuples

N-grams models of word-tuples are basically lists of word-pairs, word-triples, word-quadruplets and so on that are provided with frequency and likelihood information (Atwell 1996: 160). Likelihood, in this section, is based on the likelihood ratio in Dunning (1994).[6]

Lists are created of all bigrams, trigrams and quadrigrams in a subset of the Swedish-English component of PEDANT.[7] For examples of Swedish and English trigrams see figures 2 and 3 respectively in the Appendix.

The motivation for performing the tests was to see if "phrases", in a loose sense, in one language showed any inclination to be translated by phrases in another language. There cannot be that many translations of "as soon as possible" into Swedish other than "så snart som möjligt" and there might even be other hidden combinations of words that do not directly come to mind.

The purpose, at this stage, is not to identify which phrases are translations of each other. The tables are sorted in descending frequency according to the significance of the "tuple" in its own language, and the equivalent in the other language, when compared to its own corpus of words, cannot be expected to show the same degree of equivalence. There is simply no connection between the two with regard to the ranking the phrases get in their respective languages.

The assumption is that a phrase in one language might be translated by a phrase in another. If this is the case, then, should we search in the parallel texts, have a translation pair in front of us and find a phrase in L1 and also in L2, it is worth noting if the phrase in L2 is a translation of the phrase in L1. This method might succeed if the language pairs are not overloaded with phrases, which seems to be a fair assumption to begin with, considering the fact that the vast majority of alignments are 1-1, that is, one sentence to one sentence.

The "phrases" identified by the $-2 \log \lambda$ formula have been automatically tagged with a global search-and-replace command so that they are enclosed in the <PHR>...</PHR> chunk. This was done by taking the first 75 trigrams and thereafter the first 75 qaudrigrams.[8]

As mentioned above, this is just experimental and not a full report, but the first indications are encouraging. In the excerpt below we see the results of searching for the English phrase "with regard to." The following is the result of piping our corpus of pure links as described above through the *sggrep* utility:

```
mknsg -D sgml -D sgml/pedant Swedish-English.sgm \
  | sggrep ".*/SEG" "SEG/.*/TSEG/.*/PHR" "with regard to"
```

— in other words, search for the phrase "with regard to" only in the English half of the translation pairs.

> <SEG ID=SE-000001.S119>
> <SSEG ID=SE-000001.SS119><S ID=SE-000001.91.1 LANG=SE>- förbättra deras finansiella miljö genom en förbättrad tillgång till kapitalmarknaderna och främja utvecklingen av europeiska investeringsfondens funktion <PHR>**när det gäller**</PHR> <PHR> små och medelstora</PHR> företag.</S></SSEG>
> <TSEG ID=EN-000001.TS119><S ID=EN-000001.91.1 LANG=EN>- improve the financial environment for them by means of better access to capital markets and encourage development <PHR>**of the European**</PHR> Investment Fund function <PHR> **with regard to**</PHR> SMEs.</S></TSEG>

</SEG>
<SEG ID=SE-000004.S345>
<SSEG ID=SE-000004.SS345><S ID=SE-000004.196.2><PHR>**Små och medelstora företag**</PHR>, som utgör nästan hälften av den ekonomiska basen, möter speciella svårigheter och särskilt i <PHR>**frågor som rör**</PHR> finansiering (t.ex. den effektiva räntan är ofta 2 till 3 punkter högre än i utvecklade regioner), men även avseende möjlighet till samarbete, tillgång till teknisk kompetens eller ledningskompetens, m.m.</S></SSEG>
<TSEG ID=EN-000004.TS345><S ID=EN-000004.196.2>**The** SMEs, which make up virtually the entire economic fabric encounter special difficulties there, particularly <PHR>**with regard to**</PHR> financing (e.g. actual interest rates are often 2-3 points higher than in the more developed regions) but also <PHR>**with regard to**</PHR> cooperation opportunities, access to sources of technical or management skills, etc.</S>
</TSEG>
</SEG>

We achieve similar results by searching for Swedish phrases or parts of them, "syssel" in this case.

```
mknsg -D sgml -D sgml/pedant Swedish-English.sgm \
  | sggrep ".*/SEG" "SEG/.*/SSEG/.*/PHR" "syssel"
```

Note the change from `TSEG` to `SSEG` in the subquery of *sggrep*. A sample of the output is as follows:

<SEG ID=SE-000002.S78>
<SSEG ID=SE-000002.SS78><S ID=SE-000002.45.1 LANG=SE>Inom denna ram har Europeiska rådet uppmärksammat det italienska ordförandeskapets avsikt att inför mötet i Florens sammankalla en trepartskonferens i Rom i mitten av juni mellan regeringarna, arbetsmarknadens parter och kommissionen om <PHR>**tillväxt och sysselsättning**</PHR>.</S></SSEG>
<TSEG ID=EN-000002.TS78><S ID=EN-000002.45.1 LANG=EN><PHR>In this context</PHR>, it noted that, in preparation for the Florence meeting of <PHR>**the European Council**</PHR>, the Italian Presidency intended to hold a Tripartite Conference on <PHR>**growth and employment**</PHR>, involving governments, social partners and the Commission, in Rome in mid-June.</S></TSEG>
</SEG>
<SEG ID=SE-000003.S8>
<SSEG ID=SE-000003.SS8><S ID=SE-000003.8.1 LANG=SE>Med stöd av den strategi som det uppnåddes enighet om i Essen samt vitboken diskuterade Europeiska rådet i detalj <PHR>**tillväxt och sysselsättning**</PHR> <PHR>**på grundval av**</PHR> kommissionens meddelande "Insatser för sysselsättning i Europa: en förtroendepakt", den gemensamma interimsrapporten om sysselsättning samt de tidigare dokumenten, inklusive slutsatserna från trepartskonferensen om <PHR>**tillväxt och sysselsättning**</PHR> i Rom den 14-15 juni 1996 och Frankrikes <PHR>**memorandum om en social**</PHR> modell för Europa.</S></SSEG>

<TSEG ID=EN-000003.TS8><S ID=EN-000003.8.1 LANG=EN>Drawing on the strategy agreed in Essen and on <PHR>**the White Paper**</PHR>, <PHR>**the European Council**</PHR> held a detailed discussion on the subject of <PHR>**growth and employment**</PHR> <PHR>**on the basis**</PHR> of the Commission communication entitled "Action for employment in Europe: A confidence pact", the joint interim report on employment <PHR>**as well as**</PHR> the other documents before it, including the conclusions drawn from the Tripartite Conference on Growth and Employment held in Rome on 14 and 15 June 1996 and the French Memorandum on a European social model.</S></TSEG>
</SEG>
<SEG ID=SE-000003.S11>
<SSEG ID=SE-000003.SS11><S ID=SE-000003.9.3 LANG=SE>I överensstämmelse med kommissionens strategi gäller det att sätta igång en öppen och flexibel process som <PHR>**gör det möjligt**</PHR> för alla berörda att göra specifika åtaganden inom sina respektive ansvarsområden <PHR>**för att skapa**</PHR> en för sysselsättningen gynnsam makroekonomisk ram, maximalt utnyttja <PHR>**den inre marknaden**</PHR>s möjligheter, påskynda reformer på arbetsmarknaden och bättre utnyttja unionens politik till förmån för <PHR>**tillväxt och sysselsättning**</PHR>.</S></SSEG>
<TSEG ID=EN-000003.TS11><S ID=EN-000003.9.3 LANG=EN>In line with the Commission's approach, an open and flexible process <PHR>**needs to be**</PHR> got under way which will enable all those concerned to enter into specific commitments at their own level of responsibility <PHR>**in order to**</PHR> create a macroeconomic framework favourable to employment, to exploit to the full the potential of <PHR>**the internal market**</PHR>, to speed up <PHR>**the labour market**</PHR> reforms and to make better use of the Union's policies in the interest of <PHR>**growth and employment**</PHR>.</S></TSEG>
</SEG>

In the first translation pair, we see that "tillväxt och sysselsättning" is isolated as a multiword unit and corresponds to "growth and employment" in the English half of the translation segment. In the second segment we find the same two correspondences, the statistical processing has isolated several other multiword segments. In the English half we find "on the basis of", which corresponds to "på grundval av" in the Swedish half, and which also has been marked as significant. Similar results can be seen in the third and last translation pair. "In order to" has been isolated as a significant trigram but the infinitive falls outside the scope of the tagging. In Swedish, the translation equivalent is "för att skapa" and it has also been isolated as a significant trigram. "The internal market" corresponds to "den inre marknaden" and once again "growth and employment" corresponds to "tillväxt och sysselsättning".

Our immediate efforts will concentrate on investigating the best balance of bigrams, trigrams, quadrigrams and n-grams between the various languages. Many prepositional phrases in English, for example, are translated by fewer orthographic words in Swedish because of compounding.

We also noticed that "phrases" which end in functional words have a fairly predictable pattern of morphosyntactic tags following them and when the

equivalent phrase in another language also ends, for example, with a preposition, the same can be seen. In the previous example, for instance, "in order to" is going to be followed by an infinitive. We want to see how far this will lead us in identifying even more equivalents. In other words, if we know that we have identified phrasal equivalents between two languages and we also know the grammatical constructions that usually follow them, then we want to see if we can isolate previously unidentified equivalents — the material following the phrases — based on how well they fit the grammatical constructions that usually follow the identified phrases.

Notes

1. This is a temporary oversimplification motivated by the desire not to introduce technicalities involved with the NSL package at this early stage.
2. The most pressing need for extra annotation is in the treatment of lists where the list as a whole could be regarded as a <p> and the individual items could be equated, technically, with <s>.
3. The "document" in this case is a whole collection of texts, which make up one SGML document, a "corpus" built according to TEI's specifications. One must be careful not to mix up the terms "documents" and "files" in an SGML context.
4. The L stands for "lemmatizer" and the rest by the same analogy as PEDANT.
5. It does work, however, in some circumstances with very special texts from an industrial domain.
6. The original article appeared in *Computational Linguistics* 19: 61-74, 1993.
7. Press65, a one million word corpus of Swedish newspaper texts from 1965, has been processed by the same routines and took 12 days, nothing for someone who wants to experiment; thus the subcorpus. Over and beyond this, the $-2 \log \lambda$ is supposed to be particularly suited for small texts. The Swedish-English subcorpus that was used for this test contained 60,000+ words per language.
8. This prevents phrases such as "as soon as possible" from being tagged since "soon as possible" will be tagged first. It has been done this way in order to simplify the first experimental probes, preventing problems with SGML's inadequacy in handling overlapping structures, an inadequacy that can only be alleviated by introducing LT NSL's hyperlinking mechanism.

References

Armstrong, S. 1996. *Multilingual Corpora: Survey of Work with Multilingual Texts. Technical Report, EAGLES.* Geneva: Text Corpora Working Group, ISSCO.

Atwell, E. 1996. Machine Learning from Corpus Resources for Speech and Handwriting Recognition. Thomas, J. and M. Short (Eds.). 1996. *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech*: 151–166. London/New York: Longman.

Danielsson, P. and D. Ridings. 1996a. *Annotating Parallel Texts with the NSL Library.* Research Reports from the Department of Swedish, Göteborg University GU-ISS-96-7, Språkdata.

Danielsson, P. and D. Ridings. 1996b. *PEDANT: Parallel Texts in Göteborg*. Research Reports from the Department of Swedish, Göteborg University GU-ISS-96-2, Språkdata.

Dunning, T. 1994. Accurate Methods for the Statistics of Surprise and Coincidence. Armstrong, S. (Ed.). 1994. *Using Large Corpora*: 61–74. Cambridge, MA./London: The MIT Press.

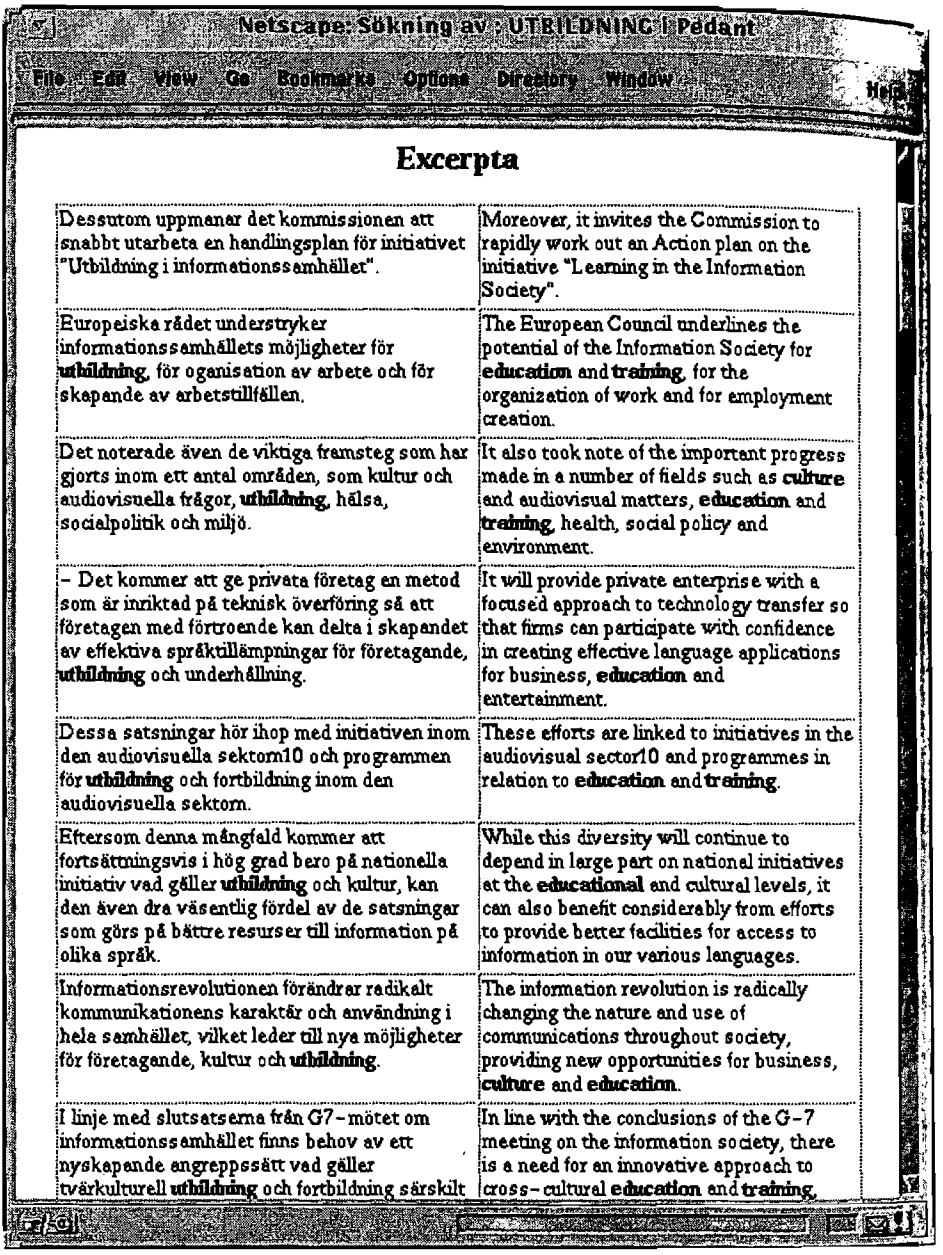
Hellberg, S. 1978. *The Morphology of Present-day Swedish: Word-inflection, Word-formation, Basic Dictionary*. Data linguistica. Stockholm: Almqvist & Wiksell International.

Karlsson, F. 1992. SWETWOL: A Comprehensive Morphological Analyser for Swedish. *Nordic Journal of Linguistics* 15: 1–45.

Sperberg-McQueen, C.M. and L. Burnard (Eds.). 1994. *Guidelines for Electronic Text Encoding and Interchange*. Chicago: ACH, ACL, ALLC.

Thompson, H., S. Finch and D. McKelvie. 1995. *Normalised SGML Library (NSL)*. Multext LRE Project 62-050, University of Edinburgh, The Language Technology Group.

Appendix



Dessutom uppmanar det kommissionen att snabbt utarbeta en handlingsplan för initiativet "Utbildning i informationssamhället".	Moreover, it invites the Commission to rapidly work out an Action plan on the initiative "Learning in the Information Society".
Europeiska rådet understryker informationssamhällets möjligheter för **utbildning**, för oganisation av arbete och för skapande av arbetstillfällen.	The European Council underlines the potential of the Information Society for **education** and **training**, for the organization of work and for employment creation.
Det noterade även de viktiga framsteg som har gjorts inom ett antal områden, som kultur och audiovisuella frågor, **utbildning**, hälsa, socialpolitik och miljö.	It also took note of the important progress made in a number of fields such as **culture** and audiovisual matters, **education** and **training**, health, social policy and environment.
– Det kommer att ge privata företag en metod som är inriktad på teknisk överföring så att företagen med förtroende kan delta i skapandet av effektiva språktillämpningar för företagande, **utbildning** och underhållning.	It will provide private enterprise with a focused approach to technology transfer so that firms can participate with confidence in creating effective language applications for business, **education** and entertainment.
Dessa satsningar hör ihop med initiativen inom den audiovisuella sektorn10 och programmen för **utbildning** och fortbildning inom den audiovisuella sektorn.	These efforts are linked to initiatives in the audiovisual sector10 and programmes in relation to **education** and **training**.
Eftersom denna mångfald kommer att fortsättningsvis i hög grad bero på nationella initiativ vad gäller **utbildning** och kultur, kan den även dra väsentlig fördel av de satsningar som görs på bättre resurser till information på olika språk.	While this diversity will continue to depend in large part on national initiatives at the **educational** and cultural levels, it can also benefit considerably from efforts to provide better facilities for access to information in our various languages.
Informationsrevolutionen förändrar radikalt kommunikationens karaktär och användning i hela samhället, vilket leder till nya möjligheter för företagande, kultur och **utbildning**.	The information revolution is radically changing the nature and use of communications throughout society, providing new opportunities for business, **culture** and **education**.
I linje med slutsatserna från G7-mötet om informationssamhället finns behov av ett nyskapande angreppssätt vad gäller tvärkulturell **utbildning** och fortbildning särskilt	In line with the conclusions of the G-7 meeting on the information society, there is a need for an innovative approach to cross-cultural **education** and **training**,

Figure 1: The Netscape view of the database

A B	A ~B	~A B	~A ~B	log λ	trigram
114	2	6	62954	1599.97	små och medelstora
89	31	163	62793	881.92	och medelstora företag
59	4	76	62937	726.15	när det gäller
40	61	1	62974	524.01	informations- och kommunikationsteknologi
40	61	8	62967	490.18	informations- och kommunikationsteknologin
42	0	1343	61691	322.02	inom ramen för
19	0	57	63000	260.61	snart som möjligt
20	0	141	62915	241.42	mellan Europeiska unionen
23	22	80	62951	238.45	forskning och utveckling
12	3	1	63060	207.55	den 1 januari
17	6	59	62994	206.25	gör det möjligt
16	9	57	62994	187.51	på nationell nivå
20	0	582	62474	186.74	i enlighet med
13	3	26	63034	181.57	den inre marknaden
10	1	2	63063	177.47	den tredje etappen
13	10	15	63038	176.50	på detta område
10	7	0	63059	171.95	hälso- och sjukvården
9	0	2	63065	166.96	formerna för arbetets
18	2	431	62625	165.77	i fråga om
11	0	30	63035	164.70	noterar med tillfredsställelse
13	7	42	63014	160.62	tillväxt och sysselsättning
11	0	43	63022	157.80	varor och tjänster
25	0	2678	60373	157.72	av informations- och
21	234	40	62781	154.98	Europeiska rådet uppmanar
11	21	4	63040	153.81	den gemensamma valutan
19	236	23	62798	153.19	Europeiska rådet välkomnar
11	12	9	63044	153.03	I detta sammanhang
11	0	62	63003	150.50	på hög nivå
9	3	3	63061	150.40	1 januari 1999
9	0	11	63056	149.86	för arbetets organisation
10	21	2	63043	145.19	de nya formerna
11	1	62	63002	143.62	på europeisk nivå
8	1	2	63065	143.27	Central- och Östeuropa
19	2	1146	61909	138.85	så snart som
25	9	1844	61198	137.51	det möjligt att
16	1	734	62325	134.58	med hänsyn till
13	6	148	62909	132.64	i Europeiska unionen
11	5	65	62995	129.68	göra det möjligt
16	239	19	62802	129.25	Europeiska rådet noterar
20	0	2683	60373	126.14	för små och
7	2	1	63066	125.92	äldre och handikappade
9	1	41	63025	123.75	att lägga fram
8	1	17	63050	121.94	Den tredje utmaningen
21	359	64	62632	121.56	för att skapa
13	0	589	62474	121.23	i samband med
8	1	19	63048	120.47	experter på hög
9	0	80	62987	119.08	på så sätt
9	17	9	63041	118.90	frågor som rör
9	4	32	63031	118.19	uttrycker sin tillfredsställelse
6	0	1	63069	117.38	Frågor som oroar
7	0	12	63057	116.48	på lång sikt
13	0	737	62326	115.46	anpassa *sig till*
6	0	2	63068	114.13	Frågor att fundera

Figure 2: Swedish trigrams

A B	A ~B	~A B	~A ~B	log λ	trigram
180	39	171	66968	1801.10	The European Council
57	2	27	67272	797.51	the Information Society
86	146	101	67025	754.48	the European Union
51	4	114	67189	602.21	the Member States
47	6	86	67219	566.93	education and training
61	2	1862	65433	418.12	in order to
24	3	3	67328	391.43	Route of Actions
55	177	296	66830	334.77	the European Council
23	18	21	67296	296.06	with a view
29	2	323	67004	292.37	as well as
50	5	2855	64448	282.11	the development of
23	209	2	67124	249.29	the European Parliament
18	2	42	67296	245.88	soon as possible
18	15	18	67307	236.81	the social partners
24	27	141	67166	221.82	the United States
34	0	2871	64453	214.15	in terms of
20	0	332	67006	211.32	as soon as
22	24	129	67183	208.20	research and development
15	36	5	67302	198.01	the United Kingdom
13	2	18	67325	194.43	notes with satisfaction
16	4	81	67257	192.16	the labour market
21	13	277	67047	184.05	to ensure that
10	0	3	67345	182.26	the Structural Funds
22	457	13	66866	172.64	in the field
14	18	23	67303	172.47	In this context
12	0	50	67296	170.26	the Intergovernmental Conference
20	227	22	67089	168.01	European Council welcomes
62	749	533	66014	167.09	of the European
9	0	3	67346	165.07	Central and Eastern
23	0	1900	65435	163.85	a view to
11	0	37	67310	162.16	this Green Paper
14	5	89	67250	161.63	at Community level
16	41	50	67251	158.32	of new technologies
26	193	161	66978	151.22	The European Union
19	15	370	66954	150.30	have to be
9	0	21	67328	141.92	Council took note
10	41	1	67306	139.12	the United Nations
8	3	2	67345	137.71	1 January 1999
9	1	18	67330	137.70	Competitiveness and Employment
9	1	19	67329	136.90	Heads of State
23	1	2882	64452	136.55	the field of
9	7	8	67334	133.13	the single currency
9	0	39	67310	132.24	the White Paper
21	0	2884	64453	132.18	the end of
9	9	7	67333	131.69	Paper on Growth
22	1	2883	64452	130.34	the use of
13	11	124	67210	129.35	growth and employment
15	192	19	67132	128.11	on the basis
24	0	4676	62658	127.91	note of the
11	4	86	67257	127.86	the internal market
19	1	1904	65434	127.43	with regard to

Figure 3: English trigrams

The Political Economy of the Harmonisation of the Nguni and the Sotho Languages

Neville Alexander, *Project for the Study of Alternative Education in South Africa, University of Cape Town, South Africa*

Abstract: The author believes that it is essential to revisit the issue of the harmonisation of the African languages of South Africa. He maintains that most people who have been writing on the subject locally have not understood the kernel of the original Nhlapo-Alexander proposal and restates the economic and political arguments for it. Because there are no "linguistic" barriers to the realisation of this proposal, he concludes that the main obstacle is the lack of political will and appeals to the relevant academics and political/cultural leadership of the country to reconsider the issue against the background of a similar movement in the rest of the continent.

