A Perspective on the Lexicographic Value of Mega Newspaper Corpora — The Case of Afrikaans in South Africa

D.J. Prinsloo, *Department of African Languages, University of Pretoria, Pretoria, Republic of South Africa (danie.prinsloo@up.ac.za)*

Abstract: The aim of this article is to assess the potential use of a mega newspaper corpus, the Media24 archive, in the absence of large balanced and representative corpora, for the compilation of major general dictionaries for Afrikaans. Firstly, an evaluation of Media24 against the lemmalists of both a major single-volume and a multi-volume monolingual dictionary for Afrikaans is undertaken to determine to what extent Media24 correlates with the lemmalists of major dictionaries. Secondly, the strength/suitability of Media24 for lemma selection in categories other than newspapers is evaluated. Finally, it is determined what the contribution could be of Media24 to lexical sense distinction, selection of examples of usage, and typical collocations.

Keywords: AFRIKAANS, MEDIA24 ARCHIVE, NEWSPAPER CORPORA, BALANCED CORPORA, REPRESENTATIVE CORPORA, WORD/LEXICOGRAPHIC/LEXICAL SENSE DISTINCTION, LEMMALIST, CORPUS DESIGN

Opsomming: 'n Perspektief op die leksikografiese waarde van megakoerantkorpusse — Die geval van Afrikaans in Suid-Afrika. Die doel van hierdie artikel is om die bruikbaarheid van 'n megakoerantkorpus, die Media24-argief, te bepaal in die afwesigheid van groot, gebalanseerde en verteenwoordigende korpusse vir die samestelling van omvattende algemene woordeboeke vir Afrikaans. Eerstens word 'n evaluering van Media24 gedoen deur dit met die lemmalyste van 'n groot omvattende enkelvolume-, en 'n multivolumewoordeboek van Afrikaans te vergelyk, ten einde te bepaal tot watter mate Media24 met die lemmalyste van groot woordeboeke korreleer. Tweedens word die gewig/toepaslikheid van Media24 vir lemmaseleksie uit kategorieë wat koerante uitsluit, geëvalueer. Ten slotte word bepaal wat die bydrae van Media24 kan wees tot leksikale betekenisonderskeiding, keuse van gebruiksvoorbeelde en tipiese kollokasies.

Sleutelwoorde: AFRIKAANS, MEDIA24-ARGIEF, KOERANTKORPUSSE, GEBALANSEERDE KORPUSSE, VERTEENWOORDIGENDE KORPUSSE, WOORD/LEKSIKOGRAFIESE/LEKSIKALE BETEKENISONDERSKEIDING, LEMMALYS, KORPUSONTWERP

Introduction

The aim of this article is to assess the contribution that a mega newspaper corpus, in the absence of large balanced and representative corpora, can make to dictionary compilation of major general dictionaries for Afrikaans.

Afrikaans lexicography finds itself in a situation where (a) a number of excellent major dictionaries are available (not traditionally based on corpus material), (b) no large balanced and representative corpora exist, but (c) a mega newspaper archive estimated at 1 000 000 000 (a thousand million tokens) can be consulted.

In this article, the Afrikaans Media24 archive is subjected to three tests in order to determine its effectiveness for the compilation or review of major Afrikaans general dictionaries.

The first test is an evaluation of the newspaper corpus on a macrostructural level against the lemmalists of a modern, major monolingual dictionary for Afrikaans, i.e. the 5th edition of the *Handwoordeboek van die Afrikaanse Taal* (HAT) and the 4th volume of the multi-volume *Woordeboek van die Afrikaanse Taal* (WAT). The intention is to establish to what extent this media archive can be used as a source for inclusion versus omission of lemmas in the revision (or compilation) of major Afrikaans dictionaries. Firstly, lemmas in HAT and WAT are compared to Media24 in order to determine to what extent a word list culled from Media24 matches the existing lemmalists of a single volume of a major dictionary such as HAT and a multi-volume comprehensive dictionary equal to the WAT. Secondly, an attempt is made to determine to what extent a word list culled from Media24 is suitable as an aid to inclusion or omission in future versions of these dictionaries. The question is, therefore, whether such a word list indicates what lemmas could be added to current lemmalists and whether non-occurrence could suggest the need for the omission of certain lemmas from existing dictionaries. Finally, it is suggested that frequency counts over three decades can assist the lexicographer to decide on inclusion or omission.

The second test evaluates the suitability of Media24 for lemmas most likely to be looked for by the target users of a general dictionary in categories not intensively covered by newspapers, for instance, religion, skills, hobbies, government, house organs and fiction which is covered in the BROWN/LOB corpora but collected as separate categories (cf. Table 1 below). Newspapers report on these fields, but the question is whether the coverage of such items is sufficient for lexicographic purposes. The purpose is therefore to determine to what extent terms from these fields, which can be presumed not to be generally associated with newspaper reporting, are covered by the Media24 newspaper archive. The randomly selected categories are gardening, quilting and embroidery. The last two contain precise subject specific terminology and therefore pose an implicit challenge in terms of coverage by a general newspaper corpus.

The third test, on the level of the microstructure of dictionaries, aims to determine the value of Media24 as an aid in sense distinction, selection of examples of usage, and typical collocations. The question is whether a presumed bias towards typical 'newspaper senses' versus more 'general senses' impedes the value of Media24 in comparison to general corpora.

A brief description of WAT, HAT and Media24 will be given, followed by a calculation of the size of the Media24 archive.

Balance and representativeness as essential but problematic aspects in corpus creation

The debate as to what entails valid/ideal/balanced/representative corpora and whether it will ever be possible to compile such corpora is ongoing (cf. Biber 1993, Summers 1993, Kilgarriff 1997, Kennedy 1998, Kruyt and Dutilh, 1997, Otlogetswe 2007 and Atkins and Rundell 2008 for detailed discussions). A few excerpts serve to illustrate these lexicographic concerns.

> Questions associated with 'representativeness' and 'balance' are complex and often intractable. (Kennedy 1998: 62.)

