http://lexikos.journals.ac.za

A Balanced and Representative Corpus: The Effects of Strict Corpus-based Dictionary Compilation in Sesotho sa Leboa*

V.M. Mojela, *Sesotho sa Leboa National Lexicography Unit, University of Limpopo, Turfloop Campus, Polokwane, South Africa (victor.mojela@ul.ac.za)*

Abstract: Theoretically the Northern Sotho language is made up of almost 30 dialects while practically it is not so, because the standard language was formed from very few of its dialects. As a result, even today the language has no corpus which is balanced or representative owing to the fact that almost all of the available corpora are compiled from the written standard language and the written dialects. The majority of the Northern Sotho dialects do not have written orthographies, and the few dialects which had written orthographies prior to standardization came to monopolize the standard language and the Northern Sotho corpora. Therefore, the compilation of a corpus-based dictionary in Northern Sotho is tantamount to a continuation of producing unbalanced and unrepresentative dictionaries, which continue to sideline and to marginalize the majority of the communities and the linguistic varieties which could potentially enrich both the Northern Sotho standard language and the Northern Sotho corpora. The main objective with this research is to analyze, to expose and to suggest ways of correcting these irregularities so that the marginalized Northern Sotho dialects can be accommodated in the standard language. This will obviously increase the size of the Northern Sotho standard language and the corpus by more than 50%.

Keywords: CORPUS, BALANCED CORPUS, REPRESENTATIVE CORPUS, STANDARDIZATION, DIALECT, ORTHOGRAPHY, MARGINALIZED DIALECTS, PRESTIGE DIALECTS, MISSIONARY ACTIVITIES

Opsomming: 'n Gebalanseerde en verteenwoordigende korpus: Die gevolge van streng korpusgebaseerde woordeboeksamestelling in Sesotho sa Leboa. Teoreties bestaan die Noord-Sotho taal uit byna 30 dialekte, terwyl dit prakties nie die geval is nie omdat die standaardtaal uit slegs 'n paar van sy dialekte gevorm is. Gevolglik het die taal selfs vandag nog geen korpus wat gebalanseerd of verteenwoordigend is nie as gevolg van die feit dat byna al die beskikbare korpusse saamgestel is uit die geskrewe standaardtaal en die geskrewe dialekte. Die meerderheid Noord-Sotho dialekte het nie geskrewe ortografieë nie, en die paar dialekte wat geskrewe ortografieë gehad het voor standaardisasie het begin om die standaard-

* This article is a revised version of a paper presented at the Seventeenth Annual International Conference of the African Association for Lexicography (AFRILEX), which was hosted by the Department of African Languages, University of Pretoria, Pretoria, South Africa, 2–5 July 2012.

taal en die Noord-Sotho korpusse te monopoliseer. Die samestelling van 'n korpusgebaseerde woordeboek kom gevolglik neer op 'n voortsetting van die totstandbrenging van ongebalanseerde en onverteenwoordigende woordeboeke wat voortgaan om die meerderheid van die gemeenskappe en taalvariëteite opsy te skuif en te marginaliseer wat potensieel sowel die Noord-Sotho standaardtaal as die Noord-Sotho korpusse kan verryk. Die hoofdoel met hierdie navorsing is om maniere te ondersoek, uit te wys en voor te stel om hierdie ongelykhede reg te stel sodat die gemarginaliseerde Noord-Sotho dialekte in die standaardtaal ondergebring kan word. Dit sal vanselfsprekend die grootte van die Noord-Sotho standaardtaal en korpus met meer as 50% vermeerder.

Sleutelwoorde: KORPUS, GEBALANSEERDE KORPUS, VERTEENWOORDIGENDE KORPUS, STANDAARDISASIE, DIALEK, ORTOGRAFIE, GEMARGINALISEERDE DIALEKTE, PRESTIGEDIALEKTE, SENDELINGAKTIWITEITE

1. Introduction

Northern Sotho, or Sesotho sa Leboa, presently has corpora which were built entirely from published materials, and as such representing only the written and documented dialects. This is a major shortcoming, because the published documents in indigenous languages like Northern Sotho are usually based on the few dialects which are restricted to certain parts of society, while the majority of the undocumented dialects are sidelined. Northern Sotho is made up of approximately 30 dialects which are found in almost all five municipal districts of the Limpopo Province. Of all these dialects, almost half are marginalized 'languages'. This simply means that these dialects (or 'languages' as the communities themselves regard their dialects) are not included in written standard Northern Sotho.