Keywords: STANDARDISATION, HARMONISATION, LANGUAGE PLANNING, STANDARD NGUNI, STANDARD SOTHO, ETHNICITY, NATION-BUILDING, AFRICAN LANGUAGES, AUSBAU LANGUAGES, LANGUAGE ENGINEERING, MUTUAL INTELLIGIBILITY, PAN SOUTH AFRICAN LANGUAGE BOARD, LANGUAGE MODERNISATION

Opsomming: Die polities-ekonomiese aspekte van die harmonisering van die Nguni- en die Sothotale. Die outeur is van mening dat dit noodsaaklik is om weer te kyk na die harmonisering van die Afrikatale van Suid-Afrika. Hy beweer dat die meeste mense wat plaaslik oor hierdie onderwerp geskryf het, nie die kern van die oorspronklike Nhlapo-Alexander-voorstel verstaan het nie, en hy stel weer die ekonomiese en politieke argumente ten gunste daarvan. Aangesien daar geen "taalkundige" grense is vir die daarstelling van hierdie voorstel nie, kom hy tot die gevolgtrekking dat die hoofstruikelblok die gebrek aan politieke wilskrag is. Hy beroep hom op die relevante akademici en politieke/kulturele leiers van die land om hierdie saak te heroorweeg teen die agtergrond van 'n soortgelyke beweging in die res van die kontinent.

Sleutelwoorde: STANDAARDISERING, HARMONISERING, TAALBEPLANNING, STANDAARD NGUNI, STANDAARD SOTHO, ETNISITEIT, NASIEBOU, AFRIKATALE, AUSBAUTALE, TAALMANIPULERING, WEDERSYDSE VERSTAANBAARHEID, PAN-SUID-AFRIKAANSE TAALRAAD, TAALVERNUWING

In view of the erroneous and misleading opinions expressed by Louwrens (1997: 248-250) and because the question of the harmonisation of the Nguni and the Sotho languages of South Africa is bound to become more, not less, relevant in the near future, it is appropriate that this still contentious matter be looked at

once again. [1]

To begin with, let me restate in my own original words what is being proposed. In the first version of the proposal, I wrote:

> The development of a written Standard Nguni and a Standard Sotho, as an initial phase of a very long-term process of "uniformation", need not and will not lead to the disappearance of Zulu, Xhosa, Ndebele, Siswati, Sipedi and Tswana and their dialects ... Indeed, subject to the availability of resources, they will be encouraged in print in literature of all kinds. The main difference will be that in all formal situations, including the crucial area of education, the Standard Nguni or Standard Sotho forms will be promoted. It is to be expected that, over time, the spoken standard — used in formal and relatively formal situations — will begin to approximate to the written standard, even though individuals will inevitably betray their regional or social origins *via* their accent and intonation as they do in all similar situations elsewhere in the world. (Alexander 1989: 64)

Compare this with the aboriginal suggestion of Jacob Nhlapo:

> Even though there are many Bantu languages in South Africa, we can agree that Xhosa, Sotho, Zulu, Tswana and Pedi are the chief ones. They are the ones which are spoken by most Bantu, and most Bantu books are written in them. Let it be said here that these books are bought mostly by school children ... From these tongues we can at first build up two languages. Zulu and Xhosa together with the branches known as Ndebele, Swazi, Baca, etc., are so much alike that, put together they can make one good strong language called Nguni. In the same way, Pedi, Tswana, and Southern Sotho, together with Kxatla, Tlokwa, etc., are so much alike that joined together they can make one good strong language called Sotho. Writing is the best way to make languages grow together ... (Nhlapo 1944)

The essential argument here is based on the mutual intelligibility of the varieties which make up the two language clusters even though one of the main reasons for the proposal is derived from the economic rationality thereof and another from its political appropriateness in the context of the avowed nation-building, antitribalist strategy of the new government. In other words, the original proposal was directed at language specialists and linguists and was intended to remind them that linguistically there was (and is) nothing to prevent the planned convergence towards a written Standard Nguni and a written Standard Sotho. This reminder was all the more necessary as many years of

[1] As background to this contribution, readers are referred to my article (Alexander 1992) in which the rationale for the harmonisation proposal is canvassed in detail.

apartheid-inspired social (and linguistic) engineering had created stereotypes of racial and ethnic separateness which would have to be weakened and even eliminated if the promotion of national unity was to have any chance of success at all. Vested interests in separateness and Ausbau strategies[2] in the domain of language policy and language development would have to be identified and countered if the notion of the two written standards was to make any headway.

The vested interests are obvious. Academics who specialise in particular varieties of the languages concerned as well as traditional leadership have a clear reason for being suspicious and sceptical. However, all of this is misplaced and it is essential that the appropriate and relevant historical and comparative information be placed before these interest groups so that they may realise that far from the proposal for harmonisation being detrimental, tendentially and actually, to their special interests, it will in fact lead to an efflorescence of dialectology and of the study of particular varieties. This is what has happened wherever such planned convergence (standardisation) has been undertaken. And it is so obvious that this must be a spin-off of standardisation initiatives that one wonders why it is necessary to argue the point at all. One reason that is usually implicitly but sometimes explicitly advanced for opposition to the proposal is the reactionary notion of "ethnicity" derived from an outmoded and dated Eurocentric paradigm of identity formation. This, in my view, lies at the heart of Louwrens's unsubtle perception of the language domain and language policy and planning in postapartheid South Africa (see especially Louwrens 1997: 248).

By way of illustrating this proposition as well as that alluded to earlier, that harmonisation will become more, not less, relevant in the near future, it is necessary to point out that within the past 12 months or so, the Pan South African Language Board has received no fewer than five different requests from relatively small groups of people seeking a change of status for their languages. Although the specifics are different in each case, the approaches by groups purporting to represent Indian, Khoi and San, Northern Ndebele, Puthi and Lovedu "linguistic communities" in some cases for recognition as "official languages" forespell the inevitability of harmonisation (or further standardisation) of the Nguni and Sotho clusters. Since it is clear that there can be no recognition of more "languages" — quite apart from the fact that the recognition of the 11 languages was a political decision based on a compromise and had nothing to

[2] Ausbau languages are those that are so similar in grammar and lexicon to other, stronger, previously recognised languages that their language authorities often attempt to maximize the differences between themselves and their Big Brothers by multiplying or magnifying them through adopting or creating distinctive paradigms for neologisms, word order and grammar, particularly in their written forms. Thus Ausbau languages are languages by effort, i.e. they are consciously built away ("ausgebaut") from other, more powerful and basically similar languages so as not to be considered mere dialects of the latter, but rather, to be viewed as obviously distinctive languages in their own right ... (Fishman 1974). Also see Msimang 1996, where evidence for the adoption of this strategy by the apartheid social engineers is given very clearly.

do with whether the 11 were or are "languages" as opposed to "dialects" — I believe that, increasingly, political and academic planners are going to have to come back to the harmonisation proposal. For this reason, among others, it is essential that the political and economic rationality of going down the road of harmonisation be demonstrated. Paradoxically, the danger of opening the Pandora's box of tribal and ethnic strife will impell those who fear the consequences of such destabilisation to turn towards harmonisation as the way to resolve the dilemma of recognising some "dialects" as official languages but not others.

In order to clear up another confusion that has crept into this debate, it is necessary to note that those of us who have consistently propagated harmonisation or unification of the "mutually intelligible" varieties of the two language clusters have never suggested that this should be a sudden, short-term, cut-and-paste operation undertaken by some faceless language engineers in smoke-filled backrooms. This image of a sinister strategy for Jacobinic homogenisation of the unsuspecting "people" is a caricature which it is all too easy to pillory and even to kill off. Fortunately, those who have actually read (and understood) the proposal realise that it is based on a wealth of comparative, historical and linguistic good sense and that it will not disappear so easily. Some suggestions about how the process of harmonisation might be initiated and sustained (see, for example, Glaughton and Gough 1996, Msimang 1996 and Cluver 1990) are eminently discussable and will become respectable as soon as the requisite sociolinguistic and political atmosphere comes into being. To this, one should add the not inconsiderable fact that the actual practice of SABC TV, which uses all the Nguni and the Sotho varieties on a specific channel as a matter of economic necessity, besides other reasons, has already taken the process a long way forward even though few people would recognise or deliberately describe this practice in terms of harmonisation.

Two years ago, when Professor Kwesi Prah of the University of the Western Cape organised a highly significant seminar in Cape Town on harmonisation and standardisation of African languages, numerous African scholars from countries as far apart as Ghana, Cameroon and Lesotho and South Africa demonstrated one after the other that there is no linguistic-technical reason why these processes cannot be promoted and that there is every political and economic reason why they should be promoted. Citation needs only be made from two of the most relevant papers delivered at that seminar. Thus, Professor Emenanjo of the National Institute for Nigerian Languages, in an erudite and elegant essay on Modern Standard Igbo showed how this language differs from previous artificial academic attempts to "harmonise" different varieties of the Igbo continuum.

> Unlike the three extinct "standards" which were artificially created to fill a vacuum, the extant standard is the product of the dynamic forces of inclusion and exclusion necessitated by the imperatives of modernisation, and engineered by the dialect-neutral Society for Promoting Igbo

> Language and Culture through its language think-tank: the Igbo Standardization Committee. Motivated and propelled by its own internal logic, with its own verifiable and quantifiable rules which include eclecticism in the choice and use of lexical items, Modern Standard Igbo is distinct from any and all of the live Igbo dialects, and any and all of the "dead" Igbo standards in its pan-Igbo acceptability and patronage as well as its valency in metalanguage. Standard Igbo has the richest lexical inventory among all other varieties of Igbo. (Emenanjo 1996: 3)

In a short contribution comparing the modernisation of Japanese and Kiswahili, Professor Miyamoto of Osaka University noted:

> (Just) as Shona in Zimbabwe is sometimes said to be "a language which everyone writes and nobody speaks", so it was with Standard Swahili at the beginning. But today the written form of Swahili has strongly influenced the spoken form, as publications increase and the number of speakers of Swahili as a second language increases. The meaning of modernisation of any language is not always clear but some radical planning seems necessary, especially in a multilingual situation or in ... critical ... (periods) of history ... The ... (most important) factor ... (for) the success of any language planning seems ... (to be the promise) that it will create larger and better job-markets for the common masses who will speak and write ... (the language concerned). (Miyamoto 1996: 10-11)

It is these political and economic moments of the argument that will in future have to become the focus of our research if we are to pursue the matter of harmonisation as a practical undertaking during the next few years.

In a multilingual ecology where English is King of the Languages, it makes eminent sense to ensure that for the sake of the vast majority of the population for whom English will always be a foreign language, there are strong African languages with a solid infrastructure of literacy, interpreting and translation competence. In South Africa, as in many other African countries, such an infrastructure will be created much more easily and effectively through the harmonisation of existing varieties of indigenous language clusters wherever and whenever this is possible and feasible. The essential next step has to be initiated by the political and cultural leadership in and across the relevant African countries. Men and women who have the vision and the courage to be "unpopular" for a while because of the unavoidable objections that will be forthcoming from those with vested interests and those who are restricted by the myopia of hegemonic agendas simply have to realise that they have to publish the information and produce the prototypical examples of texts that will persuade the "masses" of the feasibility of this historic undertaking. And, they have to realise the Biblical prophesy: "Where there is no vision, the people perish!"

For it remains a stubborn fact that the users of the varieties of a language are in the final analysis the people who decide whether theirs is a different

guage" from the variety spoken in the neighbouring village. It is, however, the prerogative of government to intervene and to shape people's consciousness differently, provided this happens in a transparent and democratic manner. If this were not the case, what right would the government of this, or any other, country have to persuade men and women to wear condoms in order to lessen the danger of spreading Aids? Politically speaking, we have to persuade people who speak a Nguni or a Sotho variety that these are in fact "dialects of each other", i.e., Zulu is a dialect of Xhosa and vice versa, for example. The implicit political agenda in such a statement ought not to be problematical to anyone committed to the democratic transition in South Africa. The kind of objection or fear raised by Louwrens (1997) is, as I have indicated, either based on a misconception of the dynamics of harmonisation as a process or it derives from a dated paradigm.

We must, it is clear, bear in mind that the strength of tradition, especially where the languages have been written for more than a century and have relatively strong literary treasuries will represent a retarding moment. This is, however, not insuperable at all as recent examples in Africa itself and in countries such as Estonia demonstrate.

Practically, for the next decade or so, this implies that a text that is composed predominantly in Xhosa but which includes many lexical, syntactical and morphological elements from one or more of the other varieties of Nguni will be labelled a "Nguni" text; similarly, if it were to be composed predominantly in Zulu or Swati or Ndebele. The necessary condition for this to happen is that comprehensive dictionaries embracing all the relevant varieties would have to become readily available in different formats that can be used for different purposes. Texts would also necessarily have glossaries or notes by means of which peculiarities deriving from one or other variety would be explicated. In an earlier period, many Dutch texts carried such glossaries in Afrikaans and it is a practice that is widespread in similar situations elsewhere. It ought to be obvious that such an approach is economically rational since the need to translate and to print different (more costly per unit) texts falls away.

The other related but enormous task that awaits us in this undertaking is the revival or the establishment of a reading culture in the African languages. This is a matter the need for which is so well attested as to require no further substantiation. What is clear is that the educational structures, especially those devoted to early childhood development, the preprimary and elementary schools, will play the major role in this regard. Besides the need to conduct radically new research into the acquisition of literacy by young children and to retrain teachers throughout the continent along appropriate lines, it is clear that one of the great virtually unthought-of tasks is the translation into the African languages for use at all levels of sophistication of the great works of world — including African — literature, science and philosophy.

Sociolinguists, lexicographers, terminologists, terminographers, translators, interpreters and other language and educational specialists, together with enlightened political leadership will have to concert their efforts in order to cre-

ate the framework within which the harmonisation of the Nguni and Sotho language varieties respectively can take place. Any African "renaissance" will be stillborn unless this process is initiated.

References

Alexander, N. 1989. *Language Policy and National Unity in South Africa/Azania.* Cape Town: Buchu Books.

Alexander, N. 1992. South Africa: Harmonising Nguni and Sotho. Crawhall, N. (Ed). 1992. *Democratically Speaking: International Perspectives on Language Planning.* Cape Town: National Language Project.

Cluver, A. (Ed). 1990. Taalbeplanning in Suid-Afrika. Lecture No. 6 of Study Guide, *Linguistiek.* Studiegids 2 vir Lng 100-105. Pretoria: University of South Africa.

Emenanjo, E. 1996. The Modernization of the Igbo Language and its Implications for Holistic Education. Paper delivered at the Colloquium on Harmonising and Standardising African Languages for Education and Development, University of Cape Town, 11-14 July 1996. *Unpublished mimeo.*

Fishman, J. (Ed). 1974. *Advances in Language Planning.* The Hague: Mouton.

Glaughton, J. and D. Gough. 1996. Standard Nguni: An Alternative Proposal. Paper delivered at the Colloquium on Harmonising and Standardising African Languages for Education and Development, University of Cape Town, 11-14 July 1996. Unpublished mimeo.

Louwrens, L. 1997. On the Development of Scientific Terminology in African Languages: The Terminographer's Dilemma in a New Dispensation. *Lexikos* 7: 245-251.

Miyamoto, M. 1996. The Modernization of the Japanese Language in Comparison to Swahili. Paper delivered at the Colloquium on Harmonising and Standardising African Languages for Education and Development, University of Cape Town, 11-14 July 1996. Unpublished mimeo.

Msimang, T. 1996. The Nature and History of Harmonisation of South African Languages. Paper delivered at the Colloquium on Harmonising and Standardising African Languages for Education and Development, University of Cape Town, 11-14 July 1996. Unpublished mimeo.

Nhlapo, J. 1944. *Bantu Babel: Will the Bantu Languages Live?* Cape Town: The African Bookman.

Lexicographic Training at the Bureau of the Woordeboek van die Afrikaanse Taal

W.F. Botha and E. Botha,
Bureau of the Woordeboek van die Afrikaanse Taal,
Stellenbosch, South Africa

Abstract: Since 1995, the Bureau of the WAT has developed several training courses for students and other persons interested in gaining insight into and practical experience of the planning, compilation and management of a dictionary. This article summarizes the courses offered.

Keywords: GENERAL LEXICOGRAPHY, COMPUTER LEXICOGRAPHY, CO-OPERATIVE TRAINING, DICTIONARY TYPOLOGY, DICTIONARY PLANNING, DICTIONARY COMPILATION, DICTIONARY MANAGEMENT

Opsomming: Leksikografiese opleiding by die Buro van die Woordeboek van die Afrikaanse Taal. Die Buro van die WAT het sedert 1995 verskeie kursusse ontwikkel vir studente en ander persone wat daarin geïnteresseerd is om insig en praktiese ervaring te verkry in die beplanning, samestelling en bestuur van 'n woordeboek. Hierdie artikel gee 'n opsomming van die kursusse wat aangebied word.

Sleutelwoorde: ALGEMENE LEKSIKOGRAFIE, REKENAARLEKSIKOGRAFIE, KOÖPERATIEWE OPLEIDING, WOORDEBOEKTIPOLOGIE, WOORDEBOEKBEPLANNING, WOORDEBOEKSAMESTELLING, WOORDEBOEKBESTUUR

1. Introduction

One of the consequences of the strategic planning of the Bureau of the WAT was the development of an in-service training course. The course, which was aimed at the rapid development of the lexicographical skills of new staff members, proved to be very successful. As the idea of co-operative lexicography gained momentum at the Bureau, a training course for both practising lexicographers and persons interested in lexicography from outside the Bureau was designed in 1995. This led to the development of other training courses. The Bureau currently offers the following:

— a ten-day training course in general and computer lexicography and in the planning and management of a lexicographic project, offered during March and September,
— a co-operative training course for language practitioner students from technikons,

- a co-operative training course for students in lexicography and for lecturers who wish to gain practical experience of dictionary-making,
- lectures dealing mainly with dictionary typology and the dictionary-making process.

2. Course in general and computer lexicography, and in the planning and management of a lexicographic project

2.1 Background

This course was first offered in 1995, and since then annually during March and September. Practising and prospective lexicographers, students and lecturers have attended the course. Attendants have come from almost all the provinces of South Africa, and also from as far as Namibia, Angola, Gabon, Zambia and Tanzania. Among the participants have been members of the isiXhosa, the isiZulu and the Sepedi Dictionary Projects, lecturers and students of the Soweto Campus of Vista University, the University of the North, the University of Venda, the University of Durban-Westville, the Technikon Northern Gauteng and members of the Language Services of the Northern Province. In 1997 two of the Bureau's staff members also offered a course at the Centre International des Civilisations Bantu (Ciciba) in Libreville, Gabon. The attendants represented several African countries.

The different components of the course are offered by staff members of the Bureau who specialize in a particular field of the dictionary-making process, such as general lexicography, computer lexicography, planning and management, and typesetting. It covers the theoretical and practical aspects of dictionary making and a wide range of topics in computer lexicography relevant to editorial and management staff of lexicographic projects. Attendants participate by identifying and examining different facets of the planning and management of a dictionary project. They are expected to do some reading on the topics treated, join group discussions and do practical exercises.

The course is structured in such a way that simple topics are introduced before more advanced ones. Consequently, there is enough variety and stimulation to keep participants interested and actively involved.

Since feedback from the participants is very important to the Bureau, evaluation forms are completed at the end of the course. This feedback has contributed to the evolution of the course over the past four years, and has for instance led to more time being allocated to the component dealing with computer lexicography.

A certificate of attendance detailing the topics covered in the training is presented to each participant upon completion of the course.

2.2 General lexicography

The first component of the course covers many aspects of the theory of dictionary making. However, the aim of the course is the practical implementation of the theory of lexicography. Consequently much time is devoted to practical lexicography. The focus of the course is on descriptive dictionaries.

All participants have access to computers during the practical side of this component. A series of exercises guide them towards the compilation of dictionary articles using a template of tags. Participants with little or no computer experience receive personal attention from the presenters and progress at their own pace.

The course is well-documented and notes are made available to the participants during the course. The medium of instruction is English or Afrikaans, or both of these languages.

The needs of each group of participants are taken into consideration during the presentation of the course, thus the course has a dynamic character. Since the number of participants is usually limited to no more than six people per session, there is an opportunity for interaction between trainer and participant.

The following themes are covered in the component on general lexicography:

- **Introduction**
  - target user
  - dictionary typology
  - data collection
  - macro- and microstructure
  - types of information in a dictionary
  - diachrony and synchrony
  - lexicography, lexicology and metalexicography
- **Criteria for inclusion**
- **Different kinds of lemmas as dictionary entries**
- **Labelling**
- **Dealing with meaning**
  - polysemic versus homonymic lemmas
  - arrangement of polysemic distinctions
  - different kinds of definitions
  - general principles of definition
  - dealing with insulting and sensitive lexical items
- **Grammatical information and its presentation**
- **Pronunciation**
- **Editing**

2.3 Computer lexicography

The contents of the computer lexicography component is changed annually to keep abreast of developments in computer technology, software and computer lexicography.

This component primarily covers matters of relevance to editorial and management staff of lexicographic projects in a general and nonspecialist manner. Issues such as the benefits of computerization, networks, hardware and software considerations, training and support are covered. The Bureau's system serves to illustrate some of these topics, and possible improvements or alternatives to the Bureau's system are also discussed.

Particular attention is given to language material collection and editorial processing. Some basic principles of database and corpus design and their use are discussed. The making of structured manuscript in a database and tagged text environment is also dealt with.

Participants have the opportunity to gain hands-on experience in manuscript making in a tagged text dictionary-making system. Electronic dictionaries on CD-ROM and the Internet are also demonstrated, as well as corpus-processing and concordancing tools.

A handbook containing exercises and a reading list is provided. Sample software and further reading matter are also provided, and participants can suggest how this component may be improved.

The Bureau has also developed prototype editorial manuscript-making systems for dictionary projects, together with manuals.

Below is a more detailed list of the topics covered:

- **Language material collection policy and techniques**
- **Compilation of a database and corpus**
- **Systems analysis and design**
- **Structured text and SGML**
- **Desktop publishing**
- **Electronic resources**

2.4 Planning and management of a lexicographic project

The component dealing with the planning and management of a lexicographic project is not only intended for managers, but for anybody interested in management. The particular needs of the trainees are taken into consideration.

Additional reading is expected from participants in order to stimulate discussion and to enable them to complete certain tasks.

Upon completion of this component, participants are able to formulate a goal, a mission, objectives, and medium- and long-term objectives for a dictionary project.

The main themes covered in this component of the course are:

— **Mission of the project**
— **Historical overview**
— **Strategic areas of focus**
— **Analysis of the environment**
  restrictive factors in the internal/external environment
  supportive factors in the internal/external environment
— **Scenario**
— **Assumptions regarding planning**
— **Strategic policy guidelines**
— **Objectives**
  ultimate goal
  long-term objectives
  medium-term objectives
  goals
— **Plan of action**
— **Values applicable to a lexicographic project**

3. Co-operative training of trainee language practitioners

The Bureau of the WAT offers annual co-operative training courses to trainee language practitioner students as part of their practical training.

The students' existing skills and knowledge are utilized and developed over a period of ten weeks. The training is of a practical nature, and any special language skills are utilized and developed.

The students are introduced to many facets of dictionary-making, including language material collection, the verification thereof for editorial use, corpus building, lexicography, the making of dictionary manuscript, desktop publishing and the utilization of computer programs to create word-lists and to do word-frequency counts.

Students are also assisted in developing their computer skills and in using the library, on-line library catalogues, the Internet, CD-ROMs and databases for language research and in dealing with language queries.

A certificate of attendance is presented to each student upon the completion of the course, detailing his/her activities.

This co-operative training course is possible as a result of an agreement between the Bureau and the tertiary institutions to which the students are affiliated.

4. Co-operative training of students in lexicography

Senior or postgraduate students in lexicography — both locally and from abroad — who wish to gain practical experience of dictionary-making, can be

accommodated at the Bureau. They receive training according to their needs in selected components of the training courses discussed above. During the final stage the students have the opportunity to compile dictionary articles on computer, utilizing the data of the Bureau.

As with the training of language practitioner students, the training course is the result of an agreement between the Bureau and the universities to which the students are affiliated. The duration of the training is between four and ten weeks.

5. Lectures dealing mainly with dictionary typology and the dictionary-making process

Lectures dealing mainly with various traditional types of dictionaries and the dictionary-making process are presented on request. Groups of scholars, students and other interested persons attend these lectures. About six lectures are presented annually.

6. Course fees and accommodation

Course fees are determined according to the circumstances of the participants. Affordable accommodation can be arranged for participants. Details on exactly how the course fees are determined can be obtained from the Bureau.