> A general corpus is typically designed to be **balanced**, by containing texts from different genres and domains of use including spoken and written, private and public […] For a corpus to be 'representative' there must be a clearly analysed and defined population to take the sample from. (Kennedy 1998: 20, 52.)

> What we mean by *representative* is covering what we judge to be the typical and central aspects of the language, and providing enough occurrences of words and phrases for the lexicographers […] to believe that they have sufficient evidence from the corpus to make accurate statements about lexical behaviour. (Summers 1993: 186, 190.)

> […] to be representative of general language. This is a bold ambition — some say one that is impossible to fulfil. (Summers *s.d.* [1996–1998]: 6.)

> COBUILD have always insisted that it is impossible to create a corpus that is truly representative of the language, and have focused on size of corpus rather than balance. (Kilgarriff 1997: 150.)

> Lexicographers traditionally aim at a 'representative' or 'balanced' corpus, that is, the corpus should be appropriate as the basis for generalizations concerning the language as a whole. (Kruyt and Dutilh 1997: 230.)

Scholars even differ in their interpretation of the terms. This debate, however, is beyond the scope of this article — the issue at stake here is simply whether a 1 000 million-word newspaper archive can be regarded as a suitable, main source for the compilation of major Afrikaans dictionaries.

The design of a pioneering corpus, such as the *Brown Corpus of Standard American English* and *Lancaster-Oslo/Bergen Corpus* (LOB), was a carefully compiled selection of American English, totalling approximately a million words drawn from a wide variety of sources for which each contained 2 000 words. The corpus was sampled from 15 text categories given in Table 1.

PRESS: REPORTAGE (44 texts)	LEARNED (80 texts)
PRESS: EDITORIAL (27 texts)	FICTION: GENERAL (29 texts)
PRESS: REVIEWS (17 texts)	FICTION: MYSTERY (24 texts)
RELIGION (17 texts)	FICTION: SCIENCE (6 texts)
SKILLS AND HOBBIES (36 texts)	FICTION: ADVENTURE (29 texts)
POPULAR LORE (48 texts)	FICTION: ROMANCE (29 texts)
BELLES-LETTRES (75 texts)	HUMOR (9 texts)
MISCELLANEOUS: GOVERNMENT & HOUSE ORGANS (30 texts)	

Table 1: Design of the Brown and LOB corpora

It could be assumed that newspaper texts represent a specific, almost homogeneous subtype that can easily skew a balanced corpus if newspaper texts are added in large quantities (cf. MacLeod and Grisham (2000) for the case of adding a vast amount of newspaper data to the Brown Corpus). They indicate how an increase in the Brown Corpus of 1 329% (thus more than thirteen times) resulted in a skewed or inadequate corpus e.g. in the representation of business-related words, such as *sell*, *rise*, *buy*, *pay*, and *increase*. Newspaper texts also contain words belonging to a slightly higher register; cf. *arts* (instead of *dokter/geneesheer*) 'doctor', *baar* (*kraam/geboorte gee*) 'give birth'. On the other hand, these texts also contain words belonging to an informal register; cf. for example *herrie* (*oproer/rusie/ontevredenheid*) 'uproar/quarrel/dissatisfaction', *grondgryp* (*grondonteiening*) 'land seizure', and *blaser* (*skeidsregter*) 'referee'. Both these types are uncommon to everyday written and oral communication. The use of such words on newspaper banners or in headlines contribute to attracting attention and are usually shorter than their equivalents, fitting into limited space. They are apparently much less frequently used in non-newspaper corpora. Preliminary tests indicate that *herrie* is used ten times more in Media24 than in a 4 million-token test corpus consisting of Afrikaans literary works. Likewise no occurrence of *blaser* referring to a referee could be found in the test corpus. The real potential corpus-skewing factor of such words should however be determined by more detailed studies.

It could also be argued that a growing newspaper corpus, such as Media24, partially qualifies for what Atkins calls an *organic corpus*, at least as far as the 'growing part' is concerned.

> A corpus builder should first attempt to create a representative corpus. […] the corpus is enhanced by the addition or deletion of material [...] This is the way to approach a balanced corpus. One should not try to make a comprehensive and watertight listing […] rather, a corpus may be thought of as organic, and must be

> allowed to grow and live if it is to reflect a growing living language. (Atkins 1997, personal communication at Salex'97 (Atkins et al. 1997.))

Building neatly designed corpora, such as the Brown corpus, was also envisaged for African languages and Afrikaans when corpus creation for these languages commenced in 1990. For the African languages, it was not possible, because many of the categories, such as the three press sections, simply do not exist as most of the languages do not even have a single newspaper and some in fact have very limited printed matter. For these languages, a more organic approach (cf. Atkins et al. 1997) was followed. For Afrikaans, the situation was more conducive, but no attempt was ever made to build a large corpus, for example, along the lines of the Brown/LOB design. An organic corpus of 10 000 000 tokens was compiled at the University of Pretoria but this corpus is dwarfed by the Media24 newspaper archive estimated at more than 1 000 million tokens.

The question remains, however, as to what extent growing in size also means growing in representativeness or, in what Leech terms its *diversity*.

> The value of a corpus as a research tool cannot be measured in terms of brute size. The **diversity** of the corpus, in terms of the variety of registers or text types it represents, can be an equally important (or even more important) criterion. (Emphasis in the original.) (Garside et al. 1997: 2.)

> Regardless of the corpus size, a corpus that is systematically selected from a single register cannot be taken to represent the patterns of variation in an entire language; […] corpora representing the full range of registers are required. […] it is important to design corpora that are representative with respect to both size and diversity. However, given limited resources for a project, representation of diversity is more important for these purposes than representation of size. (Biber 1995: 131.)

What is important, therefore, is to estimate the value of the Media24 archive for Afrikaans lexicography. Is its 'brute size' also representative of the varieties of registers or non-newspaper categories in, for example, the design of the Brown Corpus?