This written Northern Sotho language, which is derived from the few documented dialects, i.e. the 'prestige' dialects which were fortunate to have written and published materials prior to standardization, are the ones which are represented in Northern Sotho corpora today. These irregularities are the subject of investigation in this research, whose objectives can be summarized as follows:

(a) to discuss the shortcomings and disadvantages of relying solely on corpus-based dictionary compilations in indigenous languages like Northern Sotho,

(b) to demonstrate that Northern Sotho does not have a 'balanced or representative' corpus,

(c) to analyze factors leading to the marginalization of the majority of the Northern Sotho dialects, and

(d) to show how purism and monopolies influenced the standardization of Northern Sotho, thereby leading to the marginalization of the majority of its dialects.

2. What is a corpus?

The term *corpus* is defined by several linguistic and lexicographic scholars, such as Watson (1976), Kennedy (1998), Gouws and Prinsloo (2005) among others. Watson (1976: 243) describes a corpus as:

> a body of writings of a particular kind, or on a particular subject.

Kennedy's (1998: 1) definition of a corpus is as follows:

> a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description.

The most direct and straightforward description of a corpus is given by Gouws and Prinsloo (2005: 21):

> The collection of written and spoken material from the sources earmarked for the dictionary basis. Data is compiled and stored as a lexicographic data basis which should preferably be an electronic corpus. An electronic corpus can be defined in an oversimplified way as a computerized collection of texts, such a collection of texts can, for example, consist of tape recordings of conversations and written texts which have been typed into the computer.

These definitions show that corpora are supposed to be compiled from both written and oral materials. But, on the contrary, it is not always easy to compile a corpus for languages or dialects with no orthographies or written forms. Gouws and Prinsloo (2005: 21-22) say the following in this regard:

> Unfortunately most corpora around the world lack sufficient data from spoken sources. The reason for this is that there are many logistical problems and ethical factors involved in the collection of spoken data. It is also much more expensive and time consuming to enlarge the corpus with spoken data compared to data available in electronic, printed or even handwritten format. Extending the corpus with data already in electronic format such as texts downloaded from the internet or texts already available on computer disk is relatively easy. Printed matter which is not available in electronic form can also relatively easily be computerized by means of Optical Character Recognition (OCR), commonly referred to as 'scanning'.

3. The issue of a 'balanced' and 'representative' corpus

A normal and appropriate general corpus for a language needs to be balanced and representative. According to Kennedy (1998: 20), "a general corpus is typically designed to be balanced, by containing texts from different genres — including spoken and written". Kennedy (1998: 52) further emphasizes that "for a corpus to be 'representative' there must be a clearly analysed and defined population to take the sample from".

Gouws and Prinsloo (2005: 25) re-emphasize these requirements with relevancy to the South African indigenous language situation as follows:

> Important for lexicographic work in South Africa is that corpus compilers should be sensitive to all of these aspects, i.e. to build as far as possible, corpora that are big enough, well balanced and representative so that valid conclusions for lexicographic purposes can be drawn.

As far as the standard language is concerned, it could be argued that, although no spoken data has been included, the compilers of currently available corpora for Northern Sotho, did make a real effort to capture available written sources. Ideally the corpus should be extended to include large quantities of dialectal information.

3.1 The consequences of standardization on Northern Sotho

Standardization has prioritized certain Northern Sotho dialects, while it has marginalized others, as indicated by the following facts:

— Purism was used as pretext by most language committee members who dominated the official language bodies, for standardizing their own dialects to 'represent' all other dialects. The few dialects which have contributed to standardized Northern Sotho form less than half of all the Northern Sotho dialects.