7. Information on the courses

For further information on the training courses offered at the Bureau, Dr D.J. van Schalkwyk, Editor-in-chief can be contacted. The Bureau's particulars are:

Postal address: P.O. Box 245, Stellenbosch 7599
Telephone: (021) 887 3113
Fax: (021) 808 4336
E-mail: wat@maties.sun.ac.za
Website: http://www.sun.ac.za/wat/index.htm

Report on the SALEX '97 Lexicographical Training Course, Grahamstown, 15-27 September 1997

Penny Silva, *Dictionary Unit for South African English, Grahamstown, South Africa*

Abstract: The report describes the background to the SALEX '97 Lexicographical Training Course, and the reasons for its conception. It explains the constraints within which the course had to be designed, and lists its practical and theoretical aims. The financing of the course, the range of participants attending, the structure of the working day, and the course materials provided to participants are described. The report ends with excerpts from evaluations provided by participants, and with a reference to the second planned training course, SALEX '98. The simple initial framework upon which the course was based is provided as an Appendix.

Keywords: SOUTH AFRICA, LEXICOGRAPHY, TRAINING, SALEX, CORPUS, AFRICAN LANGUAGES, LANGUAGE-INDEPENDENT, PROJECT PLANNING, DICTIONARY COMPILATION

Opsomming: Verslag oor die SALEX '97 leksikografiese opleidingskursus, *Grahamstad, 15-27 September 1997.* In hierdie verslag word die agtergrond waarteen die SALEX '97 leksikografiese opleidingskursus plaasgevind het, geskets en daar word redes aangevoer vir die ontstaan daarvan. Die beperkings waarbinne die kursus ontwerp moes word, word beskryf en die praktiese en teoretiese doelwitte van die kursus word gelys. Die finansiering van die kursus, die verskeidenheid kursusgangers, die uiteensetting van die werksdag en die kursusmateriaal wat aan die kursusgangers verskaf is, word uiteengesit. Hierdie verslag eindig met aanhalings uit kursusgangers se evaluerings van die kursus, asook met 'n verwysing na die tweede beplande kursus, SALEX '98. Die eenvoudige aanvanklike raamwerk waarop die kursus gebaseer is, word as 'n Bylae weergegee.

Sleutelwoorde: SUID-AFRIKA, LEKSIKOGRAFIE, OPLEIDING, SALEX, KORPUS, AFRIKATALE, TAALONAFHANKLIK, PROJEKBEPLANNING, WOORDEBOEKSAMESTELLING

1. Background

The impetus for the SALEX '97 Lexicographical Training Course came from the South African postapartheid constitution, under which nine African languages were added to English and Afrikaans as official languages. State-funded units for the compilation of dictionaries exist at present for only English and Afrikaans, but there are moves afoot to right this imbalance. The National Dictionary Units Bill has been under consideration for over a year, and it was hoped

that several Units might have been established by the time the course was presented. However, establishment was unfortunately delayed because the Bill was withdrawn for reconsideration.

The success of the new Units, when they are established, will depend upon the development of a cadre of lexicographers with the practical and intellectual skills to tackle every aspect of running a dictionary project. Conceived in order to help prepare lexicographers, particularly in the African languages, for this task, SALEX '97 ("South African Lexicography") was organized by Penny Silva, Director of the Dictionary Unit for South African English at Rhodes University, under the auspices of the Department of Arts, Culture, Science and Technology, and of the African Association for Lexicography (AFRILEX).

The course was led by three eminent British lexicographers and teachers — Sue Atkins (course leader), Michael Rundell, and Edmund Weiner. Their enthusiasm for contributing to the training of lexicographers for the new South African society, their extensive theoretical knowledge, and their collective experience at the practical end of lexicography made them ideal leaders of the training course. Their expertise covers bilingual, pedagogical, and scholarly dictionaries, and includes corpus development and the design of systems for corpus querying, and for the electronic compilation and manipulation of dictionary text, as well as the training of lexicographers in both research and commercial institutions.

SALEX '97 took place in the St. Peter's Building, Rhodes University, Grahamstown, during the second half of September, 1997.

2. Financing the Course

The course received substantial support from donors, making the project possible, and helping to keep the cost to the participants within manageable limits. The two major donors were the British Council and the national Department of Arts, Culture, Science and Technology, and generous funding was also provided by the Anglo American and De Beers Chairman's Fund, and First National Bank. Sets of forty learners' dictionaries were donated by Addison Wesley Longman, Cambridge University Press, and HarperCollins (all in the United Kingdom), and Oxford University Press (Southern Africa). Xeratech provided free photocopying of much of the course material. Expenses were considerable, but were partly offset by the R500 registration fee paid by most of the participants. Several participants requested assistance: in these cases the registration fee was waived, and for two participants additional financial support was provided (accommodation and/or travel costs being covered).

3. The Participants

The thirty-five SALEX '97 participants were representative of all of the eleven official languages with the exception of Xitsonga. Some languages (e.g. Sepedi,

isiZulu, isiXhosa, Afrikaans, and English) were well-represented, with three or more participants each. The participants were drawn from universities, provincial and national language and terminology departments, dictionary units, publishers, and the international Summer Institute of Linguistics. Two participants were freelancers. In all, fifteen languages were represented. In addition to the South African participants, there were several lexicographers working in other African languages — Chiluba (Zaire), Mwani (Mozambique), Kiswahili (Tanzania), and Amharic/Silte (Ethiopia) — as well as participants from neighbouring African countries (Lesotho and Swaziland). Two participants came from Europe.

Mr John Orr of the SABC attended part of the course as an observer, and recorded interviews for a series of three programmes for SAfm's *Word of Mouth* language programme, broadcast during October 1997.

4. The Course: Theory and Preparation

In September 1996, the organizer sent a list of the various components of dictionary compilation to Sue Atkins for her consideration. The final version of this list (see Appendix) formed the framework upon which the course was based. The course aimed to enable participants to

— plan a dictionary project;
— design a dictionary appropriate to the needs of their language community;
— build and train a team of lexicographers; and
— edit their dictionary to publication stage.

The challenge while planning and designing the course was to make sure that it was

— a general, broad, basic training in dictionary-making, including project-planning;
— situated firmly within the empirical and descriptive tradition;
— language-independent, applicable to many different languages;
— English-medium, using English examples, but at a level which would be accessible to all;
— of a content level which would suit participants with differing levels of experience;
— practically oriented, not simply theoretical; and
— exhaustively documented, providing participants with a comprehensive record of proceedings for subsequent use.

The result was a course which is believed to be a first in world lexicography — an intensive methodological foundation-course which is language-independent. The course had four main strands:

- the practical business of planning a dictionary project and taking the editorial process through to completion;
- the metalexicographic concepts needed for discussing the structure and editorial content of dictionaries (macrostructure and microstructure);
- the concepts from linguistic theory that can usefully inform and underpin aspects of the lexicographer's work, particularly when analysing the data and drafting the entry; and dictionary entry writing.

The organiser and three trainers met in Lewes, East Sussex, in February 1997 for a two-day planning meeting, during which the South African language and lexicographical background was discussed, the course outline designed, the programme plotted out, and the lecturing tasks allocated. For the next six months, extensive four-way discussion took place on email, and electronic files containing slide-presentations and worksheets were sent to Grahamstown as attachments to email letters. Designing and planning such a course jointly without electronic communication would not have been possible.

5. The Course: Organization

SALEX '97 ran for ten working days, from 9h00 until 16h30, with an hour for lunch and two 20-minute tea-breaks daily. The working language was English. Each half-day module consisted of a lecture (with overhead slides and identical handout material) and a practical workshop (with worksheets). There were also plenary sessions, including demonstrations of corpus software.

Groups were initially mixed, in order to ensure that the participants talked across linguistic boundaries, but participants subsequently worked mainly in four language-groups: English, Afrikaans, Nguni, and Sotho. The trainers moved from one group to the next, offering assistance and comment. Participants worked extremely hard, and with great enthusiasm and dedication — something the trainers all remarked upon.

The course took participants from the first stages of project planning and the identification of their target user, through the collection and analysis of data, and the compilation of entries, to choices in publishing their dictionaries. Most of the time was spent on analysis of data (using a large computer-based corpus of English, loaned by Addison Wesley Longman UK) and on entry-compilation for a general user's dictionary, but there was time given also to specialized skills such as etymology and pronunciation.

During the lectures, the participants were able to follow the overhead slides on an identical printed version, and could thus add their own notes as the course progressed. At the end of the course, each participant possessed a large file of material which could be used for training their colleagues — a file containing close to 800 overheads, dozens of worksheets, a bibliography, a 22-page glossary of terms used during SALEX '97, and a list of contact names and addresses of all attending the course.

6. Evaluation

SALEX '97 exceeded the organizer's expectations in the quality of its content and teaching, the response of the participants, and the spirit of common purpose which developed. The degree of commitment from both trainers and participants was impressive.

The course was highly intensive for trainers and participants alike (and particularly for participants for whom English was not first language), leading the trainers to question whether any modules could have been dropped. In the end, consensus was that little if any of the course should have been omitted, given that participants had extensive documentation to take away with them. With this documentation, participants should be able to recap on the course, run local training programmes, and spend more time on workshop exercises, as required.

The following responses are excerpts from evaluation-forms completed by participants after SALEX '97:

> The theoretical background gained from the course will enhance the quality of our dictionary's results and output. A follow-up course is a dire necessity. SALEX should investigate the possibility of formulating a structured course in lexicography registered at any South African university for purposes of improving the lexicographer's qualifications.
>
> Since I already have a finished manuscript, I am now able to revisit it methodically and with confidence. My experience before and after this conference should make me a better trainer of young lexicographers. Not only did we learn about dictionary-making as such, our lecturers virtually took us on a guided tour around our own brains to help us observe just how language is computed. It was an unforgettable experience.
>
> More than ever before I am now aware of the complexities of dictionary compilation, as well as the thrill of actually realizing that I am (and have been) making progress. The sooner we have another course concentrating on the lexicographical problems peculiar to the South African indigenous languages, the better.
>
> SALEX was an historic event ... All role-players or potential role-players in the lexicography industry were able to learn for two weeks from the knowledge of the presenters. They were also able to learn from one another, discovering new points of contact between the various languages, etc. Presenting a training course for such a divergent group makes great demands on the presenters. They succeeded to a high degree in being "everything for everyone" ... A follow-up course is imperative, because a process was initiated which should not be allowed to die. Certain needs of the participants were identified, and it is imperative

that attention be given to these, in order to get the dictionary-making industry in African languages established effectively.

Reservations expressed by some participants were

- that there was too much covered in the time available;
- that some course modules were not directly relevant to the types of dictionaries being planned by individuals or groups; and
- that bilingual dictionary-making should have been covered.

There was a strong conviction expressed that further linked training courses were essential.

7. The Future

SALEX '97 was the first of three practical courses planned by AFRILEX for South African lexicographers — the initial grounding in general methodology and practice. The second course, SALEX '98, will apply the SALEX '97 principles to bilingual dictionary-making, and will consider the specific problems of the African languages (two areas of great concern which arose out of SALEX '97). The course will be organized by Professor Daan Prinsloo at the University of Pretoria in September 1998.

Further information can be obtained from:

Daan Prinsloo: PRINSLOO@libarts.up.ac.za
Penny Silva: P.Silva@ru.ac.za
Website: URL: http://www.ru.ac.za/affiliates/dsae/salex97

APPENDIX

The components of dictionary production

A. COLLECTING DATA
   1. *Choosing data: corpus design*
      a. Types of data
      b. Types of corpus
   2. *Gathering and storing data: corpus maintenance*
      a. Print corpus
      b. Electronic corpus
   3. *Accessing the data*

B. PLANNING A DICTIONARY
   1. *What kind of dictionary?* (i.e. dictionary macrostructure)
   2. *Dictionary type?* (i.e. who is the user?)
   3. *The word-list*
   4. *The entries*
   5. *Supporting material*

C. WRITING A DICTIONARY
   1. *Analysing the data*
   2. *Compiling an entry*

D. TEXT PRODUCTION

Paradigmaverskuiwings en die Afrikaanse mediese vaktaal*

H.P. Wassermann, *Voormalige Dekaan, Fakulteit Geneeskunde, Universiteit van Stellenbosch, Suid-Afrika*

Abstract: Paradigm Shifts and the Afrikaans Medical Terminology. The *Paramediese Woordeboek,* the first Afrikaans medical dictionary to appear in an age of international shifting paradigms in medicine, is reviewed against that background. An historical perspective on the influence of such shifts and their aims is presented, emphasising their effect on intercollegiate, communal and transactional communication. The enigma of the title concerns the intended target group, and inferred distinct technical terminology. The preface, with its claim to simplicity of definitions, and alleged problems with internationally recognised Afrikaans medical language is discussed. Constructional user-unfriendly aspects and incorrect definitions are pointed out, with a note on Afrikaans spelling. In spite of its title, it is largely a school dictionary.

A perspective on the future of medical dictionaries and paradigm shifts is presented; paradigm shifts probably have less terminological than strategic implications. The latter involves reaching and providing for the needs of disadvantaged target groups in accessing a highly specialised intercollegiate technical language.

A need for updating and revising the existing standard Afrikaans explanatory medical dictionary is emphasised; from such a revised edition user-defined explanatory wordlists could be abstracted by a panel of professionals from the intended target groups in accessing a highly specialised intercollegiate technical language. A challenging task awaits Afrikaans medical dictionaries.

Keywords: MEDICAL DICTIONARY, PARAMEDICAL, PARADIGM SHIFTS, HISTORIC PERSPECTIVE, FUTURE PERSPECTIVE, ENIGMATIC TITLE, TARGET GROUPS, CONSTRUCTION, PREFACE, PRESCRIPTIVE DEFINITIONS, SPELLING, INTERCOLLEGIATE COMMUNICATION, COMMUNAL COMMUNICATION, TRANSACTIONAL COMMUNICATION, MULTILANGUAGE DICTIONARY

Opsomming: Die *Paramediese Woordeboek,* die eerste Afrikaanse mediese woordeboek om in 'n era van internasionale paradigmaverskuiwings in geneeskunde te verskyn, word teen daardie agtergrond beskou. 'n Historiese perspektief op die invloed van sulke verskuiwings en hul oogmerke word gegee, met klem op interkollegiale, kommunale en transaksionele kommunikasie. Die enigma van die titel gaan oor die bedoelde teikengroep, en 'n vermoedelik afsonderlike tegniese terminologie. Die voorwoord, met sy aanspraak op eenvoudige definisies, en beweerde probleme met die internasionaal erkende Afrikaanse mediese taal word bespreek. Konstruksionele gebrui-

* Resensieartikel oor die *Paramediese Woordeboek* deur Lynette van Rensburg, 1996, 289 pp., ISBN 0-7986-3568-1, uitgegee deur Kagiso Uitgewers, Pretoria.

kersonvriendelike aspekte en foutiewe definisies word aangetoon, met 'n opmerking oor die Afrikaanse spelwyse. Ten spyte van sy titel is dit hoofsaaklik 'n skoolwoordeboek.

'n Perspektief op die toekoms van mediese woordeboeke word gegee; paradigmaverskuiwings het waarskynlik minder terminologiese as strategiese implikasies. Laasgenoemde betrek die bereiking van, en voorsiening in die behoeftes van agtergeblewe teikengroepe ten einde toegang te verkry tot 'n hoogs gespesialiseerde interkollegiale tegniese taal.

Die behoefte aan bywerking van die bestaande standaard Afrikaanse verklarende woordeboek word beklemtoon; uit so 'n hersiene uitgawe kan gebruiker-gedefinieerde verklarende woordelyste ekstraheer word deur 'n paneel vakmanne uit die bedoelde teikengroep. 'n Uitdagende taak wag op Afrikaanse mediese woordeboeke.

Sleutelwoorde: MEDIESE WOORDEBOEK, PARAMEDIES, PARADIGMAVERSKUIWINGS, HISTORIESE PERSPEKTIEF, TOEKOMSPERSPEKTIEF, ENIGMATIESE TITEL, TEIKENGROEPE, KONSTRUKSIE, VOORWOORD, VOORSKRIFTELIKE DEFINISIES, SPELWYSE, KOLLEGIALE KOMMUNIKASIE, KOMMUNALE KOMMUNIKASIE, TRANSAKSIONELE KOMMUNIKASIE, VEELTALIGE WOORDEBOEK

1. Historiese perspektief op mediese woordeboeke en paradigmaverskuiwings

Die *Paramediese Woordeboek* (PW), die eerste Afrikaanse mediese woordeboek om tydens die huidige wêreldwye paradigmaverskuiwings in geneeskunde te verskyn, moet teen hierdie agtergrond gesien en beoordeel word.

Mediese terminologie ontwikkel uit bewoordbare begrippe omtrent siektes en ongesteldhede (bv. *koue vat, maansiek, duiwelbesete*) — 'n voortgaande proses binne die omgangstaal. Die wetenskaplike benadering dateer vanaf Hippokrates (circa 480 v.C.), die Griekse "vader" daarvan. In antieke Rome het Griekse geneeshere, soos Galenus (130-200 n.C.) by uitstek, geneeskunde beoefen, in Latyn baie en oor alle mediese sake geskryf, en óók die eerste woordelys vir mediese studente saamgestel (MacNalty 1965[3]: v). Sy deurlopende invloed op die geneeskunde was nog merkbaar tot die laat- agtiende eeu. Vir Marcus Aurelius was hy "die eerste onder dokters en die voorste onder filosowe", vir die Middeleeue was hy "die mediese pous", en Renaissanceanatome en -fisioloë het hom as "mentor" erken (Magner 1992: 86 e.v.). Die Grieks-Latynse oorsprong van mediese vakterme in alle wetenskapstale verklaar waarom mediese woordeboeke hul besonderlik tot vertaling in enige ander wetenskapstaal leen.[1] Van Suid-Afrika se amptelike tale is slegs Engels en die inheemse Afrikaans nog wetenskapstale.

Gesondheidsdienste ontwikkel in die twintigste eeu deur drie fases: Vóór die tegnologiese ontploffing ná die Tweede Wêreldoorlog is optimale mannekragvoorsiening met klem op voorgraadse mediese onderrig nagestreef, vanaf die vyftigerjare volg spesialisasie en navorsing om diensgehalte te ontwikkel en uit te bou, en sedert die tagtigerjare verskuif die klem na gelykberegtiging op beskikbare dienste deur dit op bekostigbare wyse toeganklik vir almal te pro-

beer maak. Elke fase het terminologiese implikasies: Die eerste moes die wetenskaplike benadering tot siekte en gesondheid verstaanbaar aan pasiënte en belangstellendes oordra, die tweede moes nuwe teoretiese begrippe en tegniese terme en medisyne- en apparatuurname skep, en in die laaste fase moet medies-ekonomiese en medies-sosiologiese begrippe onder meer bewoord word. Die wêreldwye paradigmaverskuiwing neem sedert die Alma Ata-konferensie primêre gesondheidsorg as visie en missie (World Health Organisation and UNICEF 1978).

Tydens paradigmaverskuiwings herleef en groei die gewildheid van alternatiewe geneeskunde. *Alternatief* beskryf in dié konteks "ander gesondheidspraktisyns as die ortodokse", en sluit onder altesaam 27 ander in die VSA homeopate, herbaliste, chiropraktisyns en sjamaniste in. Die alternatiewe ideologieë gebruik nie wetenskaplike terminologie nie, want hulle is uiteraard antiwetenskapsideologieë. Alternatiewe praktisyns groei tans in die VSA vinniger aan as ortodokse geneeshere (Wardwell 1994). Tydens paradigmaverskuiwings word geprobeer om alternatiewe, tradisionele en selfhelpgeneeskunde in die hoofstroom van gesondheidsdienste in te trek.[2] Opflikkerings van alternatiewe geneeskunde kom voor wanneer ortodokse geneeskunde, om watter rede ook al, nie meer aan die gevoelde behoeftes van 'n gemeenskap voldoen nie. Anders as voriges, het die huidige alternatiewe beweging 'n stewiger infrastruktuur en finansiële borgskappe, en mag dus langer voortduur (Yankauer 1997).

Dekades lank is daar 'n internasionale soeke na 'n nuwe mediese ("holistiese" of "biopsigososiale") model (Engel 1977) wat die eertydse antiwetenskap- en antirasionaliteitsentimente aanwakker. Kommunikasie tussen leke- en professionele gesondheidswerkers word tans weer toenemend belangrik, soos met die paradigmaverskuiwings van 150 en 75 jaar gelede.

Mediese vaktaal, primêr vak- en nie taalgerig nie, is gegrond op wetenskaplik-filosofiese begrippe en nie op tradisioneel-kulturele aannames van die omgangstaal nie. Presiese woordgebruik kenmerk die eerste, en simboliese woordgebruik die tweede. Vir alle mense is hul liggaam die verwysingsraamwerk vir simboliese antropomorfismes, veral belangrik in die opkomende dissipline "transkulturele psigiatrie". Die afgelope dekade word in Engels al meer onderskei tussen *disease* en *illness* wat albei voorheen as *siekte* vertaal en eenders verklaar is. Die eerste word tans meer spesifiek *vir* 'n objektiewe patologiese entiteit gebruik, en die tweede vir 'n subjektiewe belewenis of eksistensiële ervaring van 'n siekteproses, die beste deur die Franse *malaise* beskryf.

Die Europese renaissance van mediese wetenskappe in die middel- 17de eeu val saam met die Wes-Europese besetting van die Kaap, die "medifisering" van massas ongeletterde lyfeienes in Europa, en die ontstaan van mediese glossografieë. Jan van Riebeeck (1618-1677) was 'n tydgenoot van William Harvey (1578-1657), ontdekker van die bloedsomloop, en Antonie van Leeuwenhoek (1632-1723), uitvinder van die mikroskoop — albei baanbrekers van die mediese renaissance.

Mediese handleidings ("manuals"), 'n innovasie wat die oorwegend onge-

letterde Europese bevolking in die elementêre beginsels van higiëne, diagnose en behandeling moes onderrig, was wesenlik verklarende terminologieë vir die nuwe mediese wetenskap. Die weiniges wat kon lees, moes die handleidings aan ongeletterdes voorlees en hulle daarin "katkiseer" met die doel om geneeskunde van sy mistiek te stroop en ongeletterde lyfeienes insig te gee in die nuwe mediese denke (Heller 1976). Voorskrifte en diagnoses was nog in Latyn, en wetenskaplike mededelings is in koerante gedoen, gewoonlik as langdradige, obskure filosofiese redenasies. Die mediese handleidings moes óók medici oortuig dat dit voordelig was dat pasiënte hulle verstaan ten einde 'n stewige bolwerk te vorm teen kwaksalwery en die uitbuiting van ongeletterde onkunde. Kwaksalwery het gedy aan die Kaap, en met inheemse magiese Koikoigeneeskunde en diverse tipes "helers" was sinvolle kommunikasie in die renaissanceparadigma onmoontlik — daarom dat die eerste (ongepubliseerde) medies-vakkundige manuskrip in Suid-Afrika juis só 'n handleiding was (Burrows 1958: 62). Trouens, Häszner (1793) haal die twee beroemdste Europese handleidings van Buchan en Tissot aan, en werp lig op medies-maatskaplike toestande aan die Kaap (Pretorius 1992).

Dokters het welvarende landhere en adellikes bedien, maar is gewantrou deur die armes wat aangewys was op selfhelpboererate en rondreisende "helers" van allerlei oortuigings. Medifisering van die breë gemeenskap begin éérs toe dokters met die armes in aanraking kom, meesal ten tye van epidemies (die Swart Dood in Europa; pokke aan die Kaap). Die middel- 17de-eeuse mediese glossografieë van Europa[3] was 'n uitvloeisel van die handleidingveldtog wat geprobeer het om die kommunikasiegaping in gesondheidsdienste as gevolg van geletterdheids- en sosio-ekonomiese agterstande te oorbrug (Rosenberg 1983) — 'n taak wat weereens die Afrikaanse mediese woordeboek ten laste gelê word.

Die omvorming van armehuise tot hospitale was nog 'n groeistimulus vir 'n mediese vaktaal in Europa. Die toonaangewendes is in die akademiese sfeer ingetrek as "akademiese hospitale" van mediese fakulteite.[4]

Immigrantedokters, almal produkte van die renaissance in die Europese geneeskunde, het hul mediese vaktaal na Suid-Afrika saamgebring. Onder die Britse Setlaars van 1820 was daar 19 dokters wat 'n sterk akademiese invloed in die Oos-Kaap uitgeoefen het (Blumberg 1974). Duits-Joodse dokters, vanaf 1840 deur die Karoodorpgemeenskappe geassimileer, was die oorsprong van "die slim Joodse dokter" in die Afrikaanse idioom. Duitsland was toe die toonaangewende mediese moondheid (Burrows 1958: 187-188). Die Afrikaanse mediese vaktaal het uit die ingevoerde Hollandse, Engelse en Duitse vakterminologie gegroei, tale wat almal reeds die Grieks-Latynse vakterme volgens hulle eie besondere taalgebruik geassimileer het. Snyman (1988[3]: xi e.v.) bespreek die leksikografiese aspekte daarvan.