The Media24 archive

The Media24 archive is a searchable database of Afrikaans media reports available at http://152.111.1.251/cgi-bin/s.cgi. Media24 contains among others the newspapers *Rapport*, *Beeld*, *Volksblad* and *Die Burger*, available in electronic format for the past two to three decades. A range of search functions such as basic words, fixed and semi-fixed phrases as well as the use of certain Boolean operators are allowed. Hits are presented as full reports as they were published in the newspapers, up to a maximum of 50 at a time. It also means that a report can contain more than one occurrence of a word. As for the size of Media24, no authoritative figure is available. Evaluation of Media24 frequencies is therefore difficult if the size of the corpus is unknown. An attempt was made to calculate its approximate size in a simplistic way before comparisons with HAT and WAT were made.

A random selection of 18 words was chosen for the calculation of the size of the Media24 archive given in Table 2. Statistics used for this calculation were

(a) counts in a 750 million-word subsection of the corpus (exact size: 749 553 152 tokens); Column 2,

(b) number of newspaper reports containing each of these 18 words in the 750m subcorpus; Column 3, and

(c) number of newspaper reports in the entire archive containing each of these 18 words; Column 5.

First, the relation between the number of media reports in which a specific word occurs and the total number of occurrences of the word in all reports in the 750m subcorpus was calculated; Column 4. In the case of *die*, for example, the number of reports is only 6% of the total counts for *die*, i.e. *die* occurs very frequently in each report (more than 50 million times in less than 3 million reports).

This relation was then used to calculate the total count of each word in the entire Media24 archive based on the number of reports in the entire archive; Column 6.

A basic correlation value between counts in the 750m subcorpus and the total size 750m was then calculated for each word in the 750m subcorpus by dividing 750m with the total counts for each word; Column 7.

This correlation value was finally used to calculate the size of the Media24 archive by multiplying it with the calculated total counts in Column 6.

Thus for all of the 18 keywords, a corpus size slightly exceeding 1 000 000 000 (one thousand million tokens) was independently postulated; Column 8.

1	2	3	4	5	6	7	8
Word	**750m Subcorpus counts**	**750m Subcorpus reports**	**750m Reports as % of counts**	**Media24 Reports**	**Media24 Calculated total counts**	**750m Correlation value**	**Media24 Calculated size**
boek 'book'	100 787	55 096	55	85 881	157 102	7 437.00	1 168 367 472
die 'the'	51 184 148	2 889 785	6	4 079 349	72 253 819	14.64	1 058 102 558
drink	25 450	24 524	96	44 692	46 380	29 451.99	1 365 969 233
eet 'eat'	41 741	35 329	85	58 660	69 306	17 957.24	1 244 552 291
en 'and'	16 741 151	2 606 098	16	3 718 038	23 884 073	44.77	1 069 363 893
het 'has'	16 870 704	2 443 463	14	3 449 872	23 819 378	44.43	1 058 277 711
hond 'dog'	30 571	19 020	62	30 500	49 023	24 518.44	1 201 964 834
huis 'house'	360 993	226 869	63	345 317	549 467	2 076.36	1 140 893 845
kat 'cat'	17 614	13 384	76	20 797	27 370	42 554.40	1 164 708 376
loop 'walk'	132 839	132 328	100	200 486	201 260	5 642.57	1 135 624 458
mens 'human'	413 683	280 379	68	432 056	637 474	1 811.90	1 155 039 916

Table 2: Calculation of the size of the Media24 archive

HAT and WAT

HAT is the 5th edition of the *Verklarende Handwoordeboek van die Afrikaanse Taal* containing more than 50 000 lemmas. WAT, *Woordeboek van die Afrikaanse Taal*, is a multi-volume explanatory dictionary currently published up to the letter R (13 volumes).

Comparison of types in the Media24 archive to the lemmalists of HAT and WAT

For the first test, a random sub-stretch of the arbitrarily selected alphabetical stretch 'I' was selected i.e. *ideaal* to *idioot*. There are 153 lemmas strictly alphabetical[1] in this stretch in WAT and HAT taken together. WAT has 147, HAT 48 and they have 42 in common.

Lemmas given in both WAT and HAT with overall counts in Media24:

ideaal (31 557), **idealis (298)**, **idealiseer (163)**, **idealisme (1 390)**, **idealisties (673)**, **idealiteit (2)**, **idee (58 214)**, ideëassosiasie (0), **ideëel (8)**, **idée-fixe (2)**, ideëleer (0), **ideëryk (7)**, ideëverering (0), **ideëwêreld (13)**, **idem (167)**, **identiek (7)**, **identies (2 345)**, **identifikasie (3 185)**, **identifiseer (25 680)**, **identiteit (25 053)**, **identiteitsbedrog (54)**, **identiteitsbewys (45)**, **identiteitskaart (420)**, **identiteitsplaat(jie) (13)**, **ideo- (142)**, ideofoon (0), **ideografie (2)**, **ideogram (7)**, **ideolatrie (2)**, **ideologie (5 898)**, **ideoloog (231)**, ideomotories (0), **idille (203)**, **idillies (312)**, **idio- (13)**, idiolatrie (0), **idiomatiek (37)**, **idiomaties (302)**, idiomorf (0), idiomorfie (0), **idioom (2 838)**, **idioot (923)**

Lemmas given in WAT but not in HAT, with overall counts in Media24:

ideageen (0), ideasie (0), ideatief (0), **ideëdrama (7)**, ideëfonds (0), **idee-force (128)**, ideëgeskiedenis (0), ideëkuns (0), ideëliriek (0), ideëpoësie (0), ideerigting (0), ideëskat (0), ideëskrif (0), **ideëspel (1)**, ideëtragiek (0), ideëvlug (0), ideëwaarde (0), idemfaktor (0), idempotent (0), **identifieer (2)**, identifiëring (0), identifiëringsparade (0), identifikasiebaken (0), **identifikasiebewys (4)**, identifikasiekaart(jie) (0), identifikasieletter (0), identifikasielig (0), **identifikasieparade (2)**, identifikasieplaat(jie) (0), **identifikasiesein (2)**, **identifikasieteken (1)**, **identifisering (3 534)**, identifiseringsparade (0), identiteitsafstand (0), **identiteitsbeginsel (2)**, identiteitsbrief (0), identiteitselement (0), identiteitsfilosofie (0), identiteitshipotese (0), identiteitsisteem (0), identiteitskenmerk (0), identiteitskyf (0), identiteitsmatriks (0), **identiteitsmerk (2)**, identiteitsoordeel (0), **identiteitsparade (16)**, identiteitsprinsipe (0), identiteitsreaksie (0), identiteitsteken (0), **identiteitstelsel (18)**, **identiteitsteorie (1)**, identiteitswet (0), ideofreen (0), ideofrenie (0), ideogeen (0), ideogenese (0), **ideograaf (2)**, ideografies (0), ideokineties (0), ideologieëleer (0), **ideologies (2 065)**, ideometabolies (0), ideometabolisme (0), ideomosie (0), ideomotoriek (0), ideomuskulêr (0), ideoplasie (0), ideoplastie (0), ideoplastiek (0), ideoplasties (0), ideorefleksie (0), ideosekretories (0), ideosensories (0), ideovaskulêr (0), ideovisueel (0), Idiacanthidae (0), Idiacanthus (0), idioadaptasie (0), idiobiologie (0), idioblas(t) (0), idioblasties (0), idiochromaties (0), idiochromatine (0), idiochromidie (0), idiochromosoom (0), idiofonie (0), idiofreen (0), idiogaam (0), idiogamie (0), idiogeen (0), idiogenese (0), idioglossie (0), idioglotties (0), idiograaf (0), idiografies (0), idiogram (0), idiohipnose (0), idioïmbesiel (0), idiokinese (0), idiokineties (0), idiokrasie (0), idiolalie (0), idiomatologie (0), idiomorfies (0), idiomuskulêr (0)

Lemmas given in HAT but not in WAT with overall counts in Media24:

ideëberaad (3), **identikit (1 391)**, **identiteitsdokument (3 124)**, **identiteitskrisis (831)**, idioëlektries (0), **idiolek (56)**

Media24 reflects counts for 55 of these 153 lemmas. In comparison to WAT, Media24 shows counts for 50 lemmas, i.e. roughly 30% and in comparison to HAT Media24 reflects 39 lemmas, i.e. 80%. Thus the value of Media24 for the compilation of a lemmalist for a multi-volume dictionary of the magnitude of WAT is substantially lower than for a single-volume major dictionary. The lemmas presented in WAT cover 165 993 tokens in Media24 and HAT covers 165 611. This is quite significant, i.e. that HAT, although having only one third of the lemmas compared to WAT, covers the same number of tokens in Media24. HAT fared well in comparison to WAT for lemmatising *identiteitsdokument* (3 124) 'identity document', *identikit* (1 391), *identiteitskrisis* (831) 'identity crisis' and *idiolek* (56) 'idiolect', which reflect high counts in the corpus. WAT on the other hand did well in comparison to HAT for lemmatising *identifisering* (3 534) 'identification', *ideologies* (2 065) 'ideological', *idee-force* (128) 'active idea', *identiteitstelsel* (18) 'identity system' and *identiteitsparade* (16) 'identity parade' which show high counts in the corpus.

From these comparisons, it is clear that the Media24 archive not only covers all frequently used lemmas in the dictionaries but also a significant number of low frequency lemmas. Some lemmas in the dictionary with zero occurrences in the archive could therefore be considered for omission in a forthcoming revision of the dictionary. Likewise, certain words in the archive could be considered for inclusion in the dictionary (given general considerations for lemma inclusion such as the self-explanatory nature of some morphologically complex words), such as *identiteitloos* (67) 'without identity', *identiteitloosheid* (21) 'state of being without identity', *identiteitlose* (62) 'being without identity', *identiteitsboek* (164) 'identity book', *identiteitsboeke* (125) 'identity books', *identiteitsboekie* (409) 'small identity book', '*identiteitsboekies* (323) 'small identity books', *identiteitsfoto* (20) 'identity photo', *identiteitsfoto's* (48) 'identity photos', *identiteitsnommer* (690) 'identity number', *identiteitsnommers* (255) 'identity numbers'. These words could be lemmatised as *identiteitloos*, *identiteitsboek*, *identiteitsfoto* and *identiteitsnommer*. In order to gain an impression of their frequency trajectories over two decades, the total counts of these words are expressed per 50 million tokens for five-year periods ending in 1989, 1994, 1999 and 2003[2] respectively in Table 3 and graphically illustrated in Figure 1.

	89/50M	94/50M	99/50M	2003/50M
identiteitloos (lemma)	3	8	10	13
identiteitsboek (lemma)	29	85	84	51
identiteitsfoto (lemma)	3	4	5	4
identiteitsnommer (lemma)	52	47	73	67

Table 3: Total counts (lemmas and derivations) expressed per 50 million tokens in Media24

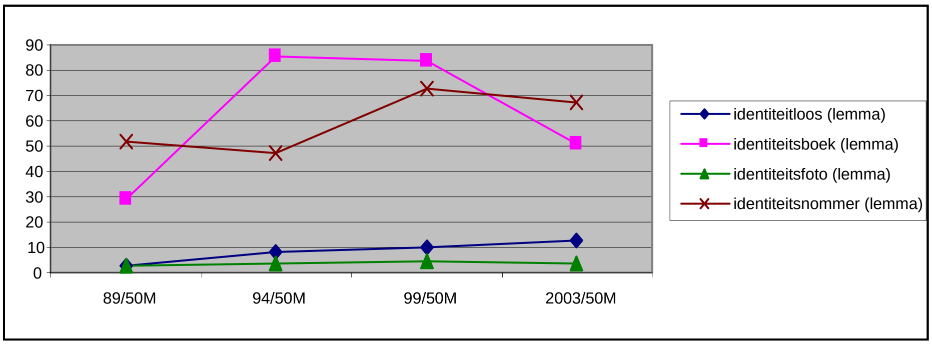

Figure 1: Trajectories of the total counts (lemmas and derivations) expressed per 50 million tokens in Media24

The frequency counts in Table 3 and the trajectories in Figure 1 suggest inclusion of these lemmas in major Afrikaans dictionaries.