— Of the approximately 30 Northern Sotho dialects, the only dialects which are represented in standard Northern Sotho and the available corpora are Sekone (which is spoken in the central and southern parts of the Waterberg district and a section of the Capricorn district), Sekopa and Sepedi, which are used in the Sekhukhune district and Sekgaga (dialect of the Gamphahlele and Gamothapo districts), as well as the few dialects spoken in the Mankweng and Mamabolo areas.

— Standardization has marginalized the majority of the 'potential' Northern Sotho dialects, i.e. the dialects which are grouped as dialects of Northern Sotho even though practically they differ by far from the Northern Sotho standard language. These include major 'languages' like Sepulana, Setlokwa, Khelobedu, Seroka, Sephalaborwa, Sehananwa, Sekgaga (of Maake and Mogoboya), Sekhutšwe, and others.

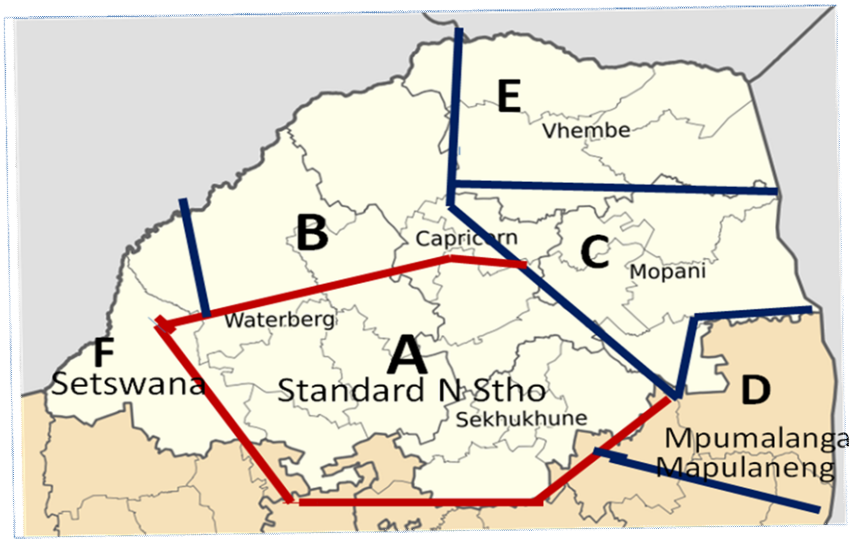
3.2 Demographic representation of the Northern Sotho dialects

Demographically the marginalized 'dialects' include all the linguistic areas of Botlokwa, Sekgosese–Lemondokop and the Senwabarwana areas in the Northern part of the Capricorn district and a small section of the Vhembe district; and the whole of the Mopani district, which encompasses, inter alia, the areas

of Bolobedu–Modjadjiskloof, Tzaneen, Trichardtsdal and Phalaborwa, as well as the Bushbuckridge–Mapulaneng area in the North Eastern part of Mpumalanga.

This means that Kennedy's principle, i.e. 'for a corpus to be 'representative' there must be a clearly analysed and defined population to take the sample from', was only considered with reference to the few dialects around Polokwane–Matlala, Lepelle–Nkumbi, Gasekhukhune and a section of the Waterberg district.

The following map explains the demography of Northern Sotho, showing dialectal distributions within the Limpopo and Mpumalanga regions:



Map showing the demography of the Northern Sotho dialectal regions

Area A: Standard Northern Sotho is formed on the basis of dialects spoken in this area. The area includes the central and south eastern part of the Waterberg district, the southern part of the Capricorn district and the whole of the Greater Sekhukhune municipal district. The dialects in this area include, inter alia, Sekopa, Sepedi, Sekgaga (of Mphahlele, Mothapo, etc.) and Sekone (of Moletši, Matlala, Bakenberg, Polokwane, Mothiba, Dikgale, etc.).