Kommunikasie[5] vorm die brug tussen abstrakte geneeskundekennis en die werklikheid van die siekbed. Die abstrakte kennis moet dáár aan elke individuele pasiënt persoonlik gekommunikeer word. Vakkennis moet uiteindelik tot

die verstaanbaarheid van gewone taal omvorm word. Volgens die sofistikasie van gespreksgenote mag kommunikasie op drie vlakke geskied:

- *interkollegiaal* tussen professionele beroepsmense onderling,
- *kommunaal* as 'n verstaanbare een-tot-baie-oordrag van gesondheidsinligting aan verskeie taalgemeenskappe, bv. oor bedreigings (rook, alkohol, eetgewoontes), epidemiologiese inligting (malaria, VIGS, cholera) en die bekendstelling van diensmodaliteite en -gebruike (voorgeboorte-, bejaarde-, en kindersorg; private-, openbare-, en daghospitaaldienste en -benutting), en
- *transaksioneel* as 'n een-tot-een vakman-leek-interaksie tussen dokter-pasiënt-vennote in 'n kliniese transaksie.

Op interkollegiale vlak het elke taal maar één mediese vaktaal met vakterme wat baie spesifieke betekenisse dra om vir die wydste moontlike professionele gebruikersgroep as gesaghebbende terminologiebron en vakkundige spellys te dien. Unieke leksikografiese vereistes word aan mediese woordeboeke gestel:

> Unlike editors of general dictionaries, lexicographers in scientific and technical fields are still expected to perform a prescriptive function, upholding standards of correctness and consistency. Here the problem is that the nomenclature of nearly every biomedical field is unstable. A term may have one meaning in formal usage and quite another in the jargon of practitioners. (Dirckx 1997: vi)

Woordeboeke in dié klas volg altyd éérs op die ontstaan van plaaslike geneeskundefakulteite, biblioteke en gevestigde vaktydskrifte. Twee omvattende Afrikaanse mediese woordeboeke voorsien in hierdie behoefte vir Afrikaans: Brink (1979) se verklarende woordeboek (WAG) en Snyman (1988[3]) se vertalende woordeboek (GW), respektiewelik produkte van mediese fakulteite gestig aan die Universiteit van Stellenbosch (1956) en die Universiteit van Pretoria (1948). Hul gewildste eweknieë is dié van Dorland (1864; 1994[28]) en Stedman (1911; 1995[26]). Deurlopende hersiening, bywerking en vertolking van internasionale vakterminologiewoordeboeke (soos gehanteer deur hul nasionale universiteite) besorg aan hierdie woordeboeke internasionale gesag en aansien. *Butterworths Medical Dictionary* (MacNalty 1965[3]: v-xii) verklaar eksplisiet dat 'n land se akademiese departemente gewoonlik samestellers lei in hulle hanteringswyse van internasionale vakterminologiese voorstelle. Na 'n meningsopname onder sy gebruikers besluit die jongste SMD (1995[26]: vi) om van sy 92 jaar lange gebruik van Latynse terme (NA) af te sien en oor te skakel na Engels (BR)[6]. Vakterminologiese hersiening en bywerking van farmaseutiese, bakteriologiese, chemiese en psigiatriese terme word gebaseer op óf amptelike farmakopeë of standaardteksboeke óf, in geval van die psigiatriese terminologie, DSM-III-R (*Diagnostic and Statistical Manual of Mental Disorders*, derde hersiening, 1987).

Die PW val egter in die kader van beknopte, spesifiek-gerigte woordeboeke ingestel op 'n breë, ongedefinieerde gebruikersgroep, soos Van der Merwe en Louw (1935) se *Mediese Woordeboek (met inbegrip van Veeartsenykundige, Tandheelkundige en Hospitaal-benaminge)*, Hansen (1962) se *Beknopte Mediese Woordeboek/Concise Medical Dictionary*, en Rompel (1975) se *Nurses' Dictionary English-Afrikaans/Verpleegsterswoordeboek Engels-Afrikaans*. Die ongedefinieerde gebruikersgroep word nogtans taamlik homogeen omskryf:

by Van der Merwe en Louw (1935: I):

> "verpleegsters en professioneel opgeleide persone wat vakkundige stukke in Afrikaans wil vertaal of wat in Afrikaans oor geneeskundige onderwerpe wil skrywe of lesings wil hou", asook "nie-vakkundiges" wat "Engelse stukke in Afrikaans" wil "oorsit",
>
> (Daar was destyds nog slegs twee Engelstalige mediese fakulteite in die land.)

by Hansen (1962: vii):

> "verpleegsters, hulp-geneeskundiges, noodhelpers en alle ander persone wat basiese geneeskundige terme magtig moet wees", en

by Rompel (1975: Preface):

> benewens die "nurses" van die titel, ook "medical auxiliaries, first-aiders, medical secretaries, and all others who require a knowledge of basic medical terms".

Sowel Hansen as Rompel erken die hulp van die Taalkomitee, Fakulteit Geneeskunde, Universiteit van Stellenbosch. Dié drie beknopte woordeboeke het elk minder as 10 000, sowat 'n kwart van die trefwoordinskrywings in WAG (1979) of GW (1988[3]). Selfs nóg korter en meer spesifiek gerig tot niemediese fisiologiestudente is die glossarium (469 fisiologiese vakterme) agter in die fisiologieteksboek van Meyer en Meij (1987) (vermeld in die PW se bronnelys). Rompel (1975) het minder trefwoorde (4 709) as die PW (6 865), maar vertaal veel meer klinies-diagnostiese terme as die PW. Die PW, anders as sy twee voorgangers, is egter verklarend en nie slegs vertalend nie.

Mediese joernalistiek gedy tans, met 'n eiesoortige behoefte aan verklarende woordeboeke. Nederlandse studies bevind dat 90% koerantberigte een of meer foute bevat, en ongeveer die helfte tydskrifte publiseer onnoukeurige sensasionele artikels — 'n eenderse situasie as in Suid-Afrika (Brummer 1986).

Die mediese vaktaal brei letterlik astronomies uit:

> Adepts of the Big Bang theory will recognise that if the exponential increase seen in the decades 1950 through 1970 continues, then some

> time during the next century Index Medicus will be growing faster than the generally accepted rate for the expansion of the universe. (Taylor 1992)

Daar is sowat 1 350 mediese fakulteite in die wêreld (126 in die VSA, 8 in Suid-Afrika) waar die meeste professionele vaktaalgebruikers opgelei word, met nuwer kategorieë hulppersoneel veral aan technikons.

2. Die Paramediese Woordeboek

2.1 Titel

Die enigma van die titel wentel om uitkenning van die teikengebruikers: Is dit bedoel vir "beroepsgroepe aanvullend tot geneeskunde" ("professions allied to medicine"), of "aanverwante beroepe" (WAG)? Is daar sprake van 'n nuwe mediese terminologie, benewens die gebruiklike? Die werk se enigmatiese titel hou ook verband met sy inhoud.

In die 1960's al het verpleegkundiges, ten regte, beswaar gehad teen die beroepsaanduiding "paramedies" vanweë hul integrerende rol in gesondheidsberoepe. Die term verwys, vanaf die 1970's, plaaslik én internasionaal, toenemend na 'n spesifieke beroepsgroep, die "paramedici", wat ambulanspersoneel, spesifiek vir noodsituasies opgelei, aandui. Die Suid-Afrikaanse Geneeskundige en Tandheelkundige Raad (SAGTR) registreer paramedici onder die omvattende beroepsgroep "noodsorgbemanning" ("emergency care personnel"), met subkategorieë ambulansnoodpersoneel en -assistente, basiese ambulansassistente, operasionele noodsorgordonnanse en noodsorgassistente — slegs één van 12 "aanvullende" professionele beroepsgroepe (SAGTR, 1997). Engelse woordeboektitels is duideliker, bv. *Medical and* Allied Health *Dictionary* (1997[3]).

Die PW definieer *paramedies* soos volg:

> wat een of ander verband met die mediese wetenskap het; aanvullend tot geneeskunde in die handhawing en herstel van normale gesondheid (paramediese werkers sluit in terapeute, verpleegpersoneel, geneeskundige maatskaplike werkers, ambulanspersoneel, aptekers, dieetkundiges, ens.).

Die pleonastiese taakomskrywing "herstel van normale gesondheid" vra by implikasie wat "abnormale" gesondheid is. Die lys paramediese vakgebiede, op die agterblad en in die eerste sin van die Voorwoord herhaal, lys dáár ook "skoliere met biologie as vak" en "studente van anatomie en fisiologie". 'n Meer beskrywende titel kon *Skoolwoordeboek van Anatomiese en Fisiologiese Terme* gewees het — 'n vermoede bevestig deur die inhoud en bronnelys (p. 288-289) waar 24 van die 30 gelyste bronne boeke is wat op skoolvlak gebruik mag word. Vir terapeute, onder die dosyn aanverwante professies van die SAGTR, is daar min

vakeie terme. Tandheelkundige terme word wel ruim gedek. Vir paramedici sou, onder meer, terme soos *anabiose, binneaarse oorgieting, resussitasie, kardioversie,* en verskeie tipes spalke, lugweë en draagbare verwag word, maar behalwe *binneaars* is geeneen opgeneem nie.

Die verwysing na paramediese *terme* verdiep dus die enigma: Watter terme word na dié kategorie gedelegeer? Dat "dit bykans onmoontlik (is) om alle terme wat in die paramediese veld bestaan, in 'n woordeboek te vervat, aangesien dit so 'n wye veld dek", soos die outeur in die Voorwoord sê, werp geen lig op die keuringskriteria of -werkswyse wat wel vir seleksie gevolg is nie. Hoe kwalifiseer die ongewone *cholemie* vir opname ten koste van bv. die meer dikwels gebruikte, maar weggelate *choledochus, -sistitis* of *-sistektomie*? Of waarom word *hidrotoraks* en *-kefalus* opgeneem, maar nie *hidronefrose* nie, of *hematofaag* en *-blast*, maar nie *hematemese* of *melena* nie? Die suggestie van 'n afsonderlike terminologie vir die paramediese veld is baie verwarrend:

> Although each profession and specialty area has its own precise terminology, most health workers share a single vocabulary. (Smith en Smith 1986: ix)

Die seleksie móét deur praktisyns aktief in die betrokke beroep gedoen word. Dit vereis 'n *seleksiepaneel* as die teikengroep verskeie dissiplines omvat. Tensy gedoen deur praktisyns met "strong backgrounds in scholarship and practice in their respective fields" (Dirckx 1997: vi), kan skerp kritiek verwag word soos destyds met resensies van Mönnig, Van der Merwe en Louw (1944) en Boshoff (1953) gebeur het. Die kritiek was afkomstig van dosente van vergelykende anatomie, embriologie en veeartsenykundige anatomie (óók paramediese dissiplines?). Die voorkeur vir vertalings na Afrikaanse omgangstaalwoorde en verafrikaansings en neologismes was onaanvaarbaar. Grobbelaar[7] (1954) lys enkeles waaronder *gorrel* ("trachea"), *hooflugpyp* ("bronchus") en *lug(pypie)blaas, asemlugpypie* ("alveolar bronchiole"), en terminoloë se langer en moeiliker spelbare verafrikaansings bv. *wandgataskruispunt* ("obelion"), *vrugpisuitleiergroef* ("urethral groove") en *saaddiertjie* ("spermatozoon").

Omgangstaalterme ingebou in verklarings kan duidelikheid vir 'n lekeleser bring, en die PW maak ruim hiervan gebruik: Akkuraat-beskrywende, gewone omgangstaalwoorde mag eventueel in 'n toekomstige "woordeboek vir primêre gesondheidsorg" tereg kom bv. *sluitspier* vir *sfinkter, slukbeswaar* vir *disfagie,* en *strekspiere* en *buigspiere* vir *ekstensors* en *fleksors*. Die ou strewe om geneeskunde te demistifiseer mag só verwesenlik, en gelatiniseerde begrippe vir die leek verstaanbaar gemaak word soos deur die eertydse "handleidings" beoog. Die opname van ongewone en ongebruiklike Nederlandse terme soos *eetsel* vir *fagosiet* of *vleëlgewrig* vir *los gewrig* moet egter liefs vermy word. 'n Titelwysiging na iets skerper gefokus op identifiseerbare teikengroepe mag van die PW 'n baanbreker maak, met natuurlik streng keuring van terme vir ('n) gedefinieerde teikengroep(e). Moontlike titelwoorde is *primêre gesondheidsorg,*

mediese leketerme, selfs *geneeskundeterme*, in die lig van die agterbladaanduiding na "in besonder vir die gewone mens".

2.2 Voorwoord

Die PW, die werk van 'n enkelouteur, bedank by name individue van die Nasionale Terminologiediens, en meld 'n ontstaansgeskiedenis van vyf jaar. Die verwysing na "min vakliteratuur (wat) in Afrikaans beskikbaar is", is klaarblyklik relatief tot wat voldoende geag word, maar is tog betwyfelbaar. Daar is tans Afrikaanse mediese teksboeke op voor- en nagraadse vlak in byna elke spesialiteitsgebied van genees- en verpleegkunde beskikbaar, sommige met verskeie oplaes.[8] Verskeie vakgerigte Afrikaanse woordeboeke bedien die (mediese) basiswetenskappe (chemie, biologie, fisika, statistiek en rekenaarterminologie) en prekliniese wetenskappe (aptekerswese). Insluiting van soveel van hierdie terme in die PW neem onnodig ruimte in beslag wat beter benut kon gewees het.

"Dit is juis die behoefte aan eenvoudige verklarings, waaraan geneeskundige woordeboeke nie aandag gee nie," sê die outeur in die Voorwoord, "wat gelei het tot die samestelling van hierdie woordeboek." Hierdie beweegrede (ter wille van skoliere met biologie as vak?) word deur verskeie definisies weerspreek wanneer hulle bv. met dié in WAG vergelyk word:

omentum
'n vou of verlenging van die peritoneum wat óf vry óf as verbinding tussen twee organe in die buikholte strek (PW)
Dubbele laag peritoneum wat tussen twee buikorgane strek (WAG)

nomogram
'n grafiese voorstelling van numeriese verwantskappe deur enige van vele metodes (PW)
Voorstelling van korrelasies d.m.v. grafieke of kaarte (WAG)

aboraal
t.o.v. dele wat aan die teenoorgestelde kant as die mond voorkom (PW)
M.b.t. streke weg van die mond af (WAG)

"Die skryfwyse van terme het heelwat probleme opgelewer vanweë uiteenlopende beskouings en beperkte Afrikaanse bronne," beweer die outeur verder in die Voorwoord. Veertig jaar gelede was so 'n stelling korrek, maar is tans kennelik onjuis. Die Afrikaanse vaktaal geniet reeds nasionale en internasionale erkenning en voorsien in *alle* vaktaalbehoeftes, soos onder meer blyk uit die gebruik daarvan in die *Suid-Afrikaanse Mediese Joernaal* (SAMJ),[9] lyfblad van die Mediese Vereniging van Suid-Afrika (MVSA). Die internasionale *Index Medicus* dui reeds sedert 1966 (vertaalde) titels van Afrikaanse mediese artikels aan met (*Afr.*) agterna. Die SAMJ, 'n internasionaal gesiene vaktydskrif, dikwels in die

Citation Index vermeld, is steeds tweetalig (soos ál die MVSA se publikasies). Afrikaans verskyn in amptelike stukke van die statutêre Suid-Afrikaanse Geneeskundige en Tandheelkundige Raad (SAGTR). 'n Groot aantal Afrikaanse verwysingsbriewe met 'n redelik vaste skryfwyse van terme word daagliks tussen medici en ander beroepsgroepe gewissel. Professionele gesprekke, voordragte, en populêre tydskrifte benut vakterme vrylik en verstaanbaar. Drie Afrikaanse mediese fakulteite gebruik daagliks Afrikaans as onderrigtaal van honderde studente, en notas in Afrikaans word uitgedeel.

Daar is by Afrikaanssprekendes wel 'n ontoepaslik puristiese gevoeligheid oor alternatiewe spellings van woorde. Soos Snyman (1988[3]: xiii) uitwys, is die verskille tussen Engels en Anglo-Amerikaans groter: Stedman (1995[26]: xx) vermeld gevalle waar Engels die alfabetisering van Anglo-Amerikaanse trefwoorde raak, en gee hulle dan aan met 'n kruisverwysing na die gebruiklike Anglo-Amerikaanse spelvorm. Die Engelse *ae*, *oe* en *ou* word in Anglo-Amerikaans *e*, *e* en *o* onderskeidelik, sodat naas *aetiology*, *oedema* en *tumour* ook *etiology*, *edema* en *tumor* bestaan. Alhoewel die PW deurgaans slegs *-ien-* en nie *-ine-* vorme gee nie, word naas *urien* en *vitamien* tog *urine* en *vitamine* gegee wat as normale wisselvorme nie (meer) steurend is nie. Aangesien ook slegs *-ied-* en nie *-ide-*vorme opgeneem word nie, is dit daarom vreemd dat *piramied* nie ook as normale wisselvorm van *piramide* gegee word nie.

2.3 Konstruksie (Ordening)

Naas die Voorwoord en die Bronnelys vermeld die inhoudsopgawe van die PW die volgende: Verduidelikende aantekeninge by die hooflys, Hooflys Afrikaans-Engels (met verklarings in Afrikaans), Verduidelikende aantekeninge by die glossarium, Glossarium (Engels-Afrikaans), Voor- en agtervoegsels, Griekse en Latynse stamme, afleidings en verbindingsvorme.

2.3.1 Die aantekeninge verduidelik die gebruik van die samestellingsmetodiek, verskillende hakies en afkortings, en die manier van aanduiding van sinonieme, kruisverwysings, meervoudsvorme en verkleiningsvorme.

In die verduidelikende aantekeninge by sowel die hooflys as die glossarium kom ongeveer dieselfde inskrywing ten opsigte van trefwoorde voor. Die een wat die hooflys voorafgaan, lui:

> Trefwoorde met dieselfde spelling maar verskillende betekenisse word genommer om kruisverwysing te vergemaklik, byvoorbeeld:
>
> **suture**[1] steek
> **suture**[2] hegdraad

(Terloops: *suture* se vertaling "beennaat (tussen skedelbene)" is nie opgeneem nie.)

Dié inskrywing hang saam met die volgende verduidelikende aantekening wat

die glossarium voorafgaan:

> Waar die Engelse trefwoord op verskillende woordsoorte dui, word dit aangedui, byvoorbeeld:
>
> **block**[1] **n.** blok
> **block**[2] **v.** blokkeer, onderdruk

Alhoewel dit miskien kruisverwysing vergemaklik, is die gebruik van verhewe syfers ongewoon en verwarrend omdat dit die verskil tussen homonimie, polisemie en funksiewisseling ophef (toegegee dat woordsoortverskille óók deur afkortings soos **n.** en **v.** aangedui word). In die glossarium sou 'n mens eerder die volgende inskrywing verwag:

> **suture** **1** steek **2** hegdraad

In die hooflys kom die volgende voor:

> **steek**[1] **n.** *suture* hegting met naald en gare
> **steek**[2] **n.** *stitch* skielike skerp pyn, soos bv. miltsteek
> **steek**[3] **v.** *stab* deurboor met 'n skerp voorwerp

In Afrikaans is dié drie inskrywings etimologies verwant. Verhewe syfers sou miskien wel gebruik kon word om **steek**[3], die werkwoord, van **steek**[1] en **steek**[2], die selfstandige naamwoorde, te onderskei, maar **steek**[1] en **steek**[2] hoort saam:

> **steek**[1] **n.** **1** *suture* hegting met naald en gare **2** *stitch* skielike skerp pyn, soos bv. miltsteek
> **steek**[2] **v.** *stab* deurboor met 'n skerp voorwerp

Die samehang tussen **brug**[1], **brug**[2], **brug**[3] en **brug**[4] sou baie duideliker deur slegs 'n enkele inskrywing soos die volgende getoon kon word:

> **brug** *bridge* **1** ⟨kyk pons⟩ **2** beenagtige verhewendheid van die neus **3** middelste deel van die voet **4** smal weefselstrook

Dit geld óók vir, onder meer, die vier afsonderlike inskrywings elk van *balans*, *basis* en *bors*.

Dat die verwysingstegniek wat in die PW gebruik word, nie noodwendig so duidelik is nie, blyk uit die inskrywings **substraat**[1], **substraat**[2], **substratum**[1] en **substratum**[2]. By **substratum**[1] word gesê: ⟨kyk substraat⟩, maar die gebruiker moet dan self agterkom dat die verwysing sowel **substraat**[1] as **substraat**[2] insluit. (Terloops: slegs **substraat**[1] se verklaring is die aanvaarde; **substraat**[2] is 'n ongewone sinoniem vir *kweekbodem* ("culture medium") wat nie opgeneem is nie.)

Die gebruik van kruisverwysings is een van die tegnieke om terme wat saamhoort en vanweë die alfabetiese ordening dwarsoor 'n woordeboek versprei word, weer bymekaar uit te bring. Die aanwending van diagramme en tabelle is 'n ander. Meyer en Meij (1987) gebruik die periodieke tabel ruimtebesparend om ál die inligting oor elemente op 'n enkele bladsy (Bylae x) aan te

gee, in plaas van inskrywings onder al die letters van die alfabet — 'n gebruikersvriendelike aanbieding wat óók vir ander mediese woordeboeke aanbeveel kan word. Maar daar moet in ag geneem word dat 'n woordeboek soos die PW slegs dié elemente sal lemmatiseer wat medies van belang is.

2.3.2 Met "selection means rejection" beklemtoon Dirckx (1997: iv) die moeilike insluitingskeuring. Verskeie alledaagse terme (soos *binneoor, bal-en-potjie-gewrig, bloedbank*) word ewe goed in algemene verklarende woordeboeke (bv. VAW, 1993[8]) gehanteer. Vakwoordeboeke sluit doelbewus "gewone" of "selfverklarende" woorde uit, maar die kriteria van "gewoon" of "selfverklarend" is onduidelik. Rompel (1975) se voorbeelde is *bed, intemperance* en *integrity*; Brink s'n *gesweldheid, haakvormig* en *inasem*.

Daar is 'n aansienlike aantal gewone selfverklarende Afrikaanse woorde wat in gesondheidsverband gebruik word, moontlik as gevolg van 'n lang geskiedenis van selfhelpvolksgeneeskunde en 'n tekort aan vakkundiges, en dalk ook bevorder deur voubiljette in Hollandse medisynekiste en oornames uit ander tale. Miskien is daar behoefte aan 'n *verklarende mediese woordeboek van leketerme* om uniforme begrip aan sommige te verleen (bv. aan die Kaap, *was* "menstrueer"). Dit is belangrik vir transaksionele kommunikasie in die praktyk, veral in die lig van die huidige klem op primêre gesondheidsorg[10].

Die PW sluit *aambeeld, hamer* en *stiebeuel* (die oorbeentjieketting) in, maar laat bv. *aambei, bloedvint* en *maagwerkings* weg. *Jig* word, baie noodsaaklik, gedefinieer, want die meeste Afrikaanssprekendes gebruik dit verkeerdelik as versamelnaam vir álle gewrigsontstekings. Besonder baie gewone woorde word slegs vertaal in die PW en na meer vakkundig-gebruiklike terme kruisverwys bv. *dodelik, dokter, eier, eiland, lug, lugleegte, neerslag, spuug, steen* en *wang*. Hulle het plek in definisies maar nie as trefwoorde in die hooflys nie. *Seer* word vertaal as *sore*, en verklaar as "pynlike plek aan die liggaam", dog nie óók as "seerplek", "ulkus", "wond" of "bedseer" nie — 'n voorbeeld van die inherente dubbelsinnigheid van leke- teenoor vakgedefinieerde woordgebruik.