The repetition factor in Media24

Frequency counts of words in a media corpus can be questioned on the basis of potential repetition of the same phrases in, for example, regional releases of the same reports or stereotypical repetitions of a word/phrase. In order to determine the extent and nature of repetition in Media24, concordance lines were generated for randomly selected words given in Table 5. For example, up to 53 repetitions of the line "*... Die woorde hieronder kom voor in die* ***blok*** *met letters ...*" '... The words below occur in the block with letters ...' occur as indicated in Table 4. The total number of concordance lines generated for each word from reports in Media24 were grouped and summed to determine the number of duplications. Consider the following instances of repetition of fixed phrases containing the word *blok* 'block' in Table 4.

Concordance lines for *blok*			**Number of repetitions**
Die woorde hieronder kom voor in die	**blok**	met letters. Hulle verskyn horisont	53
Nasionale Pers. Tenderaars moet die	**blok**	, die getal en die prys verstrek van	32

oop, of aandui dat die tender uit enige	**blok**	toegeken kan word. 'n Minimum va	23
estasies buite skoolgebied Verwys na	**Blok**	A: Teen die einde van die jaar nee	20
p die foto begrawe lê. Dui u die regte	**blok**	aan en u inskrywingsvorm is een v	12
inasies het om te wen. Indien jy 'n vol	**blok**	het (met ander woorde al die blokk	11
word. Sodra u al die nommers in die	**blok**	omkring het, kan u een van die pry	9

Table 4: Most frequent repeated concordance lines for *blok* in Media24

Concordance lines for six more randomly selected words were generated, grouped, summed and studied (cf. Table 5). The final column of Table 5 indicates that the percentage of duplication for these words range from 5%–14%.

Word	**Total lines M24**	**Different lines**	**%**	**% duplication**
identifisering 'identification'	2 333	1 997	86	14
ideologies 'ideological'	1 381	1 155	84	16
identiteitsnommer 'identity number'	691	620	90	10
identiteitsboekie 'identity book'	410	365	89	11
blok 'block'	8 905	7 427	83	17
borduur 'embroider'	562	521	93	7
steek 'stitch'	16 001	15 200	95	5
Average				**11**

Table 5: Duplication factor of words in Media24

Reports are repeated in sister newspapers, regional issues of the same newspaper or sequentially over a period of time. So, for example, concordance lines generated for *ideologies* 'ideological' rendered a number of identical lines for *ideologies* in the context *Die "ou manne" wat op alle samelewingsvlakke fanaties, krampagtig en ideologies apartheid bedink en bevorder het, moet gekonfronteer word* (The 'old men' who on all levels of coexistence, fanatically, desperately and ideologically conceptualised and promoted apartheid should be confronted) (*Beeld* 29 December 2000, *Die Burger* 30 December 2000, 2 October 2000 and *Volksblad* 29 September 2000). From a strict statistical point of view, it could be argued that these are repetitions skewing frequency counts. However, from a lexicographic perspective, a case could be made for *bona fide* use in multiple sources over large and different geographic areas, e.g. *Die Burger*, mostly southern regions of South Africa, *Beeld* mostly northern regions, etc., i.e. not true/basic repetition.

Evaluation of Media24 for categories *gardening, quilting* and *embroidery*

For the second test, the categories *tuinmaak* 'gardening', *laslappie* 'quilting' and *borduur* 'embroidery' were randomly selected as representatives of the categories Skills and Hobbies (cf. Table 1 above). These categories contain precise subject-specific terms and therefore pose a challenge in terms of coverage by a general newspaper corpus. Firstly, dedicated corpora were compiled for each of these categories from randomly selected sections of gardening (49 146 tokens); *Die Suid-Afrikaanse tuin* (Gilbert 1985); embroidery (64 968 tokens); *Borduursteke vir Suid-Afrika* (Eaton 1989) and quilting (28 131 tokens); *Die Suid-Afrikaanse boek van laslappie en appliekwerk* (Turpin-Delport 1988). The Gardening Corpus and a combination of the Quilting and Embroidery Corpus were then compared to a 5.8-million token general Afrikaans corpus using the Keyness function of WordSmith Tools. The aim was to detect the so-called *positive keys*, i.e. words used in the Gardening, Quilting and Embroidery corpora that occur more frequently than expected in comparison to a general corpus.

N	WORD	KEYNESS
1	NAT ('wet')	1 947.5
2	PLANT	1 464.6
3	SNOEI ('prune')	1 367.9
4	PLANTE ('plants')	1 327.5
5	MAAL ('grind')	1 138.4
6	KOMPOS ('compost')	1 085.9
7	BESPUIT ('spray')	940.0
8	GROND ('soil')	831.1
9	KYK ('see')[3]	767.6
10	BEMES ('fertilise')	737.3
11	DEKLAAG ('upper layer')	727.8
12	SAAI ('sow')	694.2
13	EENJARIGES ('annuals')	681.0
14	DAE ('days')	676.5
15	MAAND ('month')	657.9
16	BLARE ('leaves')	594.3
17	RADYSE ('radishes')	551.4
18	MAAK ('make, prepare')	530.4
19	KAN ('can, may')	527.2
20	SAAD ('seed')	510.2

Table 6: The top 20 keys for *gardening*

All words in Table 6 typically used in the subject field of gardening occur frequently in Media24, i.e. *kompos* (2 601), *bemes* (612), *plant* (30 777), *snoei* (1987), *deklaag* (602), *saai* (13 762), *eenjariges* (155), *blare* (10 918), *saad* (8 708) and *bespuit* (1 845). Gardening Keys occurring more than once in the Gardening corpus but not occurring in Media24 were studied in more detail. The 225 words were lemmatised rendering 216 lemmas. Of these, 23 lemmas were found in WAT. For words occurring three times or more in the Gardening Corpus, WAT has the lemmas *poeiermeeldou*, *Heuchera*, *druifhiasint*, *Helianthemum*, *naelkruid*, *pondolandklimop*, *Cryptostegia*, *Holmskioldia*, *Habranthus*, *drakebloedboom*, *blouslangkop* and *klokblom*. For words occurring two or more times, HAT has the lemmas *wolfsboontjie*, *suurdoring*, *sonrosie*, *skeefblom*, *reseda*, *klokblom*, *kardinaalsmus*, *eschscholtzia* and *bordeauxmengsel*. Thus, for the treatment of these lower ranking Gardening Keys lemmatised in WAT and HAT, the lexicographer cannot count on Media24 for support.