Areas B, C and D: All the Northern Sotho dialects in these areas are not represented in the standard language. These areas include:

the northern parts of the Waterberg and Capricorn districts (area B), inhabited by, inter alia, the Batlokwa and Bahananwa communities

the whole of the Mopani district, the north eastern part of the Capricorn district and a section of the southern part of the Vembe district (area C): These areas

are inhabited by the Batlokwa (in the Sekgosese–Lemondokop areas) and Baroka communities. These communities include, inter alia, the Balobedu, Bakgaga (ba Maake), Baphalaborwa, Banareng ba Sekororo, Batlokwa, etc.

the Bushbuckridge or Mapulaneng area, in the north eastern part of the Mpumalanga Province (area D). This is the place of residence of the Mapulana communities.

Areas E and F: Even though there are a few Northern Sotho dialects in these areas, the overwhelming majority of the indigenous communities in area E are the Venda people, while area F is dominated by the Batswana communities.

3.3 The gap between the standard languages and the marginalized dialects

All the Northern Sotho dialects or 'languages' in areas B, C and D, and a few scattered remnants in area E (in the above map) are marginalized. The dialects in these areas differ considerably from standard Northern Sotho even though the communities are forced to use the standard language for official communications in education and all official correspondence. Sometimes the gap between the standard language and the marginalized dialects is so wide that most of the communities in area A (in the above map) need interpreters to engage in effective communication with the communities of areas B, C and D, especially with the Balobedu, Baphalaborwa and Mapulana. These marginalized dialects in areas B, C and D are not only excluded from the standard language, but also from the Northern Sotho official orthography and, eventually, from the official corpora.

3.4 The influence of standardization on the compilation of the corpora

The fact that no printed matter exists for the dialects left the corpus compilers no alternative than to concentrate on the already written and standardized language in compiling the Northern Sotho corpora. Ideally these corpora should be extended by the inclusion of dialectal data. Lexicographers should then be in a position to consider corpus-based dialectal data especially for lemma selection and translation equivalents. So, for example, a comprehensive dictionary should include lemmas such as *molema*, *khobe*, *lesalabu* and *kholophana* and also such words as translation equivalents for the lemmas *bat*, *fish*, *watermelon*, etc. respectively.

In assisting to increase and to improve the University of Pretoria corpus, the Sesotho sa Leboa National Lexicography Unit staff and the University of Pretoria lexicographers in the Department of African Languages, under the guidance of lexicographic scholars like Prof. D.J. Prinsloo and Prof. G.-M. de Schryver, much relied on what Gouws and Prinsloo (2005: 22) refer to as the OCR or Optical Character Recognition:

> Printed matter which is not available in electronic form can also relatively easily be computerized by means of Optical Character Recognition (OCR), commonly referred to as 'scanning'.

This means that existing published or printed materials in Northern Sotho were used to develop the corpus. Unfortunately, these materials are only available in the 'one-sided' standard language because the marginalized dialects did not have any written or published materials, apparently because these dialects did not have any orthography. Even up to this moment, there are no available printed materials covering the marginalized dialects. Therefore, continuation of relying solely on the 'printed matter' will be tantamount to advancing and promoting the marginalization policy of the purists.

4. The role played by the other Sotho languages in the marginalization of the Northern-Sotho dialects

The gap which exists between the standard Northern Sotho language and some of its own dialects is much wider than that which exists between Northern Sotho and the other two Sotho languages, i.e. Setswana (Western Sotho) and Sesotho (Southern Sotho). The influence of these Sotho languages contributed much to the widening of the gap between the Northern Sotho standard language and its dialects. The process of the interaction and the relationships between the three Sotho languages can be divided into three phases, i.e. the missionary period, the unification or harmonization period and the separate development (apartheid) period.

4.1 The missionary period

This covers the period before 1929, when the role of developing and converting the Sotho oral languages into written 'languages' was effected only by the missionaries. The task of writing and compiling the first orthographies for the Northern Sotho language was started by the German missionaries, i.e. the Berlin Evangelical Missionary Society, during the 19th century. They established missionary stations under the Bapedi and Bakopa communities, whose dialects are much closer to Setswana and the Southern Sotho communities than the Northern Sotho dialects in the North and the Lowveld. That is why the first written so-called 'Sepedi' missionary orthography did not deviate much from the structures of the other two Sotho languages. This can be ascribed to the following factors:

— Historically the Bapedi and Bakopa communities are more aligned to the Batswana communities than to the other Northern Sotho dialectal groups.