2.3.3 Die voor- en agtervoegsels, Griekse en Latynse stamme en afleidings en verbindingsvorme in die afsonderlike lys kon, myns insiens, gebruikersvriendeliker *in situ* op hul alfabetiese plek in die hooflys opgeneem gewees het. Met 37 trefwoorde in die hooflys wat met *peri-* begin, 17 met *para-*, 24 met *anti-* en 44 met *epi-* (altesaam 122), sal 'n enkele verklaring van die voorvoegsel gebruikersvriendeliker wees aan die begin van sulke lang lyste. In die plek van hierdie afsonderlike lys (p. 276-287) sou 'n saamgegroepeerde lys afkortings, asook akronieme en eponieme, 'n belangriker probleem van die mediese vaktaal, nuttig gewees het. Vreemde woorde se betekenis kan wel nog in 'n bepaalde konteks geantisipeer word, maar akronieme en eponieme het geen onderliggende logika nie — trouens, akronieme se letters word juis gerangskik om maklik op die tong te lê. Eponieme van toetse en reagense wat vandag nie meer in die kliniek teengekom word nie (bv. *Benedict-reagens* en *Benedict-toets*), kon wegge-

laat gewees het.

Homeo-, homo-, en *homoio-* word in dié lys as sinonieme gegee met betekenisse "eenders, onveranderlik, dieselfde, net soos", maar vakkundig-terminologies word tóg tussen hulle onderskei. *Homeo-* slaan spesifiek op *eenders* (soos in *homeopatie, homeostase*) en *homo-* spesifiek op *dieselfde* (soos in *homosigoot, homoniem, homofiel*). *Homopaat* en *homostase* het nie bestaansreg as vakterme nie. Die eenderse definisies, albei gelys in die PW, kyk die verskil tussen die dinamiese *eenders* en statiese *dieselfde* mis; die Grieks is respektiewelik *homoios* "eenders" en *homos* "dieselfde".

Gebruikersonvriendelik beskryf die konstruksie van die PW.

2.4 Definisies (Verklarings)

Kenmerkend van definisies in vakwoordeboeke is hul noodsaaklike *voorskriftelikheid* (Dirckx 1997: vi). Sommige van die PW se "eenvoudige verklarings" bots egter met die aanvaarde begrip en gebruik in die vakterminologie. By *aborsie* word 'n *embrio* (en nie 'n *fetus* nie) uitgewerp, en *parturisie* is die skeiding van 'n *fetus* (en nie 'n *embrio* nie) van sy moeder. *Embrio* en *fetus* word korrek as, respektiewelik, "die eerste twee maande van ontwikkeling", en "'n menslike embrio na agt weke van intra-uteriene ontwikkeling" verklaar. *Paring* ("copulation") is seksuele omgang tussen twee heteroseksuele mense, soos korrek onder *koïtus* (met ses sinonieme) verklaar, maar nie tussen "chromosome tydens meiose nie"; die "homoloë chromosome" se "aantrekking na mekaar" word beter as *afparing* beskryf.

'n Ontoereikende verklaring kom voor by **abdominopelviese holte** (holte inferior tot die diafragma) wat identies met *abdominale holte* beskryf word; laasgenoemde sluit óók die bekkenholte in, terwyl eersgenoemde die holte tussen die borskas en die bekken is. Opvallende feitefoute is: *Addison-siekte* is 'n tekort aan bynierskorshormone, en nie aan adrenokortikotrofiese hormone nie; *aldosteroon* is beslis nie (soos beklemtoon met *veral*) 'n *adrenokortikotrofiese* hormoon nie, maar 'n bynierskors*mineralokortikoïed* (soos korrek onder *aldosteroon* verklaar). Onder **pH** word twee afsonderlike, en verwarrende verklarings van die enigste definisie van *pH* gegee: pH (akroniem vir *power of* H+) is 'n *simbool* vir die negatiewe logaritme van waterstofioonkonsentrasie (uitgedruk as mol per liter) en aangegee in eenhede (en *nie* uitgedruk as mol per liter *nie*!). (Terloops, pH = 7 is slegs neutraal by 'n temperatuur van 22° C; by liggaamstemperatuur (37° C) is die neutrale punt 6.8.)

'n Ontlening soos *vleëlgewrig* uit Nederlands bly, selfs met behulp van HAT en VAW, onverstaanbaar. WAG vertaal *flail joint* ondubbelsinnig met *los gewrig*. *Sweserik* (soetvleis) wat as sinoniem van *timus* gegee word, kon net in HAT en VAW gevind word waar dit as sinoniem van *pankreas* (alvleisklier) voorkom.

Dalk verskuil "paramedies" in die titel die weglating van verskeie siekte-

name en kliniese definisies, as synde *medies* eerder as *paramedies*! *Porfirien*, die pigment, word verklaar, maar nie die siekte *porfirie* nie. (Suid-Afrika het die hoogste voorkoms daarvan ter wêreld; Suid-Afrikaanse navorsers het beduidend tot die wetenskaplike kennis en kliniese klassifikasie bygedra.) Bepaalde "paramediese gebruikers" (mediese sekretaresses, verteenwoordigers en joernaliste) sal dalk juis kliniese definisies van siektename en sindrome in die woordeboek soek (*hematemese, melena, hemoptise* ens.) om die diagnoses waarmee hulle te make het, te verstaan (en korrek te spel!) — afhangende natuurlik van hoe "paramedies" verstaan word!

Die inkonsekwente gebruik van byvoeglike naamwoorde is steurend by gevalle van endokriene hiperfunksie: by **Conn-sindroom** is daar 'n *buitensporige*, by **Cushing-siekte** 'n *oormatige* en by **Cushing-sindroom** 'n *oortollige* hormoonsekresie. Daar is subtiele verskille tussen hulle. *Oormatig* is die mees toepaslike; *buitensporig* is meer gebruiklik in die sin van "baie meer as wat nodig is", en *oortollig* in die sin van "meer as wat nodig is". Omdat hormoonvlakke normaalweg 'n wye reikwydte het, beskryf *buitensporig* en *oortollig*, myns insiens, die betekenis minder akkuraat as *oormatig*. Aan watter een egter voorkeur verleen word, dit móét immers konsekwent gebruik word.

'n Hinderlike gebruik is om woordgroepe bestaande uit byvoeglike plus selfstandige naamwoorde, alfabeties onder eersgenoemde te plaas: daar ontstaan gevolglik afsonderlike inskrywings ver uitmekaar (bv. **anatomiese, fisiologiese** en (weggelaat) **chirurgiese dooieruimte** i.p.v. **dooieruimte 1, 2** en **3**. By **dooieruimte**, die kernbegrip, word *anatomiese* en *fisiologiese dooieruimte* nie vermeld nie.

Die verouderde terme *innominaat arterie* en *innominaat been* is nie opgeneem nie, maar vir eersgenoemde word *onbenoemde slagaar* in plaas van die huidig gebruiklike *bragiosefaliese arterie* opgeneem en verklaar, terwyl vir laasgenoemde die tans korrekte *os coxae* opgeneem is en na *koksiks* verwys word waar die verklaring voorkom.

2.5 Spelling

Solank algemene Afrikaanse woordeboeke (VAW 1993[8], HAT 1994[3]; AWS 1991[8], en die *South African Multi-Language Dictionary and Phrase Book* 1991) steeds diftonge soos die dubbel-r in *diaree* gebruik, en die voorgestelde spelwyse van vaktaalwoordeboeke (WAG 1979 en GW 1988[3]) ignoreer, sal spelonsekerheid bly voortbestaan. GW, hieronder volledig aangehaal, gee suiwer beredeneerde riglyne vir die spelling van Afrikaanse vakterme. Die spelreëls van die AWS (1964) het óók vir WAG as grondslag gedien by die spel van Afrikaanse terme, maar waar dit egter nie altyd uitsluitsel kon gee nie, "is ander gesaghebbende Afrikaanse bronne gebruik" (Snyman 1988[3]: x):

> Die groot behoefte by die Afrikaanse weergawe, veral wat terme uit Grieks of Latyn betref, is om dit morfologies korrek maar so bondig en

beeldend moontlik te stel. Met die inlywing van Griekse of Latynse woordwortels in Afrikaans dra die betekenis van die oorspronklike woord besondere gewig. ... Diftonge is dus tot een fonetiese letter verminder, en die verbindings-r of -l verkieslik weggelaat bv. *diarrhoea* is saamgestel uit:

dia — deur rhein — vloei.

In die Griekse spelling is 'n verbindings-r volgens hul fonetiese benadering ingevoeg. In Afrikaans is die dubbel-r beide oorbodig en verwarrend. ... Ons spel dus die woord *diaree*.

Dieselfde geld vir die volgende voorbeelde:

menorrhoea — menoree
otorrhoea — otoree
haemorrhage — hemoragie
allorrhythmia — alloritmie.

'n Woord soos *papilloma* bestaan uit papil en -oma. Ons spel dit dus *papiloom*. Dieselfde geld vir die ander afleidings en samestellings van *papil*. Ook *tonsil* met sy afleidings het 'n oorbodige l wat weggelaat kan word, bv. *tonsillitis*: *tonsilitis* in Afrikaans. (Snyman 1988[3]: xi)

Met die uitsondering van *diarree*, spel die PW *hemoragie* en *alloritmie* korrek met een r, maar bevat nie die ander in bogenoemde lys nie. Die beginsel word wel gehandhaaf wanneer *omfaloree* met 'n enkele r gespel word; maar ongelukkig word die verklaring van *omfaloragie* dáár aangegee! Die spelling van *tonsilitis* sluit egter nie hierby aan nie: dit word met 'n dubbel-l opgeneem.

3. Toekomsperspektief op mediese woordeboeke

Die toekoms van Afrikaanse mediese woordeboeke sal ongetwyfeld dieselfde wees as dié van mediese woordeboeke oor die algemeen, en almal van die uitkoms van internasionale rigtingsoekende neigings en eksperimente met vaktaalwoordeboeke in die heersende era van paradigmaverskuiwings.

Die Nasionale Terminologiediens het kennis geneem van die hoë nasionale prioriteit op primêre gesondheidsorg en die dringende terminologiese behoefte, en 'n veeltalige publikasie word saamgestel (Jordaan-Weiss 1995), skynbaar met die doel om interkollegiale kommunikasie te bevorder. Die formele wetenskaplike onderrig van geregistreerde beroepslui geskied steeds in één van die wetenskapstale van Suid-Afrika, en gevolglik word probleme op daardie vlak nie voorsien nie. Dalk kan 'n kommunikasieleemte ontstaan as alternatiewe geneeskunde in die breë dienslewering betrek word. Die ter plaatse opleiding en gebruik van tolke in klinieke was ten spyte van groot leemtes tot dusver tog taamlik bevredigend vir transaksionele kommunikasie in Suid-Afrika, maar in 'n Kanadese eksperiment[11] was *stetoskoop* bv. nie vertaalbaar nie, omdat die

enigste Ojibwawoord vir *luister* beteken "om na die hele persoon te luister", en nie na sy spesifieke anatomiese dele nie! (Rafuse 1993). Kommunale kommunikasie is waarskynlik die hoër onmiddellike prioriteit in Suid-Afrika, juis om die veranderde klem en betekenis van primêre gesondheid sinvol te verduidelik.

Daar is toenemende voorkeur vir gewone taal in mediese woordeboeke en die PW, onder meer "vir die gewone mens" bedoel, steek sy kleim hier af. Die voorkeurverskuiwing van Latyn na Engels vir anatomiese terminologie is nóg 'n wyse waarop die aandrang op demistifikasie van geneeskunde met eenvoudige, maklik verstaanbare omgangstaalekwivalente tegemoet gekom word. Namate die paradigmaverskuiwing op dreef kom, ontplooi daar 'n ál wyer gesondheidsgesprekskring. 'n Bydraende faktor tot negatiewe kritiek op die PW lê by die ongelukkige feit dat die omvattende Afrikaanse verklarende woordeboek (WAG, vermeld in die PW se bronnelys) nie sedert sy verskyning hersien of bygewerk is nie. Die groei van nuwe terme word gebalanseer deur die aanwas van Afrikaanse mediese vakkundiges, wat onderleg in die immunologie, molekulêre biologie en genetika, sekerlik by hierdie groeipunte van die mediese terminologie ingespan kan word. Die hersiening en bywerking van WAG, en sowel die herdefiniëring as uitdunning van argaïese, ongebruiklike terme móét deurlopende en dringende aandag geniet, net soos by sy Engelse eweknieë reeds aansienlik lank (92 jaar by SMD en 130 jaar by DIM) gebeur. Die heraktivering van 'n taalkomitee soortgelyk aan dié van destyds, verkieslik saamgestel uit al drie Afrikaanse mediese fakulteite, is gebiedend noodsaaklik om terminologiese koers aan te dui tydens die onstabiele oorgangsfase na 'n nuwe kommunikasiestyl. Dit is wenslik om, soos die jongste internasionale mediese vakwoordeboek (SMD 1995: vi) doen, *hoëprofielterme*, beskryf as "terms (which) have so profoundly altered the way medicine is practised and consumed that they warrant more than the standard dictionary definition", te identifiseer en ruim toe te lig. Daar word 125 sodanige terme gedefinieer, en (in die alfabetiese lys) tussen dik groen horisontale lyne aangedui. 'n Omvattende standaardwerk behoort die gesaghebbende bron te bly waaruit dissiplinegerigte medewerkers deurlopend uittreksels en wysigings na behoefte kan maak ten einde begripsuniformiteit vir die vaktaal te verseker.

Die paradigmaverskuiwings sal waarskynlik nie soseer nuwe terminologie- as nuwe gebruikersbehoeftes skep nie. Die behoefte aan veeltalige mediese woordeboeke met eenvoudiger, alledaagse verklarings sal aanvanklik, soos in die geval van die PW, klaarblyklik op skoliere gerig wees, met vrylike gebruik van omgangstaal in die definisies van vakterme (*kieliebeen, ontwater, oorspeekselklier, streepliggaam* ens. in die PW) — 'n gebruik wat kommunikasie op primêre gesondheidsvlak sal bevorder. Dit sal enersyds die demistifikasie en verduidelikende doel van die 18de-eeuse mediese lekehandleidings en glossografieë soos dit tans weer op die voorgrond kom, in die 21ste eeu voortsit. Die eras is vergelykbaar ten opsigte van heersende onderwysagterstande, sosio-ekonomiese ongelykhede, en wisselende sofistikasievlakke, in groot dele van die huidige (derde)wêreldbevolking (insluitend Suid-Afrika). Die aandrang klink die

luidste juis in daardie gebiede waar die grootste verskille is, en die derde wêreld vorm drie kwart van die totale mensetal.

Die wêreldwye verandering in gesondheidsdienste is dinamies, en daar moet bygebly word by vordering ten opsigte van geletterdheid, leesvaardigheid, en snelstygende sofistikasievlakke onder agtergeblewenes wat daarop sal volg. 'n Hoër sofistikasievlak is reeds merkbaar onder staatshospitaalpasiënte. Mediese joernalistiek se (onversadigbare) behoefte aan artikels oor mediese sake gaan dikwels gepaard met 'n verduideliking van onderliggende teoretiese en fundamentele beginsels. Eerstetaalsprekers kommunikeer gemaklik met pasiënte uit hul eie taalgroep omdat die omgangstaal ook hul spreektaal is, maar anderstaliges is nie noodwendig vertroud met ekwivalente vir vakkundige terme in die gewone taal nie. In 'n groot akademiese hospitaal se buitepasiëntafdeling is dit opvallend dat Afrikaanssprekende pasiënte vandag tegniese en diagnostiese terme soos *hipertensie, diabetes, anemie, kardiogram* en *rekenaartomografie* gemaklik (en korrek) gebruik in teenstelling met dieselfde pasiëntbevolking en hospitaal 'n paar dekades gelede (eie waarneming). Met die werklas beter versprei tussen gemeenskapsklinieke en daghospitale sal professionele gesondheidswerkers dalk meer tyd bestee aan, en klem lê op, die mentor-aspek van gesondheidsorg.

Die ontplooiende paradigma verskuif ook interkollegiale behoeftes en gebruike. Die inhoud van plaaslike vakkundige tydskrifte het van oorwegend gevallestudies na byna uitsluitlik oorspronklike hoogs gespesialiseerde navorsingsartikels en terapeutiese proewe verskuif. Eersgenoemdes, veral gerig op 'n groot Afrikaanssprekende praktisynsgroep, het geleidelik geswig voor laasgenoemdes met, as teikengroep, 'n klein internasionale eliteleserskring van hoogs gespesialiseerde vakkundiges. Afrikaanse artikels in die *SAMJ* het radikaal afgeneem soos gespesialiseerde vakkundigheid toegeneem en Afrikaanssprekendes deskundige navorsing begin onderneem het. Prof. Daniel J. Ncayiyana (1998), redakteur, haal Mathews Phosa (*Zuid-Afrika*, Des. 1997) aan oor die gebruik en groei van Afrikaans: "Dis hoe 'n taal groei — hy moet gepraat word, geskryf word, gebruik word." Die volwaardige Afrikaanse vaktaal wat vir interkollegiale gebruik juis op hierdie wyse oor 'n halfeeu gegroei en ontwikkel het, staan nou voor die uitdaging om dit ook vir kommunale en transaksionele gebruik te doen.

'n Veeltalige (niemediese) woordeboek (Reynierse 1991) spreek tans gedeeltelik behoeftes op die gebied van kommunale en transaksionele kommunikasie aan, onder meer met geïllustreerde makroanatomiese terme. Visuele verduidelikings voorkom deels veroudering van verbale definisies, en is onmisbaar vir transaksionele kommunikasie met ongeletterdes.[12] Bepaalde begrippe (*allergie, bloeddruk* e.d.) en sinsnedes wat tydens die kliniese interaksie gebruik word, word aangegee.

Elektroniese telekommunikasietegnologie behoort op behoeftegedrewe wyse strategies benut te word, maar vervang nie 'n kernwoordeboek van presies gedefinieerde vakterme nie. Dit mag teikengebruikers skei op grond van

rekenaargeletterdheid of -beskikbaarheid, en toegang tot internetkommunikasie en -interaksie. Dié wat toegang het tot elektroniese telekommunikasietegnologie, mag dalk woordeboeke minder nuttig vind. Die chronologiese verskyning van die SMD (1995) se uittrekselwoordeboeke dui waarskynlik aan hoe 100 000 of meer trefwoorde en definisies kleiner teikengroepe selektief mag bereik. CD-ROM's en beknopte sakwoordeboeke gee gebruikers van vaktaalwoordeboeke toegang tot 'n groter terminologiese bron. Spesifieke konsultante is nodig vir die ekstrahering van terme vir die aanverwante beroepe ("allied health professions") sodat 'n vaktaal-"Bybel" nie tydens paradigmaverskuiwings in 'n vaktaal-"Babel" ontaard nie.

Die PW gee waarskynlik sy eerste treë op 'n lang en avontuurlike pad.

Aantekeninge

1. SMD (1995[16]: iv) het, benewens 'n internasionale edisie, ook vertaalde uitgawes in Grieks, Indies, Japannees, Portugees en Spaans.
2. In Suid-Afrika oorweeg die nuwe raad wat die Medisynebeheerraad vervang het, onder meer om voorsiening te maak vir jurisdiksie oor tradisionele en volksmedisynes. Dit mag mettertyd terminologiese implikasies hê. Die SAMNR se Navorsingsgroep vir Tradisionele Medisyne (Tramed) het pas (1998), in medewerking met tradisionele helers, 'n *South African Traditional Healers' Primary Health Care Handbook* gepubliseer.
3. Die eerste was dié van Thomas Blount (1656) *Glossographia ... hard words together with Divinity Terms, Law, Physick, Mathematicks and other Arts and Sciences explicated*, en Robert James (1743) *Medical Dictionary*, waaraan dr. Samuel Johnson, as "an amateur of physick" meegewerk het (MacNalty, 1965: v-vii). Met *physick* word hier "geneeskunde" bedoel.
4. Nederland word ná die Hervorming die toonaangewende akademies-mediese moondheid toe die Leidense mediese fakulteit 'n liberale toelatingsbeleid volg ten koste van die vroegste Italiaanse en Spaanse mediese fakulteite wat uitsluitlik Katolieke toegelaat het.
5. 'n Taakgroep van die Royal College of Physicians, Londen, is onlangs aangewys om die leemtes in dokter-pasiënt-kommunikasie te ondersoek, en het aanbevelings gemaak om dit reg te stel ("Improving Communication between Doctors and Patients: Summary and Recommendations of a Report of a Working Party of the Royal College of Physicians", *Journal of the Royal College of Physicians (London)* 31: 258-259, 1997).
6. In die 1950's was gelyktydig hersienings van anatomiese nomenklatuur aan die gang in Engeland (Birmingham Revision, BR), Switserland (Basle Nomina Anatomica, BNA) en Frankryk (Paris Nomina Anatomica, PNA). Eventueel maak akademiese inrigtings die keuse of hulle die Latynse (Nomina Anatomica hetsy van Basel hetsy van Parys) of die Engelse (BR) nomenklatuur wil gebruik, d.w.s. in die landstaal, óf Engels óf Anglo-Amerikaans. Natuurlik mag die landstaal óók Afrikaans of enige ander wetenskapstaal wees. Die NA term *articulatio talonavicularis* word in beide Engels en Anglo-Amerikaans *talonavicular joint* en in Afrikaans *talonavikulêre gewrig*. Die neiging van die Afrikaanse vakterminologie tot latinisering was destyds sterk, maar vandag word die internasionale tendens met die verafrikaansing van Latynse (NA) terminologie in die spreektaal gevolg.
7. Destyds senior lektor, Soölogie en Fisiese Antropologie, Universiteit van Stellenbosch.

8. 'n Veertigtal teksboeke uit die laaste kwarteeu kon getel word, slegs twee waarvan vertalings was van Engelse teksboeke. Afrikaanse teksboeke in verloskunde geniet voorkeur bo hul Engelse eweknieë omdat die lae geboortesyfers in Europese lande outeurs minder ervare op hul vakgebied maak as hul Suid-Afrikaanse eweknieë. Plaaslik is die opleiding in verloskunde sodanig dat pasgekwalifiseerdes hul beperkings ken, maar enige gewone verlossing tuis kan doen, iets wat in Kanada byvoorbeeld tot bepaalde inrigtings beperk word.
9. 'n Naamsverandering van die tydskrif, voorheen die *Suid-Afrikaanse Tydskrif vir Geneeskunde*, en afgekort as SATvG, is in die 1980's gemaak ter wille van 'n tweetalige akroniem, SAMJ, op die buiteblad. Koste was 'n belangrike oorweging by die ontwerp en modernisering van die buiteblad. 'n Meningsopname het getoon dat die hoofbeswaar teen *joernaal* was dat dit met die Engelse *magazine* verwar mag word. Ander wys daarop dat in Van Riebeeck se *Joernaal*, anders as in 'n tydskrif, slegs belangrike gebeurtenisse opgeteken is. Prof. H. W. Snyman se beswaar spesifiek was dat, soos *HAT* in 'n opmerking sê, *joernaal* in die betekenis van "wetenskaplike tydskrif" onder Engelse invloed is. Kosteoorwegings het uiteindelik die deurslag gegee!
10. In ligte, satiriese trant vra David Gunn ("An Appeal for Simpler Medical Terminology", *Canadian Medical Association Journal* 149: 1553-1667, 1993) o.a. waarom 'n *trommelvlies* juis 'n *tympanum* genoem moet word, waarom *hoë (bloed)druk* juis *hipertensie* moet wees, of *musculus lattissimus dorsi* meer sê as die *grootste rugspier*, en of *patella* enigiets anders as *knieskyf* kan beteken. Die Engelse argumente geld enige taal, óók Afrikaans, soos in bogenoemde (vertaalde) voorbeelde. Die kernbeswaar teen latinismes is: "It perpetuates the mystique that surrounds medicine: obscure medical terminology distances doctors from their patients" — 'n moderne stelling van die probleem wat die 17de-eeuse handleidings tydens die mediese renaissance wou aanspreek!
11. In Manitoba, Kanada word tradisionele Chinese geneeskunde tans gewild (vlgs. 'n brief van Edward J. Sheffmann "Traditional Chinese Medicine", *Canadian Association Medical Journal* 149: 1379, 1993). Daar het ook 'n veeltalige mediese woordeboek (Engels/Ojibwa/Cree/ Island dialect/ Deens) verskyn onder leiding van die Manitoba Association of Native Languages (MANL). Dié woordeboek sal, behalwe in klinieke, ook in skole gebruik word, maar die doel is nie soseer terminologies nie as om taalbewaring en trots op 'n nasionale erfenis te bevorder — in dié geval tradisionele of "aboriginal" geneeskunde (Rafuse 1993).
12. Daar is nog geen Afrikaanse mediese woordeboek met illustrasies of diagramme nie, maar dit word, dikwels ook in kleur, toenemend en op gespesialiseerde wyse gebruik in woordeboeke van Engelse mediese vaktaal. Stedman (1995) bevat illustrasies en diagramme met erkenning ontleen aan die elektroniese anatomie-atlas, A.D.A.M. Sommige diagramme en geanimeerde sketse in hierdie werk spreek 'n duisend woorde, en is van onskatbare waarde vir kommunikasie met ongeletterdes.