N	WORD	KEYNESS
1	GEWERK ('worked')	10 769.2
2	STEEK ('stitch')	10 543.0
11	VERTIKALE ('vertical')	2 463.9
12	HORISONTALE ('horisontal')	2 405.7

3	STEKE ('stitches')	7 140.9	13	VORM ('form')	2 292.4
4	WORD ('be (done)')[4]	4 934.5	14	KRUISSTEEK ('cross stitch')	2 223.5
5	RYE ('rows')	3 464.5	15	GAAS ('canvas')	2 142.9
6	GARING ('cotton')	3 299.4	16	KETTINGSTEEK ('chain stitch')	2 125.8
7	EWEDRAADSTOF ('even weave')	3 030.0	17	P ('p.')[5]	2 075.1
8	CM	2 805.5	18	VULLING ('filling')	1 866.2
9	EFFEBINDING ('simple (plain) weave')	2 761.2	19	GEBRUIK ('use')	1 856.7
10	STOF ('material')	2 647.7	20	BL. ('p.')[6]	1 781.0

Table 7: The top 20 keys for *quilting* and *embroidery*

With the exception of *effebinding* (0), the other Quilting and Embroidery Keys were found in Media24, i.e. *steke* (3 459), *ewedraadstof* (2), *garing* (319), *kruissteek* (51), *kettingsteek* (11), *gaas* (191) and *vulling* (170). Quilting and Embroidery Keys occurring more than once in the Quilting and Embroidery Corpus but not occurring in Media24 were studied in more detail. The 473 words were lemmatised rendering 397 lemmas. Of these, 66 lemmas were found in WAT. For words occurring 10 times or more in the Quilting and Embroidery Corpus, WAT has the lemmas *effebinding*, *lynsteek*, *knoopsteek*, *komberssteke*, *graatsteek*, *glansgaring*, *legdrade*, *lussteek*, *applikee*, *kabelsteek*, *koraalsteek*, *buitelynsteek*, *legwerk*, *krielsteek*, *gekriel*, *kussingsteek*, *komberssteekvulling* and *diamantsteek*. For words occurring three times or more, HAT has the lemmas *visgraatsteek*, *veersteek*, *sluitsteek*, *siersteek*, *randsteek*, *oormekaarslaan*, *lussteek*, *knoopsgatsteek*, *kabelsteek* and *hegsteek*. As in the case of lower ranking Gardening Keys, these Quilting and Embroidery Keys lemmatised in WAT and HAT are not supported by Media24 texts.

Evaluation of the media corpus on microstructural level

The third test aims to determine whether concordance lines culled from a major media corpus could provide sufficient aid to the compiler of a major dictionary on a microstructural level. The focus is on the contribution towards sense distinction, authentic examples and collocations — all typically regarded as areas where the corpus gives valuable support to the lexicographer in compilation of the article (cf. De Schryver and Prinsloo 2000). The polysemous words *borduur*, *steek*, *patroon*, *knop* and *blok* were selected from the Quilting and Embroidery list and their treatment in HAT and WAT as well as their use in context in Media24 was studied.

HAT

borduur ww. (geborduur) [...] **1** Met naaldwerk versier: *Blomme op 'n kussing borduur.* **2** *(fig.)* Op oordrewe wyse opsier: *'n Verhaal borduur met romantiese verdigsels.*

borduur: ~draad, ~gaas, ~garing, ~kant, ~naald, ~patroon, ~raam, ~ster, ~werk, ~wol.

WAT

borduur

I s.nw. Borduurkatoen of -wol.

II ww. **1.** Bestaande stowwe met naaldwerk versier i/d vorm van rande, figure, festoene, ens. — in teenst. met weef, tapytwerk, ens., waarby dieselfde stof se drade gebruik word: *'n Geborduurde kleedjie*. **2.** (*fig.*) Opsmuk op 'n oordrewe, verdigtende manier; uitbrei: *Die skrywer het verder daarop voortgeborduur.*

In the case of *borduur*, both HAT and WAT distinguish two basic senses, i.e. *met naaldwerk versier* 'decorate with needlework' and the figurative meaning *op oordrewe wyse opsier* 'elaborate in an exaggerated way'. The word *borduur* occurs 562 times in the Media24 archive. A random pick of 10 occurrences, i.e. every tenth line (keyword-in-context (KWIC) lines 10, 20, 30, 40, … 100) are given in Table 8.

10	En op hierdie fototentoonstelling **borduur** sy voort op dié onderwerp, met 'n
20	57) 352-9211 gerig word. Wedstryde **borduur** dwelms ...
30	051) 522-2130. Foto: Michelle Cahill **borduur** handwerk ...
40	het aan BBC Sport gesê: "Robinson **borduur** voort op Engeland se reputasie
50	am net in Engels agter op die trui te **borduur** . Die sterk Afrikaanse ondersteun
60	Los Angeles. Na die Grammys toe, **borduur** ek. "What are you going to do ther
70	borduur vir baba.Gesinsafdeling: **borduur** vir baba Arina du Plessis Maak ?
80	rksessies in Australiese kruissteek, **borduur** op wol, kralewerk op klere en eksp
90	wat jou hart begeer," sê hy. En hy **borduur** voort op wat hy haar sal gee en
100	gesorg, maar DEON GELDENHUYS **borduur** by. WAT het die gepeupel, geeste

Table 8: Concordance lines for *borduur*

The literal sense of *embroidering* is depicted by lines 30, 50, 70 and 80 and the figurative sense of *elaboration* in lines 40, 60, 90 and 100. Clear examples of usage for possible inclusion into the article of *borduur* in a new dictionary or a revised edition are available in abundance in the concordance lines, for example *borduur 'n naam/prentjie op die trui* 'embroider a name/picture on the jersey', *borduur voort op Engeland se reputasie* 'elaborate further on England's reputation'. Typical collocations such as *borduur voort/verder* 'elaborate further', *borduur en stik/brei* 'embroider and stitch/knit' can also easily be detected.