— The Bapedi and Bakopa communities are in closer proximity to the Batswana and the Basotho communities when compared to the Northern

Sotho communities in the North and the Lowveld areas, who are, in turn, bordering on the Vatsonga and the Vhavenda communities.

4.2 The unification or harmonization period

This is the period between 1929 and 1961 when the Union Government took over the administration of the indigenous schools from the missionaries for the first time, and started regulating the development of the indigenous languages. As a result, when the Transvaal Education Department (TED) took over responsibility for organising the Sotho languages after 1929, the focus was more on uniting all the Sotho languages, i.e. Western Sotho (Setswana), Southern Sotho and Northern Sotho, into one standard language. These languages were, by then, regarded as the Sotho dialects. The Transvaal Sotho District Committee, which was formed by the TED, compiled and introduced its first Sotho orthography in 1930. After holding several meetings and conferences between 1930 and 1950, like the Somerset House Conference of 1947, the Orthography Sub-Committee of the Sotho Language Board revised and adopted the official Sotho orthography for all the Sotho 'languages' in 1950. After passing the Bantu Education Act in 1953, the Union Government took over the responsibility of running formal education from the missionaries, and the 1950 Sotho orthography became official in the Transvaal. Mojela (2008: 121) comments as follows in this regard:

> It was only after 1929 that the Transvaal Education Department (TED) started making attempts at standardizing the Sotho languages in the former Transvaal which eventually led to the formation of the Language Boards (Mojela 2005: 46). In South Africa, for instance, it was only after the passing of the Bantu Education Act in October 1953 (Act No. 47 of 1953) that the South African government took over control of formal education from the missionaries.

The 1950 orthography brought the three Sotho 'languages' close together and this gave advantage to all the Northern Sotho dialects, like Sepedi, Sekopa, Sekone, etc., which are structured more closely to the Southern Sotho and Setswana languages, because the Northern Sotho standard language came to be based on this orthography. As a result, this compromised the Northern Sotho dialects, like Setlokwa, Selobedu, Sepulana, Seroka, etc. which are found in the far North and the Lowveld. All those dialects whose structures were too remote from Setswana and Southern Sotho were thus marginalized. Most of the early Northern Sotho publications and printed materials, which later came to be used as important sources in the standardization of the Northern Sotho language were written according to the 1950 orthography and the 1953 unified Sotho standard language. The following examples from Mojela (2008: 127) demonstrate the remoteness of the standard Northern Sotho language from its own dialects, when compared to its closeness to the Setswana language:

Sesotho sa Leboa	Setswana	Selobedu (NS dialect)	English
mopani	*nato/mopani*	*mo̱hanare*	mopani tree
leribiši	*lerubisi*	*mmankho̱to*	owl
mmankgagane	*mmamanthane*	*molema*	bat
hlapi	*tlhapi*	*khobe*	fish
betha/itiya	*betsa*	*mo̱ta/tiya*	hit
legotlo	*legotlo*	*leho̱to/peba/maṉtoro*	mouse
legapu	*legapu*	*lesalabu*	watermelon
nona	*nona*	*kholophana*	be fat/gain weight
mokgaditswane	*mogaditswane*	*mphekwa*	lizard

These examples demonstrate how the harmonization and unification of the Sotho languages brought the Northern Sotho standard language closer to Setswana and Southern Sotho, while at the same time distancing itself from its own dialects.

4.3 The separate development (apartheid) period

The attempt to unify the Sotho languages was destroyed by the policy of separate development, which aimed at developing all the South African societies separately. With this policy, the apartheid regime wanted to use a 'divide and rule' strategy to keep power in the hands of the white minority. As a result, the three subcommittees of the unified Sotho language, i.e. the subcommittees for Setswana, Southern Sotho and Northern Sotho were converted to fully-fledged autonomous Language Boards for the respective languages to develop each of them separately. Even though separate development gave the standardizing authorities enough chance to pay attention to the marginalized dialects in every language, in Northern Sotho too much damage had already been done because: (a) the written language and the available orthographies were dominated by the 1953 standard Sotho language, (b) the educated elite who came to dominate membership of the newly established Language Boards were educated on the basis of the 1953 Sotho standard language, (c) all the publications which were used in the standardization of Northern Sotho were totally foreign to the marginalized dialects, and (d) ultimately, the only available materials for the corpus compilers in the compilation of the Northern Sotho corpora were still those which excluded the marginalized dialects.