Bronnelys

Woordeboeke

Anderson, D.M. 1994[28]. *Dorland's Illustrated Medical Dictionary*. Philadelphia: W.B. Saunders.

Boshoff, S.P.E. 1953. *Lys Ontleedkundige Terme Engels-Afrikaans-Latyn.* Pretoria: Vaktaalburo, Suid-Afrikaanse Akademie vir Wetenskap en Kuns.

Brink, A.J. 1979. *Woordeboek van Afrikaanse Geneeskunde Terme.* Kaapstad: Nasou.

Dirckx, John H. (Red.). 1997[3]. *Stedman's Concise Medical and Allied Health Dictionary.* Baltimore: Williams & Wilkins.

Hansen, Ralph. 1962. *Beknopte Mediese Woordeboek / Concise Medical Dictionary.* Kaapstad: Sarel Marais.

Labuschagne, F.J. en L.C. Eksteen. 1993[8]. *Verklarende Afrikaanse Woordeboek.* Pretoria: J.L.van Schaik.

MacNalty, A.J. 1965[3]. *Butterworths Medical Dictionary.* Londen: Butterworths.

Mönnig, H.O., F.Z. van der Merwe en J.D. Louw. 1944. *Voorlopige Geneeskundige Woordelys.* Pretoria: Suid-Afrikaanse Akademie vir Wetenskap en Kuns.

Odendal, F.F. et al. 1997[3]. *Verklarende Handwoordeboek van die Afrikaanse Taal.* Midrand: Perskor.

Reynierse, Cecile (Red.). 1991. *South African Multi-Language Dictionary and Phrase Book.* Cape Town: Reader's Digest Association S.A.

Rompel, H. 1975. *Nurses' Dictionary English-Afrikaans / Verpleegsterswoordeboek Engels-Afrikaans.* Kaapstad: Tafelberg.

Smith, Brian and Bentley E. Smith. 1986. *Medical Teminology for the Health Professions.* New York: Academic Press.

Snyman, H.W. 1988[3]. *Geneeskundige Woordeboek.* Durban: Butterworth.

Spraycar, Marjory (Red.). 1995[26]. *Stedman's Medical Dictionary: Illustrated in Color.* Baltimore: Williams & Wilkins.

Taalkommissie. 1991[8]. *Afrikaanse Woordelys en Spelreëls.* Kaapstad: Tafelberg.

Van der Merwe, F.Z. en J.D. Louw. 1935. *Mediese Woordeboek (met inbegrip van Veeartsenykundige, Tandheelkundige en Hospitaal-benaminge).* Kaapstad: Nasionale Pers.

Ander bronne

Avery Jones, Francis (Red.). 1972. *Richard Asher Talking Sense.* Londen: Pitman Medical.

Blumberg, C. 1974. *The Provision of Medical Literature and Information in the Cape. 1827-1973.* Ongepubliseerde M.Bibl.-skripsie. Universiteit van Kaapstad.

Brummer, Tobie. 1986. Mediese joernalistiek onder die loep. *Suid-Afrikaanse Mediese Joernaal* 69: 275.

Burrows, E.H. 1958. *A History of Medicine in South Africa.* Kaapstad: A.A.Balkema.

Engel, G.L. 1977. The Need for A New Medical Model: A Challenge for Biomedicine. *Science* 196: 129-136.

Grobbelaar, C.S. 1954. Afrikaanse ontleedkundige terme: 'n resensie oor die lys uitgegee deur die Suid-Afrikaanse Akademie. *Suid-Afrikaanse Tydskrif vir Geneeskunde* 28: 480-483.

Häszner, J.F. 1793. *Huislijk geneeskundig handboek voor de ingesetenen van Nederlands Afrika.* Kaapse Argief, MOOC 14/143.

Heller, R. 1976. Priest-Doctors as a Rural Service. *Medical History* 20: 361-83.

Jordaan-Weiss, Milde. 1995. The National Terminology Services: A New Paradigm. *Lexikos* 5: 279-285.

Magner, Lois N. 1992. *A History of Medicine.* New York: Marcel Dekker.

Meyer, B.J. en Hester S. Meij. 1987. *Fisiologie van die mens: 'n Algemene oorsig.* Pretoria: HAUM.

Ncayiyana, Daniel J. 1998. Soweto se sjebeens en die toekoms van Afrikaans. Mini-redaksionele artikel. *Suid-Afrikaanse Mediese Joernaal* 88: 99.

Pretorius, J.C. 1992. Johann Friedrich Häszner (1764-1820). *Suid-Afrikaanse Tydskrif vir Kultuurgeskiedenis* 6: 124-134.

Rafuse, Jill. 1993. New Dictionary Provides Native-Language Equivalents of English Medical Terms. *Canadian Medical Association Journal* 149: 1537-1540.

Rosenberg, C.E. 1983. Medical Text and Social Context. *Bulletin of the History of Medicine* 57: 22-42.

SAGTR. 1997. *Registrasiegelde vir Beroepsrade, 1998.* Omsendbrief.

Taylor, C.R. 1992. Great Expectations — The Reading Habits of Year II Medical Students. *New England Journal of Medicine* 326: 1436-9.

Wardwell, W.I. 1994. Alternative Medicine in the United States. *Social Science and Medicine* 38: 1061-1068.

World Health Organisation and UNICEF. 1978. *Primary Health Care.* Geneva: WHO.

Yankauer, A. 1997. The Recurring Popularity of Alternative Medicine. *Perspectives in Biology and Medicine* 41: 132-138.

Afkortings gebruik in verwysings na woordeboeke

AWS *Afrikaanse Woordelys en Spelreëls* (Taalkommissie. 1991[8])
DIM *Dorland's Illustrated Medical Dictionary* (Anderson, D.M. 1994)
GW *Geneeskundige Woordeboek* (Snyman, H.W. 1988[3])
HAT *Verklarende Handwoordeboek van die Afrikaanse Taal* (Odendal, F.F. et al. 1997[7])
SMD *Stedman's Medical Dictionary: Illustrated in Color* (Spraycar, Marjory. 1995[26])
SCD *Stedman's Concise Medical and Allied Health Dictionary* (Dirckx, John H. 1997)
VAW *Verklarende Afrikaanse Woordeboek* (Labuschagne, F.J. en L.C. Eksteen. 1993[4])
WAG *Woordeboek van Afrikaanse Geneeskunde Terme* (Brink, A.J. 1979)

R.W. Burchfield. *The New Fowler's Modern English Usage.* 3rd edition 1996, xxiii + 864 pp. ISBN 0-19-869-126-2. Oxford: Clarendon Press. Price £16.99.

1. Introduction

R.W. Burchfield's revision (1996) of H.W. Fowler's *Modern English Usage* is an important book by a distinguished scholar. It provides an opportunity to assess both a large and useful new text, and, in passing, the largely prescriptive Fowler tradition up to its transformation by Burchfield.

Burchfield's aim is "to guide readers to make sensible choices in linguistically controversial areas of words, meanings, grammatical constructions and pronunciations" (Burchfield 1996: xi). For this he is well qualified by his long service as Chief Editor of the Oxford English Dictionaries (1971-1984), his *magnum opus* the four-volume *Supplement* to the *Oxford English Dictionary*, his wide array of publications on the English language and the range of his human experience, common sense and humour.

Numerous entries in his new Fowler bear on such topics as *Estuary English, gobbledegook, officialese, political correctness, racialism* ("one of the key words of the 20th century") and *sexist language*. The problem term *Mid-Atlantic* would have been a useful addition here. His intended readers are clearly sophisticated speakers of English as L1, with a grasp of the language well above that of the "foreign students of English" addressed by Michael Swan (1980) in his *Practical English Usage* or John Sinclair (1992) and his team for the *Collins Cobuild English Usage* text, heavily dependent on "actual examples not invented ones" (Sinclair 1992: iv). On this point Sinclair and Burchfield agree, but there are obvious differences between usage texts for L1 and L2 readers.

2. Usage Texts for L1 Speakers

Usage problems arise, of course, in all the three main fields of language study: grammar, semantics and phonetics/phonology. Early grammars tended to deal in passing with usage problems. Thus Lowth (1762) remarks in the Preface to his *Short Introduction to English Grammar*:

> Our best authors have committed gross mistakes

some of which he proceeds to correct.

Throughout the eighteenth and nineteenth centuries many factors, but perhaps chiefly the early efforts in education of missionary societies in their schools culminating in the Education Acts of 1870-1921, brought many thousands of dialect speakers into contact with standard English. *Standard* here is taken to mean the more or less codified form of the written language, now

fairly uniform in major English-using communities. There is perhaps no standard accent across the Englishes of the world, though broadcasting, governmental and upper-class usages cohere in a variety of local standards.

The demand for English usage texts for speakers of English as L1 is thus likely to have risen steadily from about the mid-nineteenth century. By the time that the Clarendon Press at Oxford published *The King's English* (1906) by the brothers H.W. and F.G. Fowler a fairly large range of usage-related texts was available, and usage publication had emerged as a *genre* of its own.

3. H.W. Fowler's *Modern English Usage*

H.W. Fowler's *Dictionary of Modern English Usage* appeared in 1926, eight years after the death of F.G. Fowler as a result of war service. It is a substantial book (viii and 742 pages) in its 1940 edition, basically a set of articles on words and phrases. The list of General Articles includes, however, such items as *Avoidance of the Obvious, Battered Ornaments, Swapping Horses* and *Word Patronage*. These suggest that a competent reader should ideally know the whole book when he consults it. Burchfield (1996) has no such list.

Fowler, as *Modern English Usage* came to be called, was widely read and cited. It was republished in 1965 in an edition lightly revised by Sir Ernest Gowers.

Despite the great popularity of the book, it was criticised by many linguists. Thus Hilda M. Murray (1926: 42) writes:

> The plan and execution are alike admirable and the matter excellent reading, though the reader may sometimes fail to distinguish between the voice of authority and that of private opinion.

Kemp Malone (1927: 201) remarks:

> In Mr Fowler's chosen field of activity, viz., linguistic science, sound and abiding work cannot be done by a man weak in phonetics and neglectful of the historical approach to the problems of which he writes.

The Gowers edition (1965) drew expressions of the traditional favourable view of "the unique Fowleresque quality which has made the book perennial" (*British Book News*, July 1965, 475) but also such condemnations as that of Barbara Strang (1966: 264):

> Fowler's attitude is not a possible one for a good mind in the 1960s.

Burchfield (1991: 101-106) from whom the last four quotations have been taken, sees the conflict as primarily one between linguists and nonlinguists. It has perhaps another dimension, that of social class. Thus Fowler in a letter of 1911 to his publisher writes:

> In point of fact we have our eyes not on the foreigner but on the half-educated Englishman of literary proclivities (Burchfield 1991: 96).

Fowler happens to have been educated at Rugby and Balliol College, Oxford, and to have taught English and Classics at Sedburgh School (1882-1899) where he was described as "a stickler for etiquette" (Burchfield in McArthur 1992: 414). It seems that his monument to the standard form of English attracted support in particular from the older members of society. Burchfield himself hints at this in his remark that "in the space of a few weeks a judge, a colonel, and a retired curator of Greek and Roman antiquities at the British Museum told me on separate social occasions that they have [Fowler] close at hand at all times" (Burchfield 1996: ix).

4. Databases

Burchfield (1996: ix) remarks in his Preface:

> From the start it was obvious to me that a standard work on English usage needs to be based on satisfactory modern evidence and that a great deal of the evidence could be obtained and classified by electronic means.

He proceeded to establish on a PC a database of ten independent fields with a numbering system within each, for example for gerunds:

> 3 = possessive with gerund
> *I was proud of his being accepted at such a good school — New Yorker* 1986.
>
> 4 = gerund without possessive
> *How could she think of a baby being born in the house* — A.S. Byatt 1985.

"In the end," he says, "my gerunds field contained examples of more than 100 types of gerundial constructions."

Materials were gathered "from a systematic reading of British and American newspapers, periodicals and fiction of the 1980s and 1990s in approximately equal proportions" (Burchfield 1996: x).

Burchfield had also access to the electronic and paper-slip files of the *Oxford English Dictionary*. This is an extremely rich collection of basic materials, worked over as he tells us "for nine years", during which he undertook several other projects.

The Fowlers had a much smaller database. *The King's English* gives sources of examples, citing the *Times* leading with five hundred and fifty, followed by the *Daily Telegraph* with ninety-six and the *Spectator* ninety-four. The editors

drawn upon are mainly nineteenth century British: "Scores of important writers of the Victorian period remained unexamined, or, at any rate, uncited" (Burchfield 1991: 99), and only four American writers are cited. American materials, indeed, are largely excluded from both *The King's English* and *Modern English Usage*. "I know absolutely nothing about American," wrote Fowler in 1927 in reply to an enquiry by his publisher about a possible Americanised version of *Modern English Usage* (Burchfield 1991: 97). *Modern English Usage* gives no sources for its quotations.

5. Grammar

Fowler uses the terms of traditional grammar without hesitation. In 1926 they had few competitors. Sixty years later Burchfield (1996: xi) writes:

> I judged it essential to retain the traditional terminology of English grammar: there are no tree-diagrams, no epistemic modality (except to explain what the term means), no generative grammar.

There is in fact one tree-diagram in his article on *clause*. Burchfield occasionally uses the term *determiner*, but does not define it.

I also miss *noun phrase*. Though neither *clause* nor *sentence* is a particularly good article, there are some good points in his article on *grammar*, notably the following plea:

> Ideally every English-speaking person should begin to distinguish the several parts of speech at an early age and continue to study the subject in a graduated manner throughout his or her time at school.

Both Fowler and Burchfield, however, are careful not to build complete mini-grammars into their texts. Neither text, for instance, has a "defining" article on *adjective, adverb, noun* or *verb*, though both have *clause, grammar* and *sentence*. Burchfield, moreover, has a short article on *standard English*, which is not in Fowler, even in the 1965 edition. The OED dates the phrase back to a *Quarterly Review* of 1836.

6. Styles of Treatment

These differ widely.

For *constable* Fowler (1926) has simply:

> *constable*. Pronounce kŭn-.

The brevity of this entry is not characteristic. Burchfield (1996: 175) has:

> *'constable.* Pronounce /'kʌnstəbəl/, but don't be surprised if you hear some standard speakers saying /'kɒn-/.

Here Fowler's amateurish phonetic rendering without a stress mark contrasts with Burchfield's IPA with stresses, though his second schwa instead of a syllabic /l̩/ is questionable. Burchfield also offers a possible variation and his style is characteristically relaxed.

Fowler condemns "our mutual friend", a favourite shibboleth of usage writers. Burchfield, however, points out that *mutual* has three long-established senses:

(a) "Reciprocal", as in *Wilde and Yeats reviewed each other's work with mutual regard.*
(b) "Common", as in *a mutual friend* in contexts in which *common* might imply vulgarity.
(c) "Pertaining to both parties", e.g. *of mutual benefit to both the Scots and the English.*

In constructions of type (c) *common* or *in common* are preferable if they fit idiomatically.

He makes no concessions to the dreaded *like* with which he deals with true Burchfieldian crispness. He gives four uses of this particular *like* among several articles on its forms:

1. As a conjunction.
2. As a preposition.
3. A hated parenthetic use.
4. Idiomatic phrases.

Under 3 there is a wealth of examples and a comment of which part follows:

> By the mid-20th century however, its use as an incoherent and prevalent filler had reached the proportions of an epidemic, and is now scorned by standard speakers as a vulgarism of the first order.

A pleasing example quoted is:

> *Naa, I was all into that last year, but like I don't think it's so relevant now* — M. du Plessis 1983

7. Conclusion

The appearance of this book is very well-timed. Towards the end of the century, at a time when achievement in English is so poor for millions of its learn-

ers even in L1 communities, it is of great value to have such a readable and far-reaching text on the standard English of England which takes regular cognisance of that of the United States.

Burchfield in fact inhabits a much more extensive world than did Fowler. Fowler died in 1933 having seen only a third of the twentieth century. Burchfield has seen far more and has rewritten a major text which will be an important guide to speakers and writers in the new millennium and an important auxiliary reference text for lexicographers of English.

References

Burchfield, R.W. 1991. The Fowler Brothers and the Tradition of Usage Handbooks. Leitner, G. (Ed.). 1991. *English Traditional Grammars*. Amsterdam/Philadelphia: John Benjamins.

Burchfield, R.W. 1996. *The New Fowler's Modern English Usage*. Oxford: Clarendon Press.

Fowler, H.W. 1926. *A Dictionary of Modern English Usage*. Oxford: Clarendon Press / London: Humphrey Milford.

Fowler, H.W. and Francis G. Fowler. 1906. *The King's English*. Oxford: Clarendon Press.

Lowth, Robert. 1762. *A Short Introduction to English Grammar*. Edited by M.V. Aldridge. 1973. Grahamstown: Institute for the Study of English in Africa.

Malone, Kemp. 1927. Review of Fowler *A Dictionary of Modern English Usage*. *Modern Language Notes* 42(3): 201-202.

McArthur, Tom (Ed.). 1992. *The Oxford Companion to the English Language*. Oxford: Oxford University Press.

Murray, Hilda M.R. 1926. Review of Fowler *A Dictionary of Modern English Usage*. *The Year's Work in English Studies 1926* 7: 42-43.

Sinclair, John (Ed.). 1992. *Collins Cobuild English Usage*. London: HarperCollins.

Strang, Barbara M.H. 1966. Review of Fowler *A Dictionary of Modern English Usage*, second edition. *Modern Language Review* 61(2): 264-265.

Swan, Michael. 1980. *Practical English Usage*. Oxford: Oxford University Press.

Bill Branford
Emeritus Professor
Rhodes University
Grahamstown
South Africa

Morton Benson, Evelyn Benson and Robert Ilson. *The BBI Dictionary of English Word Combinations*, revised edition 1997, xl + 386 pp. ISBN 90 272 2167 7 (Eur.) / 1-55619-521-4 (US). Amsterdam: John Benjamins. Prys *f*40,00 / $19,95.

The years 1986 and 1997 witnessed two important lexicographical events, the publication of the first edition of the *BBI Dictionary of English Word Combinations* and the appearance of the revised edition. English language teachers, students and translators are pleased that an urgently needed dictionary of collocations came into being and are delighted to see the new revised edition. The edition is enlarged and completely revised. It comprises 18 000 entries and 90 000 collocations which single it out as the most comprehensive English collocational dictionary available (cf. Kozlowska and Dzierzanowska, 1993). It is true that in the recent editions of *OALD* (1995) and *LDOCE* (1995) collocations have been brought into focus. However, foreign learners of English and translators were in serious need of a specialized collocational dictionary as the provision of collocations is not the primary aim of a general dictionary. They looked forward to a collocational dictionary which is comprehensive and up-to-date with quick access to the English word combinations which are, in the main, arbitrary and unpredictable. Fortunately this new edition fulfils all their needs.

A comparison of the 1986 and 1997 editions shows how whole articles have been reorganized and enlarged.

attack I *n.* ['assault'] (also fig.) 1. to carry out, *make*; *launch*, *mount*; *lead*, *spearhead*; *press an* ~ 2. to provoke an ~ 3. to blunt; break up, repel, repulse an ~ 4. (often mil.) an all-out, concerted, full-scale; coordinated; *mock*; *pre-emptive*; sneak, surprise ~ 5. (usu. mil.) an air; enemy; flank; frontal; torpedo ~ 6. a bitter, blistering, savage, scathing, sharp, violent; scurrilous, vicious; unprovoked; wanton ~ 7. an ~ fails, fizzles out; succeeds 8. an ~ against, on (our forces launched an all-out ~ against the enemy; he made a blistering ~ on his opponent) 9. under ~ ['onset of an *ailment*'] 10. to have an ~ (she had an ~ of hiccups) 11. an acute; light, slight; recurrent; sudden ~ 12. a fatal; heart ~
attack II *v.* to ~ viciously

The *BBI*, 1st edition, 1986

attack I *n.* ["assault"] (usu. mil.; also fig.) 1. to carry out, make; launch, mount, unleash; lead, spearhead; press an ~ 2. to provoke an ~ 3. to come under ~ 4. to survive, withstand an ~ 5. to blunt; break up, repel, repulse an ~ 6. an all-out, concerted, full-scale; coordinated; major ~ 7. a preemptive; retaliatory ~ 8. a mock; sneak, surprise ~ 9. an air; seaborne ~ 10. an enemy; terrorist ~ 11. a flank; frontal ~ 12. a nuclear; torpedo ~ 13. an ~ succeeds 14. an ~ fails; fizzles out 15. an ~ against, on (our forces launched an all-out ~ against the enemy) 16. under ~ ["a belligerent action"] (often verbal) 17. to make an ~ 18. a verbal ~ 19. a bitter, blistering, brutal, savage, scathing, sharp, vehement, vicious, violent; scurrilous; unprovoked; wanton ~ 20. an ~ on (he made a blistering ~ on his opponent) 21. (misc.) the leaked document left us open to ~ ["onset of an ailment"] 22. to have an ~ (she had an ~ of hiccups) 23. an acute; fatal; light, slight; recurrent; sudden ~ 24. a heart ~
attack II *v.* to ~ brutally, savagely, viciously; physically

The *BBI*, revised edition, 1997

The two salient features of the *BBI* in terms of which the degree of excellence and achievement must be measured are *coverage* and *lexicographical treatment*. In

both cases the *BBI* made a breakthrough. As regards coverage, I have been using the *BBI* (1997) while listening to the radio or reading English newspapers and translating from Arabic into English, and it has never failed me. The word combinations used in this edition reflect up-to-date occurrences in material from sources dealing with a wide range of subject matter: medicine, linguistics, commerce, music, politics, computer science. Many new collocations are added. They cover all the patterns, V+N, Adj.+N, N+V, V+Adv., e.g. *a cellular, mobile, pay telephone; human rights abuses; a software package; financial crunch; a computer virus; to log into; to shout abuse at; elder, spousal abuse; alcohol, substance abuse; seal off an area; a no-smoking area; to unleash an attack; a cerebral thrombosis; a desk-top, handheld, laptop computer; a machine-readable text; a crash, interdisciplinary, remedial course; a computer freezes up; to clamp sanctions on; an arms embargo.*

Some new headwords are also added, e.g. *terrorism, AIDS, bomber, stabilization, backlash, cross-country, thumbs-up, elected, electrical appliance, civil unrest, civil disobedience, civil disorder, cleansing, differing, flash point, freak, free will, ground rules, volatile.*

It has to be noted that the *BBI* is the most comprehensive dictionary available on lexical collocations. However, certain lexical collocations in English do not only require a V+N, Adj.+N or N+N but also call for a qualifying adjective or noun, thus forming what may be called "tripartite collocations":

(a) V+Adj.+N e.g. *to acquire a good command of a language, to have an impressive command of a language*
(b) N+N+N e.g. *World Health Organization*
(c) Adj.+Adj.+N e.g. *International Monetary Fund*

Such collocations should be included and an efficient method for looking them up should be devised.

Collocations of a different kind comprise a phrasal verb (compound or two-element verbs), e.g. *pick up / a skill, language.* These, it is suggested, may be treated like other collocations since as a special type of verb syntactically and semantically, phrasal verbs can be considered as heads or headwords that need collocates. Phrasal verbs such as *black out, screw up* and *cut down* could be presented according to the dictionary format as follows:

(a) **black out** *ph.v.* ~ news, information ["suppress"]
(b) **screw up** *ph.v.* ~ an arrangement, planning ["mishandle, mismanage"]
(c) **cut down** *ph.v.* ~ expences, consumption ["to reduce"]

In the revised edition many grammatical collocations are also included either in the entry of the new headwords introduced or as additions to the ones already treated in the first edition, e.g. *to rampage through; touched to + inf.; touched that + clause; condescending to; by computer; to tingle with; to top off with; to terminate in; to throb with; abusive to, toward; up above;* etc.