In the case of the lemma *blok* (noun), HAT distinguishes 9, and WAT 26 senses.

HAT

blok […] s.nw. (-ke) **1** (Groot) stuk hout, metaal, ens.: *'n Blok beton, marmer. Die seuntjie sit heerlik met sy blokkies en speel,* speelgoed wat 'n reëlmatige vorm het. **2** *(fig.)* Swaargeboude persoon: *'n Blok van 'n vent.* **3** *(hist.)* Strafwerktuig om die bene of nek van 'n gevangene vas te sluit: *Iemand in die blok sit.* **4** Groep, afdeling: *'n Hele blok huise. Blokke en blokke woonstelle langs mekaar in Hillbrow. 'n Politieke blok vorm.*

5 Stuk hout aan die poot van diere; ook fig., bv. *dis 'n blok aan my been,* 'n belemmering. **6** Vierkantige of reghoekige figuur of voorwerp: *Die blokke van 'n skaakbord, van rokmateriaal. 'n Skryfblok.* **7** Ruimte in 'n stad, dorp tussen strate: *Hy woon twee blokke verder.* **8** *(filat.)* Groep (van vier) seëls, gewoonlik in die boonste regterkant van 'n vel; kontroleblok. **9** (Son)blokker. UITDR.: *'n Blok aan die **been** hê,* iets wat jou belemmer, in jou vooruitgang strem. *Jou **kop** op 'n blok sit,* iets met klem bevestig.

WAT

blok. I s. 1. Swaar stuk hout, steen of metaal, min of meer reëlmatig van vorm, gew. met een of meer vlak sye: *'n Blok om vleis op te kap. 'n Blok marmer.* **2.** *(fig.)* Lywige persoon: *'n Blok van 'n kêrel. 'n Blok van 'n kind.* **3.** Swaar houtstrafwerktuig waarin eertyds die bene en soms ook die arms en nek van 'n gevangene gesluit is: *Iem. i/d blok sit (sluit).* **4.** 'n Stuk hout wat aan 'n perd of bees se voorpote gebind word om te belet dat die dier oor slote spring, ens. **5.** 'n Gerwehoop, netjies op die oesland gepak sodat dit bv. teen reën beveilig is. **6.** *(skoenm.)* Houtlees om aan 'n skoen sy vorm te gee, soms gebruik i.p.v. die ysterlees; ook genoem *leesblok.* **7.** Klein, kubusvormige stukkie hout, ens.: *Blokke (blokkies) uit 'n kind se speelgoeddoos. 'n Blok(kie) sjokolade.* **8.** Kompleks geboue; ook, groot, samegestelde gebou: *'n Blok woonstelle, huise.* **9.** Ruimte, gew. min of meer reghoekig, in 'n stad of dorp, deur strate omgrens: *Ons woon drie blokke van die markplein af.* **10.** Vierkantige of langwerpige blaadjies papier wat a/d een kant op mekaar vasgeheg is vir skryf- of ander doeleindes: *'n Skryfblok.* **11.** Skrop: *Wiel- en skepblokke by dammakery.* **12.** Hout- of metaalliggaam waarin een of meer skywe om een as draai soos by 'n katrol: *Blok en loper vorm saam 'n takel.* **13.** *(spoorw.)* Een v/d aansluitende baangedeeltes waarin die baanvakke verdeel word i/d sisteem van beveiliging v/d treinloop, en waarin daar op enige tydstip nie meer as een trein hom op dieselfde spoor mag bevind nie. **14.** *(wam.)* **a.** 'n Stukkie yster met twee gate daardeur waar die twee punte van 'n klou deurgaan en a/d ander kant moere kry. **b.** Sien BRIEKBLOK. **15.** *(delw.)* Een v/d klein, min of meer kubusvormige, swart klippies in diamanthoudende gruis — gew. as vkw.: *Blokkies is 'n goeie teken van diamante.* Ook genoem *blokkiesbantom.* **16.** *(filat.)* 'n Aantal posseëls wat aan mekaar vas is i/d vorm van 'n reghoek en wat uit meer as een ry bestaan: *'n Blok van vier. 'n Blok van twaalf, in drie rye van vier elk. Twee posseëls kan nie 'n blok vorm nie.* **17.** Min of meer reghoekige stuk grond: *Ons het twee blokke geploeg met 'n wenakker aan elke ent.* Vgl. GEWEN. **18.** Plek waar die persoon wat "aan" is, sy oë toehou by aspaai en wegkruipertjie; ook genoem *bof.* **19.** *(tolspel)* Mislukte gooihou, d.w.s. sonder dat die tol draai: *'n Blok gooi.* **20.** Gevoellose persoon: *Die vrou is 'n koue blok.* **21.** Aaneensluiting van partye of groepe: *'n Politiek-ekonomiese blok. 'n Militêre blok.* **22.** Onderdeel van 'n skaaf waarin die beitel bevestig is. **23.** Vierkantige of reghoekige figuur of vak: *Die blokke in 'n blokkiesraaisel. Die blokke van 'n dambord. Blokke op 'n stuk geweefde goed, op plakpapier, ens.* **24.** *(landbou)* Werktuig bestaande uit drie houtbalke i/d vorm van 'n gelykbenige driehoek wat gebruik word om omgeploegde grond gelyk te sleep. **25.** Sien DRUKBLOK. **26.** Groep plase wat gelyktydig uitgemeet en uitgegee is: *Die Hertzogblok in S.W.A. is met behulp van fondse van regeringsweë tot beskikking v/d Angola-Boere gestel.*

Sense	**HAT**	**WAT**	**Media24 KWIC**	**HAT Sense no.**	**WAT Sense no.**
Heavy object of wood/metal/brick	✓	✓	✓	1	1
Heavy person	✓	✓	✓	2	2
Heavy object of punishment (stocks)	✓	✓		3	3
Restraint for horses	✓	✓		5	4
Bulk of grain		✓			5