5. The repercussions of dialectal marginalization

The issue of the 'unbalanced' and 'unrepresentative' corpora, which resulted from a one-sided standardization has led to the emergence of a nationalistic spirit among the marginalized communities:

— The emerging elite groups and the *magoši* (traditional leaders) from areas

such as Bolobedu, Makhutšwe, Botlokwa, Senwabarwana, Mapulaneng, etc. have already started questioning the validity of incorporating their dialects in the Northern Sotho standard language, which is not only lexically and morphophonologically foreign to them, but in all practical respects, too different from their dialects.

— At the same time, most communities whose dialects were favoured by standardization started claiming ownership of the standard language, while those groups who were sidelined by the standardization started disowning and opting for withdrawal from the standard language, because they believe their 'languages' are misplaced.

— The Balobedu under the leadership of the 'self-imposed' Archbishop Prince Madlakadlaka, and influential activists such as Mr Phetole Mampeule, under the influence of the philosophies of the Kara Heritage Institute of Dr Mathole Motshekga, started questioning the inclusion of Khelobedu into the Northern Sotho standard language. They have already submitted several petitions and requests to the Government and the Constitutional Court to have Khelobedu declared an official language. Some are even insisting that Khelobedu is more aligned to Tshivenda than to Northern Sotho, and as such, rather a Tshivenda dialect than a dialect of Northern Sotho.

— The Mapulana communities under the leadership of high profile personalities and their *magoši* are also demanding official status for Sepulana because they insist they are not Bapedi, but Basotho (in the east).

6. Conclusion

This research demonstrates that corpus lexicography in the South African indigenous languages, like Northern Sotho, is not always possible, because there are still many dialects within these languages which are not included in both the official orthographies and the standard languages. Since the corpora are compiled mostly from the written languages and from published materials, the dialects which did not have written forms or published materials will not always be included in the corpora. Even though sometimes oral materials were collected and included into the corpora, very few of these published materials were recorded in languages like Northern Sotho because, up to this moment, almost all the available Northern Sotho corpora do not have anything related to the marginalized dialects like Khelobedu, Sepulana, Setokwa, Sehananwa or Sephalaborwa. As a result, the available Northern Sotho corpora do not conform to the lexicographic principles of 'balance' and 'representativeness'. The standard language is neither balanced nor representative because it reflects less than half of its dialects.

In conclusion, this research recommends further research and a thorough

revision of the official orthography and the standard language to incorporate all the omitted dialects into the Northern Sotho language, before prescribing strict corpus-based lexicography. This will not only silence the emerging nationalistic spirit which threatens to divide the language, but will also double the size of the Northern Sotho corpus.

References

Gouws, R.H. and D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography.* Stellenbosch: SUN PReSS.

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

Mojela, V.M. 1991. *Semantic Changes Accompanying Loan-words in the Northern Sotho Lexicon*. Unpublished M.A. Thesis. Pretoria: Vista University.

Mojela, V.M. 2005. Standardization and the Development of Orthography in Sesotho sa Leboa — A Historical Overview. Webb, V. 2005. *The Standardization of African Languages in South Africa.* Pretoria: University of Pretoria.

Mojela, V.M. 2008. Standardization or Stigmatization? Challenges Confronting Lexicography and Terminography in Sesotho sa Leboa. *Lexikos* 18: 119-130.

Rooney, K. (Ed.-in-chief). 2001. *Encarta Concise English Dictionary*. London: Bloomsbury.

Watson, O. (Ed.). 1976. *Longman Modern English Dictionary*. Harlow: Longman.