The dictionary is also provided with a "Visual Guide" to indicate how entries are structured. Such a guide, missing from the first edition, is badly needed in an unfamiliar type of dictionary such as the *BBI*. It is a successful approach to use a different typeface in the definition to make it prominent, e.g.

> **appointment** *n.* ["agreement to meet"] 1. to have; keep; make, schedule an ~ with 2. to break; cancel; miss an ~ 3. by ~ (she sees patients by ~ only) 4. an ~ to + inf. (she had an ~ to see the dean) ["selection"] 5. to confirm; make an ~ 6. to block an ~ 7. ~ to (they announced her ~ to the commission) 8. an ~ as (an ~ as professor) ["position"] 9. to have, hold; receive an ~ 10. an interim; permanent; temporary ~ 11. a political ~ ["designation"] 12. by ~ to Her Majesty

The *BBI*, revised edition, 1997

In its "Style Guide" the dictionary sticks to the use of the colon and semicolon. The use of the semicolon is confusing for the user. It is suggested that either numbers should be used, or better still, short phrases should be placed within brackets, e.g.

the revised edition:

answerable: *adj.* ~ for; to (we are ~ to our superiors for our actions)

suggested treatment:

answerable: *adj.* 1. ~ for 2. ~ to

or

answerable: *adj.* ~ for (the decision he made); ~ to (government).

Besides, the semicolon does not help the user distinguish between *synonymous* and *nonsynonymous* collocations, e.g.

booth *n.* an information; listening; phone, telephone; polling, voting; projection ~.

It is suggested that slanting lines be used, e.g.

booth *n*, an information/phone, telephone/polling, voting/projection ~.

They serve as distinct demarcation lines between synonymous or related collocations.

Grammatical collocations in the dictionary consist among others of *nineteen* English verb patterns designated by nineteen capital and small letters (see p. xxix), e.g.

abandon II *v.* (D; tr.) to ~ to (they abandoned us to our fate).

It is suggested that explicit references be used, e.g.

abandon II *v* (+object)

which is more readily seen, and more quickly comprehended and will help the user assimilate the syntax.

The dictionary is also provided with a "Practical Guide" which is essential for the user. It divides collocations into *grammatical collocations* and *lexical collocations*. Grammatical collocations are those which are governed by a dominant word. The dominant word is either a noun, an adjective or a verb and the user will look them up to find the preposition or the grammatical construction, an infinitive or *that* clause which may go with them, e.g.

(a) **adhere** *v.* 2. (d. int.) to adhere to (to ~ strictly to a plan)
(b) **affinity** *n.* 4. ~ between; for; to; with (he always felt a close ~ with the underdog)
(c) **stress II** *v.* (L) the police ~ed that all regulations would be strictly enforced

Lexical collocations differ in their structure: they may comprise V+N, Adj.+N or N+N, Adv.+V or N+V, and Adv.+Adj. Hence the noun, the verb and the adjective constitute the heads of the lexical collocations or the headwords of the dictionary. The user of the dictionary will look them up to find the *collocates* that go with them. These collocates are either nouns, adjectives, verbs or adverbs.

Since the collocates that are sought are different for grammatical and lexical collocations, it is suggested that the *BBI* be divided into two parts, one for grammatical collocations and the other for lexical ones. It will facilitate the job of the user who may look for the grammatical collocates which are dispersed across the entry (note grammatical collocations nos. 16, 17, 18, 19, 25, 31, 35 in the entry **order**).

order I *n.* ["request for merchandise or services"] 1. to give, place, put in an ~ 2. to make out, write out an ~ 3. to fill; receive, take an ~ (has the waiter taken your ~?) 4. to cancel an ~ 5. a back; mail; prepublication; rush; *shipping*; *standing* ~ 6. a side (esp. AE; in a restaurant) ~ 7. (new) ~s are falling off; are picking up 8. on ~ (the merchandise is on ~) 9. to ~ (made to ~) 10. (misc.) a tall ~ to fill ("a difficult task to carry out") ["command"] 11. to give, hand down (AE), issue an ~ 12. to carry out, execute; obey, take an ~ 13. to cancel, countermand, rescind, revoke; violate an ~ 14. a direct; executive; preservation (BE); specific ~ 15. doctor's; marching; sealed; standing; verbal; writ-

ten ~s 16. an ~ to + inf. (we received an ~ to attack) 17. an ~ that + clause; subj. (headquarters issued an ~ that the attack be/should be resumed) 18. at, by, on smb.'s ~ (by whose ~ was this done?) 19. under ~s (we were under ~s to remain indoors) ["court decree"] 20. to issue an ~ 21. an affiliation (BE); cease-and-desist; court; gag; maintenance (BE), support (AE); restraining ~ ["association, group"] 22. a cloistered; Masonic; mendicant; monastic; religious; secret ~ ["system"] 23. an economic; pecking; social ~ (he's at the bottom of the pecking ~) ["proper procedure"] 24. (a) point of ~ 25. in ~; out of ~ (the senator was out of ~) 26. to call a meeting to ~ ("to begin a meeting"; "to reestablish proper procedure at a meeting") ["state of peace"] 27. to establish; keep, maintain; restore ~ 28. public ~ ["state in which everything is in its proper place or condition"] 29. to put smt. in/into ~ 30. apple-pie, good, shipshape ~ 31. in; out of ~ (everything is in good ~; this machine is out of ~ again) ["condition"] 32. working ~ (in working ~) ["sequence"] 33. alphabetical; chronological; logical; numerical ~ 34. ascending; descending ~ 35. in; out of ~ (in ~ of importance; in alphabetical ~; these entries are out of ~) ["military formation"] 36. close; extended; open ~ ["instructions to pay"] 37. a money (AE), postal (BE) ~ ["misc."] 38. law and ~; a new ~; a new world ~; an old ~; of the ~ of (BE)/on the ~ of (AE) ("approximately"); research of the highest ~

order II *v.* 1. (C) ~ a copy for me; or: ~ me a copy 2. (D; tr.) to ~ from (to ~ merchandise from a mail-order house) 3. (d; tr.) to ~ from, out of (she ~ed him out of the house) 4. (d; tr.) to ~ off (the referee ~ed the player off the field) 5. (H) the sergeant ~ed his platoon to fall in 6. (L; subj.) the mayor ~ed that free food be/should be distributed 7. (M) the judge ~ed the prisoner to be transferred 8. (esp. AE) (N; used with a past participle) the judge ~ed the prisoner transferred to the county jail 9. (misc.) the doctor ~ed her to bed

The *BBI*, revised edition, 1997

Dividing the *BBI* into two parts, one for entries for grammatical collocations and the other for entries for lexical collocations, will make the entries shorter and more manageable for the user who may lose his bearings in his search for the grammatical collocations in long entries.

Better still, grammatical and lexical collocations could be distinctly marked either by shadowing or using different colours to indicate grammatical collocations or the closed set, and black for lexical collocations or the open set.

Finally, to exemplify the differences between the two editions (1986 and 1997) the entry **test** has been chosen.

test **I** ***n.*** **['*examination, set of questions*'] 1. to administer, conduct, give a ~ 2. to draw up, make up, set (BE) a ~ 3. to take a ~ 4. to fail; pass a ~ 5. a demanding, difficult; easy ~ 6. an achievement; aptitude; intelligence; loyalty (esp. AE); placement; proficiency ~ 7. a completion; free-association; lie-detector; multiple-choice; objective; true-and-false ~ 8. a competency; means ~ 9. *a ~ *in, on (*a ~ *in mathematics; a ~ *on new material*) ['trial, experiment'] ['examination'] 10. to carry out, conduct, do, run a ~ 11. exhaustive, extensive, thorough ~s 12. *an acid, demanding,* exacting, severe ~ 13. a blood; diagnostic; endurance; laboratory; *litmus; means; nuclear; paraffin; Pap; patch; paternity; personality;* psychological; road; saliva; screen; skin; tuberculin ~ 14. a ~ for (to do a skin ~ for tuberculosis) 15. a ~ on (they conducted a series of ~s on me at the health center) 16. (misc.) to stand the ~ of time; the ~ turned out to be positive; to put smb. to the ~**

test II *v.* **1. (D; intr., tr.) to ~ for (to ~ for excessive air pollution; to ~ the urine for sugar) 2. (D; tr.) to ~ *in (we ~ed them in* English) 3. (P; intr.) some of our students ~ed in the top percentile 4. (esp. AE) (s) some students ~ high, others low**

The *BBI*, 1st edition, 1986

test I *n.* ["examination, set of questions"] 1. to administer, conduct, give a ~ 2. to draw up, make up, prepare, set (BE) a ~ 3. to sit (for) (BE), take a ~ 4. to fail, flunk (colloq.; esp. AE) a ~ 5. to pass a ~ 6. a demanding, difficult ~ 7. an easy ~ 8. an achievement; aptitude; intelligence ~ 9. a placement; proficiency ~ 10. *a cloze; completion; multiple-choice; objective; true-false* ~ 11. a driving; road ~ 12. a breath ~ 13. a lie-detector, polygraph ~ 14. a competency; means ~ 15. a ~ in, of, on (a ~ in mathematics; a ~ on new material) ["ordeal; trial"] 16. an acid, demanding, exacting, litmus, rigorous, severe ~ 17. an endurance ~ ["experiment, trial"] ["examination"] 18. to carry out, conduct, do, perform, run a ~ 19. to have, undergo a ~ 20. exhaustive, extensive, thorough ~s 21. a blood; breathing; diagnostic; DNA; drug; PAP; patch; saliva; skin; scratch; tuberculin ~ 22. a personality; psychological ~ 23. a laboratory; nuclear ~ 24. a road ~ 25. a ~ for (to do a skin ~ for tuberculosis) 26. a ~ on (they conducted a series of ~s on me at the health center) 27. (misc.) to stand the ~ of time; the ~ turned out (to be) negative/positive; the ~ was negative/positive; to put smb. to the ~

test II *v.* 1. (D; intr., tr.) to ~ for (to ~ for excessive air pollution; to ~ the urine for sugar) 2. (D; tr.) to ~ in, on (we ~ed them in English/on their knowledge of English) 3. (P; intr.) (esp. AE) some of our students ~ed in the top percentile 4. (esp. AE) (s) some students ~ high, others low; to ~ negative/positive for a disease

The *BBI*, revised edition, 1997

Without any doubt, the 1986 *BBI* has been largely expanded and updated to be a unique treasure of English word combinations. The more I use it, the more I appreciate the effort put into it.

References

Benson, Morton, et al. 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations.* Amsterdam: John Benjamins.

Benson, Morton, et al. 1997. *The BBI Dictionary of English Word Combinations.* Revised edition. Amsterdam: John Benjamins.

Kozlowska, Christian Douglas and Halina Dzierzanowska. 1993. *Selected English Collocations.* Revised and enlarged edition. Warszawa: PWN.

LDOCE *Longman Dictionary of Contemporary English.* 1995. New edition. London: Longman.

OALD *Oxford Advanced Learner's Dictionary.* 1995. New edition. Oxford: Oxford University Press.

Mohamed H. Heliel
Department of English
Kuwait University
Kuwait

B. Kirsch, S. Skorge and N. Matsiliza. *An English–Xhosa Companion for Health-Care Professionals.* 1996, xix + 537 pp. ISBN 0-7021-3452-X. Kenwyn: Juta. Price R99,00.

Introduction

This review is an evaluation of one of the first phrase books of its kind to be used in the teaching and learning of Xhosa as a second language. The medical or health-care field needs the most effective means of communication in order to realise its mission. African languages have for a long time been neglected as a means of communication. Afrikaans and English have instead been the languages mostly used. This book is therefore assessed to establish whether it can facilitate communication between a non-Xhosa-speaking professional and the patient who speaks only Xhosa. Each chapter or section is scrutinised for its merits and demerits. Commentary on each one of these is then given. Finally the worth of the book is established.

The author's extensive experience of teaching Xhosa as both a second and a first language at various levels and for different professions informs this review.

Evaluation

As stated in the preface (p. iv) this English–Xhosa companion, compiled in close consultation with a range of health-care professionals, is much more than just a phrase book. It covers introductory exchanges such as greetings, getting acquainted, putting patients at ease and expressions for regret at one's inability to speak Xhosa. In consecutive chapters extensive vocabulary is provided for (1) obtaining patient details, (2) taking a patient's history, (3) physical examination, instructions and explanations, (4) special investigations and procedures, (5) obstetrics, (6) communicating with hospitalised and clinic patients, and (7) health education.

The preface presents the reader with an outline of the contents of the publication. This part can also be referred to as the user's guide to the companion. In this section the reader is provided with such information as the use of contrasting typefaces to highlight the different constituents of Xhosa words. There is also information about charts on pronunciation of Xhosa sounds. Readers' attention is also drawn to the presence of a reference section, an alphabetical vocabulary list and an index of medical terms and illnesses.

Acknowledgement of contributions by various individuals towards the publishing of the companion is made. Sisters Michele Rolfe and Marie McKeown showed interest in the work and thus encouraged the authors to embark on the project. Dr. Nosisa Matsiliza, Miss Theresa Soci and Dr. Bongani Mayosi are the speakers of Xhosa who provided the Xhosa equivalents for the English. Prof. Ralph Kirsch seems to have been the main force of influence behind the writing of this companion. From the acknowledgements it is clear that the work

is a product of a team of specialists. This view echoes the commentary made by Prof. Daniel J. Ncayiyana in the foreword (p. iii).

The poem "Kusa Kusihlwa" (pp. vi-ix) by E. Jobodwana from *Uncuthu maZangwa* translated into English by Sindiwe Magona is a very appropriate icebreaker to this work. It is therefore clear that the work is targeted at intermediate learners of Xhosa. Beginners will definitely be scared away by this poem. Its translator is highly commended for being so innovative in providing English equivalents.

The supplements "Some key words and phrases", "A working guide to Xhosa pronunciation" and "The clicks" on the front and back cover flaps are well placed. It is extremely important for a non-Xhosa speaker to be equipped with basic expressions in Xhosa before attempting to learn the language. Pronunciation is absolutely essential for learning the sound system of Xhosa as it is totally different from that of the European languages. The guide provided should therefore serve a very useful purpose to the learners in this regard.

The authors persuade the learners to try and memorise the key words and phrases on the front flap. One would have reservations about any hint of memorisation as it suggests bad learning habits. As these are basic phrases of Xhosa, the learners ought to be persuaded rather to try and remember applying them as they are useful icebreakers in a situation requiring communication in Xhosa.

The guide to Xhosa pronunciation provided on both the front and back flaps is very user-friendly. The authors show a lot of experience in the teaching of Xhosa as a foreign language. They know all the problems of pronunciation experienced by non-Xhosa speakers. The guide on the pronunciation of clicks is placed separately on the back flap. This is very strategic indeed as the three basic clicks are often regarded as a terror by most non-Xhosa speakers. The placing of these sounds at the end of the pronunciation guide ought to reduce the difficulty of learning them. For English speakers the explanations with examples of similar sounds in English definitely eliminates the fear of these sounds.

However, on the front page flap the use of a technical term like "coalescence" is not recommended as the health-care professionals, being the target readership, are less likely to be grammatically sophisticated. The use of less technical terms seems to have been a consideration though as the authors opted for "group" when dealing with the noun classes. This is another wise decision on the part of the compilers.

A learner who is a highly analytic, critical reader, will be well served by the section entitled "Introduction to Xhosa sentence construction". It provides extensive explanations of the grammar of Xhosa. The inclusion of ideophones and tone is another indication of the authors' awareness of the problem areas in the learning of Xhosa. Ideophones and tone are always problematic to learners.

The difficulty of providing English equivalents for Xhosa ideophones can be a nightmare to learners of this language. Clear explanations of these ideophones are given in the companion. It requires a lot of imagination to be able to come up with explanations of ideophones like "twatse = snugly" and "tyokololo

= limp" (p. 47). The English equivalents provided for these ideophones are quite apt.

Very little information is given on tone (p. 51). As regards accuracy of tone, the authors state: "Learners of Xhosa should not let concern regarding the use of the correct tone inhibit their efforts to speak. Context will help to impart the intended meaning." Whilst this ought to relieve the learners from the fear of employing incorrect intonation, one feels that more examples with tone markings would be of immense help. An explanation of the tone markings would also be necessary.

That this manual is intended to improve on relations and forge that special bond between the health-care professionals and the patients becomes clear in the section "Getting acquainted" (pp. 56-63). This section begins with the professionals courteously introducing themselves to the patient rather than asking the patient "What is your name?" — the "interrogative" approach which was used in the works of most authors, revealing a domineering attitude towards the Xhosa speakers. The authors of this work are highly commended for shifting away from this position.

Terms for professions like radiographer, dietician, etc. are too technical for a non-Xhosa speaker to try in Xhosa. Why not "ndiyi-radiographer" or "ndiyi-reyidografa"? This ought to facilitate communication better. It is preferable to move away from too formal constructions that are likely to impede communication. When enquiring about religious affiliation, for instance, asking "Leliphi ihlelo lakho lenkonzo?" (p. 68) is too formal for an ordinary Xhosa speaker. A more frequently used form is "Ungena kuyiphi icawe?" The same applies to such requests as "Nceda, khulula" (Please undress) (p. 161). It would be more effective if one says "Khawukhulule" or "Ungakhulula" (Please undress or You may undress). It is best to opt for more commonly used forms than too formal ones.

The exclamation "Heke!" (Good!) is essential for the health-care professional to know and use. This is given in the section "General examination" (pp. 160-163). It is however a pity that the companion does not incorporate more or additional exclamations made by the patients to express various other emotions. Knowing the various exclamations would help the health professional in gauging, for instance, the intensity of pain or even relief from pain where appropriate.

The sections "Parts of the body and anatomical terms" and "Amalungu omzimba" (pp. 376-391) can be regarded as being central to this work for health professionals. It seems quite logical for these professionals to want to learn Xhosa words for the different parts of the body. One would even be inclined to suggest that such lists be placed quite early in or at the beginning of a text like this companion. This is nevertheless a highly valuable section with useful and usable vocabulary. The English–Xhosa section is even more user-friendly than its Xhosa–English counterpart. Although the explanation of the locative construction is given in the English–Xhosa section, the lack of such information in the Xhosa–English list is likely to be confusing to non-Xhosa-speaking learners. It would be better to list all the nouns under the headings of the locative form

in both the English and Xhosa sections.

The vocabulary lists provided on pp. 394-499 constitute an inventory of all the words used in the companion. The translations provided, determined according to the context in which they are applied, are very helpful. The authors add a caution by stating that there might be other meanings which these words acquire in other contexts. This is necessary as most Xhosa words often change meaning according to context. Tone is another factor that changes the sense of a word in Xhosa.

Most learners of Xhosa as a second language are often thrilled when they come across such lexical entries as "imajarini" (margarien), "ifestile" (venster), "ibhulukhwe" (broek) etc. To them these words seem to give the assurance that it is not at all impossible to learn Xhosa. The authors are therefore commended for including lists of adoptives from English and Afrikaans. In the same section (pp. 498-499) they have also included lists of Xhosa words that have an Anglicised version e.g. "umandlalo"/"ibhedi" (bed).

The section "Euphemisms for some medical terms and bodily functions" (pp. 502-503) is quite informative, as it is, but it would have been complete had some Xhosa cultural explanations been provided, perhaps some elaboration on A.C. Jordan's articles in the 1961 *Cape Argus*. Some other sources could also have been consulted as regards this aspect. Its misplaced inclusion in the alphabetic index should be corrected in subsequent editions.

The alphabetic index (pp. 504-535) serves a very useful purpose. It ought to enhance the user-friendliness of this work. A health-care professional who is at the intermediate level of learning Xhosa, will definitely carry this companion in his/her coat pocket (p. iii) all the time. With the help of this index, it should be easy to search for whatever information one is looking for.

A bibliography, listing medical books and dictionaries, is to be found on pages 536-537. These references should help the learners of Xhosa to trace more useful sources of information. It is however not clear why A.C. Jordan's contribution is not included in the list.

Conclusion

This companion is a valuable contribution towards the learning of Xhosa as a second language. A lot of hard work was put into its compilation. A publication of this stature ought to stimulate a lot of interest in writing more quality works within the same area for second language learners of Xhosa.

M.W. Jadezweni
Department of African Languages
University of Stellenbosch
South Africa

William Fox and Ivan H. Meyer. *Public Administration Dictionary*, 1995. viii + 139 pp. ISBN 0 7021 3219 5. Juta. Price R69,00.

As a member of the Dictionary for Political and Associated Sciences Terms for the past 16 years and consequently having some idea of the work involved in compiling such a dictionary, I congratulate Proff. Fox and Meyer. Their openness towards comments can only be commended. However, this invites the inevitable reservations.

The book would have been more valuable if the entries (not necessarily the explanations) were translated into Zulu, Xhosa and Afrikaans, since translation often clarifies meaning and forces a rethinking of the explanation as well.

Typical South African terms are also absent, e.g. there is a wide variety of terms surrounding the basic term, *executive*, which is not in the dictionary. I also find the dictionary regional. Proff. Fox and Meyer are born and bred Southern university academics. This is apparent in the selection of terms. The term *organisation* which is preferred by the Southern universities, is used throughout the dictionary. The Northern universities prefer *institution*.

Furthermore, the dictionary presupposes Stellenbossian systems and structuralist approaches to Public Administration. The regionalism of the dictionary could lead overseas academics in Public Administration, students, members of the public, journalists, public servants, and lexicographers who use the dictionary and who are not aware of this, to think that the terms in the dictionary represent an overall picture regarding Public Administration terminology. This will create a false impression of the nature and scope of Public Administration in South Africa.

To comment on some specific terms: *Body* and *organ* (*institution, branch, section, instrument* are better alternatives) are archaic terms from the days when Herbert Spencer applied Darwin's theory of evolution to politics. *Power* and *authority* are defined as meaning the same, while there is a clear distinction between them. Power is a personal concept, perhaps even a physical one. It is a unitary concept. Power can never be written into a constitution. Authority is a structural or organisational concept which can and should be written into a constitution where it acts to prevent excesses of power on the part of officials. The President, when he acts officially, has the authority to appoint a minister. He may have the power to appoint a person he prefers, contrary to the wishes of everyone else, but to authorise the appointment, he must follow prescribed rules. Power alone is insufficient. The incorrect use of the terms *power* and *authority* is encountered in constitutions and books on politics and Public Administration which refer to the "three powers": legislative, executive and judicial. These are not powers, they are authorities, or to use the newly introduced term, *competencies* — a term used in the previous South African constitution and which should be included in the dictionary. Authority is granted or allocated to someone, and can be withdrawn; power is never granted, it is taken. A "Power of Authority" does not grant power, it grants authority. It can

be withdrawn, while power cannot be withdrawn. Authority can be shared, but power is not shared.

Terms which are central to the dictionary, *administration, public administration, public service* (which does not appear in the dictionary, although *civil service* does — but the 1993 constitution does not carry the term *civil service*) and *Public Administration* are used in a very confusing manner. In 1947 Dwight Waldo suggested that a distinction be made between Public Administration and public administration: the former the science, and the latter the practice. The definition of the term *administration* drifts off into a description of an administrative system, which has a different meaning from administration. (This, incidentally, is an example of the influence of the systems thinking of the two compilers.) The term *public administration* is described as "the executive branch of government", which is wrong. The executive branch of government can only be the Cabinet. I maintain that there is a clear difference between the public service and the executive. The public service assists the executive, the Cabinet. In the dictionary, *public administration* is used as a synonym for the term *public service.* I know this is done all over in books and lectures on Public Administration, but it leads to confusion. One sometimes comes across such utterly ridiculous writing as: "In (P)public (A)administration ..." I have heard speakers say: "Public Administration with capital letters ..." It would be better if the terms *Public Administration* and *public service* are used. We should get rid of the habit of using the term *public administration* as a replacement for *public service.* In the Preface to the book, *public administration* is used in reference to the science of Public Administration. The authors promise the reader, in their Foreword, that the *subject language* (their emphasis) of Public Administration will be dealt with, but the term *Public Administration* does not appear in the dictionary.

When a term is included in a dictionary, the compilers should take the full flow of its meaning into consideration when explaining it. In this regard, also see the truncated manner in which the terms *dualism, gatekeeper* and *meaning* were explained. *Bureaucracy* should be looked at again. The British, and their former colonies, use the term *bureaucracy* in a negative sense, indicating that public officials go beyond their authority and assume a position of power. The Europeans (the French, Italians, Germans, Dutch, etc.) use the term as we would use the term *public service,* that is, they use *bureaucracy* instead of *public service.* The Americans distinguish between positive and negative bureaucracy — the latter having the meaning which South Africans append to it. *Audi alteram partem* is incorrectly defined. It means to listen to the other party before making a decision.