Last (shoemaking)		✓			6
Toy	✓	✓	✓	1	7
Group of apartments	✓	✓	✓	4	8
Street sections	✓	✓	✓	7	9
Stack of writing paper		✓			10
Dam scraper		✓			11
Block (and tackle)		✓			12
Railway section		✓			13
Object used in wagon building		✓			14
Black cubical stones (mining/digging)		✓			15
Group of stamps	✓	✓	✓	8	16
Square sections of ground	✓	✓		7	17
In hide-and-seek game		✓			18
In top (toy) game		✓			19
Apathetic person		✓			20
Parties/groups	✓	✓	✓	4	21
Stock of a plane (tool)		✓			22
Square/rectangular figure	✓	✓	✓	6	23
Object used to flatten soil (agriculture)		✓			24
Pressure block		✓			25
Group of farms		✓			26
Prevention object	✓			9	
Rubik's cube			✓		

Table 9: Senses of *blok* in WAT and HAT and occurrences in Media24

Media24 occurrences were found in support of six of the nine senses given in HAT and eight of the 26 senses given in WAT. Once again the evidence suggests that the Media24 archive could be a sufficient tool for the compilation of a major dictionary but insufficient as sole corpus for the compilation of a dictionary of the magnitude of WAT.

Conclusion

It can be concluded that in the current situation where no large designed corpus for Afrikaans exists, the Media24 archive is an excellent substitute. In fact, its value goes far beyond a limited component of a corpus design pattern, i.e. 'press'. The Media24 archive is so vast and versatile that it can be regarded as a world of information in its own right and its success in terms of broad coverage can probably be attributed to the fact that virtually all aspects of modern life in South Africa are covered in the daily reporting of these newspapers.

The question could even be asked whether the stage has not been reached in corpus-based lexicography where media coverage is so comprehensive in reporting on all spheres of everyday life that mega newspaper corpora have indeed become a world in one medium, i.e. a corpus sufficient, or at least going a long way as a basis for the compilation of general dictionaries.

Endnotes

1. Related lemmas such as *idioties* ('idiotic'), *idiotisme* ('idiotism') falling alphabetically outside the stretch *ideaal–idioot* were not considered.
2. Strategy utilised by Prinsloo and Gouws (2006) to express the increasing number of tokens per 5-year period in the Media24 archive as equal comparable sections, i.e. frequency counts of words per 50 million tokens.
3. Key status as a result of its frequent use as a reference marker in the source text.
4. Key status as a result of (over)use of the passive in describing each activity.
5. Key status as a result of its frequent use as a reference marker in the source text
6. Key status as a result of its frequent use as a reference marker in the source text.

References

Dictionaries and corpora

BROWN = *Brown Corpus of Standard American English* http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html.

HAT = Odendal, F.F. and R.H. Gouws. 2005. *HAT. Verklarende Handwoordeboek van die Afrikaanse Taal.* Fifth Edition. Cape Town: Pearson.

LOB = *Lancaster-Oslo/Bergen Corpus* http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html.

WAT = Schoonees, P.C. (Ed.-in-chief). 1972. *Woordeboek van die Afrikaanse Taal. Volume 4 H–I.* Pretoria: Government Printer.

Other literature

Atkins, B.T. Sue and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography.* Oxford/New York: Oxford University Press.

Atkins, B.T. Sue, Michael Rundell and Edmund Weiner. 1997. *Salex'97. A Training Course in the Compilation of Monolingual Dictionaries.* Unpublished course material of a tutorial held at the Dictionary Unit for South African English, Rhodes University, Grahamstown, 15–26 September 1997.

Biber, D. 1993. Using Register-diversified Corpora for General Language Studies. *Computational Linguistics* 19: 219-241.

Biber, D. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison.* Cambridge: Cambridge University Press.

De Schryver, G.-M. and D.J. Prinsloo. 2000. Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 2: The Microstructure. *South African Journal of African Languages* 20(4): 310–330.

Eaton, Jan. 1989. *Borduursteke vir Suid-Afrika. 'n Volledige gids*. Cape Town: Delos.

Garside, Roger, Geoffrey Leech and Tony McEnery (Eds.). 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London/New York: Longman.

Gilbert, Zoë. 1985. *Die Suid-Afrikaanse tuin. Maand vir Maand*. Second Edition. Johannesburg: Central News Agency.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London/New York: Longman.

Kilgarriff, Adam. 1997. Putting Frequencies in the Dictionary. *International Journal of Lexicography* 10(2): 135-155.

Kruyt, J.G. and M.W.F. Dutilh. 1997. A 38 Million Words Dutch Text Corpus and its Users. *Lexikos* 7: 229-244.

MacLeod, Catherine and Ralph Grishman. 2000. The Influence of Corpora on Lexicons: Corpora Use in the Creation of COMLEX Syntax and NOMLEX. Heid, Ulrich et al. (Eds.). 2000. *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000, Stuttgart, Germany, August 8th–12th, 2000*: 141-148. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Media24: http://www.media24.co.za.

Prinsloo, D.J. and R.H. Gouws. 2006. Fashion Words in Afrikaans Dictionaries: A Long Walk to Lexicographic Freedom or Just a Lexical Fly-by-Night? Corino, E., C. Marello and C. Onesti (Eds.). 2006. *Proceedings XII EURALEX International Congress. Turin, Italy, September 6th–9th, 2006*: 301-312. Alessandria: Edizioni dell'Orso.

Otlogetswe, T.J. 2007. *Corpus Design for Setswana Lexicography*. Unpublished Ph.D. Thesis. Pretoria: University of Pretoria.

Summers, Della. 1993. Longman/Lancaster English Language Corpus — Criteria and Design, *International Journal of Lexicography* 6(3): 181-208.

Summers, Della. *s.d.* [1996–1998]. Corpus Lexicography — The Importance of Representativeness in Relation to Frequency. *Longman Language Review* 3: 6-9.

Turpin-Delport, Lesley. 1988. *Die Suid-Afrikaanse boek van laslappie en appliekwerk*. Cape Town: C. Struik.