The use of the term *government* should be looked at throughout the book. In the strict sense of the word, there is only one government, the Cabinet. A government governs, it does not rule. A government makes policy; the legislature sanctions it in its Acts. There is a difference between *government* and *executive.* A Government, Cabinet, decides on policy. The *Executive,* although the very same Cabinet or Government members, comes into effect after Parliament

had approved the Acts based on the proposals of the Government. Parliamentary approved policy proposals of the Government are then executed by the Executive. Furthermore, the term *government* cannot, and should not be equated with legislature, as is the case in this dictionary. Two entries should be made: *Government* (see *Executive*), and *government* (see *executive*). The reason for this is that the English use the term *government* in a broad sense; it covers more or less everything from the legislature to the public service. In Afrikaans, the term *regering* is used in its restricted, and correct sense as related to the composition, authority and actions of the government of the day, the Cabinet. *Governance* is *Regeerkunde* in Afrikaans. It is a science and needs to be re-defined. Therefore, there should be two entries: *Governance*, the art of governance; and *Governance*, the science of government.

The definition of *staff personnel*, "all the active members of an organisation", implies that there are also nonactive members. The explanation of *decision process* is similarly vague, as is that of *spoils system*. "A system of recruiting and appointing personnel in the public sector on the basis of political reasons" is *political nepotism*, which is not in the dictionary. The need for meticulous care with language in a dictionary will be dealt with later, but to point something out at this stage: one does not recruit personnel from within in the public service for appointment. It should read: "a system of recruitment of personnel for political reasons to be appointed to the public service".

I add some remarks on *Highlights on the History of Public Administration* which precedes the dictionary entries:

- Plato did not promote principles of specialisation, it was Aristotle who introduced specialisation into the academic world.
- Diocletau: The date must be AD.
- Frank Goodnow was the first Professor in Municipal Government, and the first to take Public Administration out of the classroom into the real world — he was advisor to the Chinese Government on matters relating to their public service.
- Max Weber did not introduce the bureaucratic type of organisation. It existed long before him; he merely expounded it.
- Henry Fayol is not mentioned. His *Principles of Administration*, renamed *administrative processes* by J.J.N. Cloete in 1967, was published in 1915.
- The founding of the League of Nations is not mentioned.
- The first Professor in Public Administration in Britain was Edgar Wallas, appointed in 1919 at the London School of Economics.
- Public Administration was offered at Aylesbury in England from about 1840. Senior officials to be placed in colonies were trained there.
- It is incorrect to say that Dahl was in favour of "universal principles of administration"; he was actually opposed to such principles. He argued in favour of the establishment of a "science of Administration" which could not proceed as long as a few principles were adhered to.
- Herbert Simon: the sentence contains an error. It should be "attacked" and not "attached".

- In 1947 Waldo wrote on the need for a distinction between *public administration* and *Public Administration*.
- Cloete: Public Administration has been taught in South Africa since 1920. His book, and not Public Administration, became prominent.
- Kuhn introduced the term *paradigm shifts*, not *paradigms*.

Although this is a dictionary for Public Administration, I find that the net which had been cast, is too small. If the boundaries of the public service is taken into account, then those factors which have an appreciable influence within the public service, must be added to the mix. The two authors already included economic and computer terms. Politics also play an important role within the public service, and political terms related to the public service should be included. The dictionary, somehow, reflects a strong belief in the politics–administration dichotomy, which leads to the exclusion of political terms. The politics–administration approach is clearly outdated anyway: see Marx, F.M., 1946. *Elements of Public Administration*. Englewood Cliffs: Prentice-Hall.

In a dictionary, unless unavoidable, the singular should always be used. The authors use both the singular and plural of terms, also in their explanations. E.g. *administrative institutions, conference committees, demands* and *informal groups*. See also *delegates non potest delegare* where "delegates" is plural and should be singular. The plurals "bureaux" and "groups" in the definition of *bureaucracy* "rule by bureaux or by groups of appointed officials" should also be in the singular. An example of incorrect explanation due to the use of the plural is *loosely coupled systems* which should read: "a system ... among its members", and not "systems ... among their members". The use of the plural, "systems", and "their", gives the wrong impression, that of a loose level of connectedness among different systems, which, within each one, may not be loosely bound. The correct meaning is a loose connection between members of the same system.

A dictionary of Public Administration terms should be all-encompassing rather than regional or biased towards an approach. *A dictionary should never be based on the knowledge, idiosyncracies, preferred approaches, or lived-in academic world of the compilers,* however encompassing it may be. It should be based on collective knowledge within the science, hence the term "dictionary compilers".

To conclude: at this stage the dictionary of Proff. Fox and Meyer is no more than a starting-point, and only then if meticulous care is taken with further editions.

Donavon Marais
Emeritus Professor
Public Administration
University of South Africa
Pretoria
South Africa

Publikasieaankondigings / Publication Announcements

Joey Basson, Frans van Niekerk en Kobus Grobler. *Wilde Woordeboek: Alternatiewe Nuutskeppinge deur Geesdriftige Medewerkers Landwyd*, 1997, 55 pp. ISBN 0 7993 2356 X. Pretoria: J.P. van der Walt. Prys R19,95.

Morton Benson, Evelyn Benson and Robert Ilson. *The BBI Dictionary of English Word Combinations*, revised edition 1997, xl + 386 pp. ISBN softcover 90 272 2167 7 (Eur), 1-55619-521-4 (US); hardcover 90 272 2166 9 (Eur), 1-55619-520-6 (US). Amsterdam / Philadelphia: John Benjamins. Price softcover *f*40,00 ($19,95); hardcover *f*76,00 ($38,00). (Review in this issue)

H. Crompvoets en J. Goossens met medewerking van J. van Schijndel. *Woordenboek van de Limburgse Dialecten I Agrarische Terminologie Aflevering 12*, 1998, XVI + 147 pp. ISBN 90 232 3356 5. Assen: Van Gorcum. Prijs *f*45,00.

F. de Tollenaere, bezorgd door Hans Heestermans. *Etymologica & Lexicographica*, 1997, 180 pp. ISBN 90-6412-112-5. Leiden: Internationaal Forum voor Afrikaanse en Nederlandse Taal en Letteren / Dimensie Boeken. (Leidse Opstellen 28.) Prijs *f*33,95.

Graham Elliott and Jeffrey Rowlands. *Juta's Concise Dictionary of Accounting Terms with their Afrikaans Equivalents*, 3rd edition 1996, ix + 134 pp. ISBN 0 7021 3588 7. Kenwyn: Juta. Price R59,00.

William Fox and Ivan H. Meyer. *Public Administration Dictionary*, 1995, 141 pp. ISBN 0 7021 3219 5. Kenwyn: Juta. Price R69,00. (Review in this issue)

B. Gilmour. *Glossary of Terms in Marketing Research and Related Fields*, 1996, iii + 144 pp. ISBN 0-947459-72-3. Pretoria: Bureau of Market Research, University of South Africa. (Research Report 229.) Price R11,40.

Beverley Kirsch and Silvia Skorge. *An English-Xhosa Companion for Healthcare Professionals*, 1996, xix + 537 pp. ISBN 0 7021 3452 X. Kenwyn: Juta. Price R99,00 (Review in this issue)

Dorothea Mantzel and Bernd Schulz. *Francolin Illustrated School Dictionary for Southern Africa*, 1997, xvi + 336 pp. ISBN 1-86859-015-1. Cape Town: Francolin Publishers. Price R29,95.

Nasionale Terminologiediens van die Departement van Kuns, Kultuur, Wetenskap en Tegnologie / National Terminology Services of the Department of Arts, Culture, Science and Technology. *Drukkerswoordeboek / Printing Dictionary*, Afrikaans-Engels / English-Afrikaans, hersiene uitgawe, 1995, xviii + 309 pp. ISBN 0-621-17323-1. Pretoria: Staatsdrukker. Prys / Price R20,20.

A.F. Prinsloo. *Afrikaanse Spreekwoorde en Uitdrukkings met Engelse Ekwivalente en 'n Omvattende Tweetalige Indeks*, 23ste uitgawe 1997, 358 pp. ISBN 0 627 02229 4. Pretoria: J.L. van Schaik. Prys R79,95.

Lynette van Rensburg. *Paramediese Woordeboek Afrikaans-Engels (met Verklarings in Afrikaans)*, 1996, 288 pp. ISBN 0-7986-3568-1. Pretoria: Kagiso Tersiêr. Prys R129,99. (Resensieartikel in dié uitgawe)

Leo Wanner (Editor). *Lexical Functions in Lexicography and Natural Language Processing*, 1996, xix + 355 pp. ISBN 90 272 3034 X (Eur.), 1-55619-383-1 (US). Amsterdam / Philadelphia: John Benjamins. (Studies in Language Companion Series 31.) Price *f*158,00 / $79,00.

Leo Wanner (Editor). *Recent Trends in Meaning-Text Theory*, 1997, xx + 202 pp. ISBN 90 272 3042 0 (Eur.), 1-55619-925-2 (US). Amsterdam / Philadelphia: John Benjamins. (Studies in Language Companion Series 39.) Price *f*118,00/$59,00.

Herbert Ernst Wiegand. *Wörterbuchforschung: Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*, 1. Teilband, 1998, XIX + 1162 pp. ISBN 3-11-013584-1. Berlin / New York: De Gruyter. Preis DM578,00 ($343,00).

VOORSKRIFTE AAN SKRYWERS

(Tree asseblief met die Buro van die WAT in verbinding vir 'n uitvoeriger weergawe van hierdie instruksies, of besoek ons webblad: http://www.sun.ac.za/wat/index.htm)

A. REDAKSIONELE BELEID

1. Aard en inhoud van artikels
Artikels kan handel oor die suiwer leksikografie of oor implikasies wat aanverwante terreine, bv. linguistiek, algemene taalwetenskap, rekenaarwetenskap en bestuurskunde vir die leksikografie het.

Bydraes kan onder enigeen van die volgende rubrieke geklassifiseer word:

(1) **Navorsingsartikels**: Grondige oorspronklike wetenskaplike navorsing wat gedoen en die resultate wat verkry is.
(2) **Beskouende artikels**: Bestaande navorsingsresultate en ander feite wat op 'n oorspronklike wyse oorsigtelik, interpreterend, vergelykend of kritiek evaluerend aangebied word.
(3) **Resensieartikels**: Navorsingsartikels wat in die vorm van 'n kritiese resensie van een of meer gepubliseerde wetenskaplike bronne aangebied word.

Bydraes in kategorieë (1)-(3) word aan streng anonieme keuring deur onafhanklike akademiese vakgenote onderwerp ten einde die internasionale navorsingsgehalte daarvan te verseker.

(4) **Resensies.** 'n Ontleding en kritiese evaluering van gepubliseerde wetenskaplike bronne en produkte, soos boeke en rekenaarprogramme.
(5) **Projekte.** Besprekings van leksikografiese projekte.
(6) **Leksikovaria.** Enigeen van 'n groot verskeidenheid artikels, aankondigings, praktykgerigte inligting en nuusvrystellings van leksikografiese verenigings wat veral vir die praktiserende leksikograaf van waarde sal wees.
(7) **Verslae.** Verslae van konferensies en werksessies.

Bydraes in kategorieë (4)-(7) moet almal aan die eise van akademiese geskrifte voldoen en word met die oog hierop deur die redaksie gekeur.

2. Wetenskaplike standaard en keuringsprosedure
Lexikos is deur die Departement van Onderwys van die Suid-Afrikaanse Regering as 'n gesubsidieerde d.w.s. inkomstegenererende navorsingstydskrif goedgekeur.

Artikels sal op grond van die volgende aspekte beoordeel word: taal en styl; saaklikheid en verstaanbaarheid; probleemstelling, beredenering en gevolgtrekking; verwysing na die belangrikste en jongste literatuur; wesenlike bydrae tot die spesifieke vakgebied.

3. Taal van bydraes
Afrikaans, Duits, Engels, Frans of Nederlands.

4. Kopiereg
Nóg die Buro van die WAT nóg die African Association for Lexicography (AFRILEX) aanvaar enige aanspreeklikheid vir eise wat uit meewerkende skrywers se gebruik van materiaal uit ander bronne mag spruit.

Outeursreg op alle materiaal wat in *Lexikos* gepubliseer is, berus by die Beheerraad van die Woordeboek van die Afrikaanse Taal. Dit staan skrywers egter vry om hulle materiaal elders te gebruik mits *Lexikos* (AFRILEX-reeks) erken word as die oorspronklike publikasiebron.

5. Oorspronklikheid
Slegs oorspronklike werk sal vir opname oorweeg word. Skrywers dra die volle verantwoordelikheid vir die oorspronklikheid en feitelike inhoud van hulle publikasies.

6. Gratis oordrukke en eksemplare
Skrywers ontvang vyf gratis oordrukke van elke navorsings-, beskouende of resensieartikel van hulle wat gepubliseer is asook een gratis eksemplaar van die uitgawe waarin sodanige artikel(s) verskyn het. Skrywers van suiwer evaluerende resensies en van bydraes tot die rubrieke Leksikovaria, Projekte en Verslae ontvang vyf gratis oordrukke van hulle bydraes. In laasgenoemde vier kategorieë kan die redaksie egter, afhangend van die aard en omvang van die bydraes, besluit om ook 'n eksemplaar van die betrokke uitgawe aan 'n skrywer toe te ken.

7. Uitnodiging en redaksionele adres
Alle belangstellende skrywers is welkom om bydraes vir opname in *Lexikos* te lewer en aan die volgende adres te stuur:

Die Redakteur
LEXIKOS
Buro van die WAT
Posbus 245
7599 STELLENBOSCH
Republiek van Suid-Afrika

B. VOORBEREIDING VAN MANUSKRIP

Die manuskrip van artikels moet aan die volgende redaksionele vereistes voldoen:

1. Lengte en formaat van artikels
Bydraes moet verkieslik nie 20 getikte A4-bladsye met teks in dubbelspasiëring en ruim kantlyne (ongeveer 2,5 cm) oorskry nie. Manuskrip moet verkieslik in elektroniese formaat as ASCII-teks, as volledig geformateerde Microsoft Word (DOS of Windows) lêers of as WordPerfect (DOS of Windows) lêers op rekenaarskyf (360 KB tot 1.44 MB) voorgelê word. 'n Rekenaardrukstuk van die artikel moet die skyf vergesel. Elke artikel moet voorsien wees van 'n Engelse opsomming van tussen 150 en 250 woorde, sowel as tussen 10 en 30 Engelse sleutelwoorde.

2. Grafika
Een stel duidelike oorspronklike illustrasies, tabelle, grafieke, diagramme, of kwaliteitsafdrukke daarvan, moet voorgelê word. Die plasing van grafika binne die teks moet duidelik aangedui word.

3. Bibliografiese gegewens en verwysings binne die teks
Kyk na onlangse nommers van *Lexikos* vir meer inligting.

INSTRUCTIONS TO AUTHORS

(For a more detailed version of these instructions, please contact the Bureau of the WAT or refer to our web page: http://www.sun.ac.za/wat/index.htm)

A. EDITORIAL POLICY

1. Type and content of articles
Articles may deal with pure lexicography or with the implications that related fields such as linguistics, general linguistics, computer science and management have for lexicography.

Contributions may be classified in any one of the following categories:

(1) **Research articles:** Fundamentally original scientific research that has been done and the results that have been obtained.
(2) **Contemplative articles:** Reflecting existing research results and other facts in an original, synoptic, interpretative, comparative or critically evaluative manner.
(3) **Review articles:** Research articles presented in the form of a critical review of one or more published scientific sources.

Contributions in categories (1)-(3) are subjected to strict anonymous evaluation by independent academic peers in order to ensure the international research quality thereof.

(4) **Reviews**: An analysis and critical evaluation of published scientific sources and products, such as books and computer software.
(5) **Projects**: Discussions of lexicographical projects.
(6) **Lexicovaria:** Any of a large variety of articles, announcements, practice-oriented information and press releases by lexicographic societies which are of particular value to the practising lexicographer.
(7) **Reports**: Reports on conferences and workshops.

Contributions in categories (4)-(7) must all meet the requirements of academic writing and are evaluated by the editors with this in mind.

2. Academic standard and evaluation procedure
The Department of Education of the South African Government has approved *Lexikos* as a subsidized, i.e. income-generating research journal.

Articles will be evaluated on the following aspects: language and style; conciseness and comprehensibility; problem formulation, reasoning and conclusion; references to the most important and most recent literature; substantial contribution to the specific discipline.

3. Language of contributions
Afrikaans, Dutch, English, French or German.

4. Copyright
Neither the Bureau of the WAT nor the African Association for Lexicography (AFRILEX) accepts any responsibility for claims which may arise from contributing authors' use of material from other sources.
Copyright of all material published in *Lexikos* will be vested in the Board of Control of the Woordeboek van die Afrikaanse Taal. Authors are free however to use their material elsewhere provided that *Lexikos* (AFRILEX Series) is acknowledged as the original publication source.

5. Originality
Only original contributions will be considered for publication. Authors bear full responsibility for the originality and factual content of their contributions.

6. Free offprints and copies
Authors will receive five free offprints of each of their research, contemplative or review articles published, as well as one complimentary copy of the issue containing such article(s). Authors of purely evaluative reviews and of contributions to the categories Lexicovaria, Projects, and Reports receive five free offprints of their contributions. In the case of the latter four categories, the editors may, however, depending on the nature and scope of the contributions, decide to grant the author a copy of the issue concerned.

7. Invitation and editorial address
All interested authors are invited to submit contributions for publication in *Lexikos* to:

The Editor
LEXIKOS
Bureau of the WAT
P.O. Box 245
7599 STELLENBOSCH
Republic of South Africa

B. PREPARATION OF MANUSCRIPTS

Manuscripts of articles must meet the following editorial requirements:

1. Length and format
Contributions should not exceed more than 20 typewritten A4 pages with double spacing and ample margins (about 2,5 cms). Manuscript should preferably be in electronic form on a (360 KB to 1.44 MB) floppy disk as either ASCII text, fully-formatted Microsoft Word (DOS or Windows) or WordPerfect (DOS or Windows) files. A computer printout of the article should accompany the disk. Each article must be accompanied by an English abstract of 150 to 250 words, and between 10 and 30 English keywords.

2. Graphics
One set of clear original drawings, tables, graphs, diagrams or quality prints thereof must be submitted. The locations of graphics must be clearly indicated in the text.

3. Bibliographical details and references in the text
Examine recent issues of *Lexikos* for details.

HINWEISE UND RICHTLINIEN FÜR AUTOREN

(Nehmen Sie bitte mit dem Büro des WAT Kontakt auf für eine ausführlichere Wiedergabe dieser Hinweise, oder besuchen Sie unsere Webseite: http://www.sun.ac.za/wat/index.htm)

A. REDAKTIONELLE ZIELSETZUNGEN

1. Art und Inhalt der Artikel
Es können Artikel aufgenommen werden, die sich mit Themen der Lexikographie befassen oder mit Zusammenhängen, die zwischen der Lexikographie und benachbarten Fachgebieten wie z.B. Linguistik, allgemeiner Sprachwissenschaft, Lexikologie, Computerwissenschaft und Management bestehen.

Die Beiträge sollten einer der folgenden Kategorien entsprechen:

(1) **Forschungsartikel**, die grundlegend über neue Forschungsansätze und deren Ergebnisse berichten.
(2) **Kontemplative Artikel**, die bestehende Forschungsergebnisse und andere Informationen selbständig, interpretativ, vergleichend oder kritisch bewertend wiedergeben.
(3) **Rezensionsartikel**, die in der Form eines Forschungsartikels eine oder mehrere veröffentlichten wissenschaftlichen Quellen kritisch rezenzieren.

Beiträge in Kategorien (1)-(3) werden streng anonym von unabhängigen wissenschaftlichen Experten begutachtet, um ein internationales fachliches Niveau in *Lexikos* zu gewährleisten.

(4) **Rezensionen**, die veröffentlichte wissenschaftliche Quellen und Produkte, wie z.B. Bücher und Software, analysieren und kritisch bewerten.
(5) Lexikographische **Projekte**, die vorgestellt werden.
(6) **Lexikovaria**, die unterschiedliche Beiträge, Ankündigungen, praxisbezogene Informationen und Pressemitteilungen lexikographischer Vereinigungen, die dem praktischen Lexikographen wichtig sein können, einschließen.
(7) **Berichte** über Konferenzen und Workshops.

Beiträge in Kategorien (4)-(7) müssen im akademischen Stil abgefaßt werden. Sie werden von der Redaktion unter diesem Gesichtspunkt beurteilt.

2. Wissenschaftliche Standards und das Beurteilungsverfahren
Das Erziehungsministerium der südafrikanischen Regierung hat *Lexikos* als eine subventionierte, d.h. einkommenerzeugende Forschungszeitschrift anerkannt.

Artikel werden auf Grund der folgenden Gesichtspunkte bewertet: Sprache und Stil; Sachlichkeit und Verständlichkeit; Problembeschreibung, Argumentation und Schlußfolgerung; Hinweise auf die neueste und wichtigste Literatur; wesentlicher Beitrag zum besonderen Fachgebiet.

3. Sprache der Beiträge
Afrikaans, Deutsch, Englisch, Französisch oder Niederländisch.

4. Das Urheberrecht
Weder das Büro des WAT noch die African Association for Lexicography (AFRILEX) übernehmen Verantwortung für Ansprüche, die daraus entstehen könnten, daß Autoren Material aus anderen Quellen benutzt haben.

Das Urheberrecht aller in *Lexikos* publizierten Artikel wird dem Aufsichtsrat unseres Büros übertragen. Es steht Autoren jedoch frei, ihren Beitrag anderweitig zu verwenden, vorausgesetzt, *Lexikos* (AFRILEX-Serie) wird als Originalquelle genannt.

5. Originalität
Nur Originalbeiträge werden begutachtet. Autoren tragen die volle Verantwortung für die Originalität und den sachlichen Inhalt ihrer Beiträge.

6. Sonderdrucke und Freiexemplare
Autoren erhalten fünf Sonderdrucke ihrer veröffentlichten Forschungsartikel, kontemplativen Artikel oder Rezensionsartikel gratis sowie ein Freiexemplar der betreffenden Ausgabe. Rezensenten und Autoren von Beiträgen zu den Kategorien Lexikovaria, Projekte und Berichte erhalten fünf Sonderdrucke ihrer Beiträge gratis. Die Redaktion kann sich jedoch, abhängig von der Art und dem Umfang der Beiträge der letztgenannten vier Kategorien, vorbehalten, dem Autor ein Freiexemplar der Ausgabe zu überlassen.

7. Einladung und redaktionelle Adresse
Alle Autoren, die interessiert sind, Beiträge für *Lexikos* zu liefern, sind herzlich willkommen. Sie werden gebeten, ihre Artikel an die folgende Adresse zu schicken:

Der Redakteur
LEXIKOS
Buro van die WAT
Postfach 245
7599 STELLENBOSCH
Republik Südafrika

B. VORBEREITUNG DES MANUSKRIPTS

Ein Artikelmanuskript muß den folgenden redaktionellen Anforderungen entsprechen:

1. Umfang und Format
Beiträge sollen nicht länger als 20 getippte A4-Seiten in zweizeiligem Abstand und mit Randabständen von ca. 2,5 cm sein. Das Manuskript sollte möglichst als elektronischer Text auf einer (360 KB bis 1.44 MB) Diskette vorgelegt werden, entweder im ASCII-Format, oder in formatiertem Microsoft Word (DOS oder Windows) bzw. WordPerfect (DOS oder Windows). Ein Ausdruck des vollständig formatierten Artikels soll mit der Diskette eingereicht werden. Jedem Artikel ist eine Zusammenfassung im Umfang von 150-250 Wörtern beizufügen. Ferner sollen etwa 10-30 inhaltskennzeichnende Stichwörter zu jedem Artikel angegeben werden.

2. Abbildungen
Ein reproduktionsfähiger Satz der originalen Abbildungen, Illustrationen, Tabellen, Graphiken und Diagramme oder Qualitätsabdrucke muß vorgelegt werden. Der Text selber sollte klare Hinweise auf die Position der Abbildungen enthalten.

3. Bibliographische Einzelheiten und Hinweise im Text
Zu Einzelheiten des bibliographischen Systems sind neuere Ausgaben von *Lexikos* einzusehen.