Nordic Journal of African Studies 10(3): 374-398 (2001)

Corpus-based Activities versus Intuition-based Compilations by Lexicographers, the Sepedi Lemma-Sign List as a Case in Point*

GILLES-MAURICE DE SCHRYVER**
Ghent University, Belgium & University of Pretoria, South Africa
& D.J. PRINSLOO
University of Pretoria, South Africa

ABSTRACT

The authors of this article firmly believe in the advantages of utilising a corpus for lemma-sign list creation. However, one should not overreact and assume that alternative methods for the creation of a dictionary's macrostructure have no virtues, or that alternative methods are in principle per definition marred by inconsistencies. What is called for is a perspective on corpus-based activities versus intuition-based compilations by lexicographers. Therefore, while the supremacy of a corpus remains undisputed in compiling a lemma-sign list, this article also intends to show that a well-planned combination of a variety of existing lists that were assembled manually, results in a lemma-sign list with a remarkable internal consistency. Hence, the aim of this article is twofold. Besides a brief illustration of typical macrostructural inconsistencies, the main focus will be on a series of consistencies encountered in the compilation of lemma-sign lists for different sub-dictionaries at various stages of the *Sepedi Dictionary Project* (SeDiPro). Some attention will also be devoted to the so-called Miraculous Consistency Ratio '(x 1.25)^4 = x 2.4' – being a sequence of four 25% increases which result from a collation of five manually compiled Sepedi lemma-sign lists.

Keywords: lexicography, Sepedi (S32), corpus, intuition, macrostructure, lemma-sign list, (in) consistencies, Sepedi Dictionary Project (SeDiPro), *Pretoria Sepedi Corpus* (PSC), *Miraculous Consistency Ratio*

INTRODUCTION

The aim of this article is twofold. Besides a brief illustration of typical macrostructural *inconsistencies* in existing Sepedi dictionaries, the main focus

[*] An earlier version of this article was read by the authors at the 11th International Biennial Conference of ALASA (the African Language Association of Southern Africa), University of Port Elizabeth, South Africa, 11-13 July 2001.

[**] G-M de Schryver is currently Research Assistant of the Fund for Scientific Research – Flanders (Belgium).

will be on a series of *consistencies* encountered in the compilation of lemma-sign lists for different sub-dictionaries at various stages of the *Sepedi Dictionary Project* (SeDiPro). SeDiPro is a project that already led to the publication of one large bilingual Sepedi – English dictionary (Prinsloo & De Schryver 2000). It is also a project in which the lemma-sign list for a multi-volume monolingual dictionary is currently being prepared for the Sepedi *National Lexicography Unit* (NLU). Sepedi, also known as Northern Sotho, belongs to the Bantu language family (S32 in Guthrie’s classification) and is one of South Africa’s eleven official languages.

1. BRIEF THEORETICAL CONSPECTUS

Regardless of size, any general dictionary and certainly any learners’ dictionary should at least cover the basic or core vocabulary. For the English language, the 1930s saw the first attempts to limit ‘essential vocabulary’ to 1000 words (Whitcut 1988; McArthur 1989), and the “earliest English dictionaries for foreign learners ... were developed in the 1930s from the vocabulary studies of Harold E. Palmer, Michael West, and A.S. Hornby of the UK and Edward L. Thorndike of the US” (Landau 2001: 74). Already in 1921 Thorndike had published his *Teacher’s Word Book*. Based on a (pre-electronic) word count of 4.5 million words, this book “consists of several lists of words showing their relative frequency ... designed to help educators and teachers determine which words are common enough to be used” (Landau 2001: 273). Ever since, frequency counts derived from (electronic) corpora have been instrumental in setting up a language’s basic or core vocabulary. Recently, Hartmann & James (1998: 13) defined basic vocabulary as “[t]hose words selected by frequency counts and similar means”, while Bussmann (1996: 49) maintains that “the most important criterion for determining the basic vocabulary is the frequency of use”. It is thus not surprising that present-day lexicographers increasingly consult frequency counts derived from a well-designed electronic corpus in order to compile a lemmatised frequency list. This ordered list of canonical forms then constitutes the backbone of the lemma-sign list of their dictionaries.

One could say that setting up a dictionary’s lemma-sign list is the first major problem with which any lexicographer is confronted. This is well echoed in the literature:

> One of the basic problems of lexicography is to decide what to put in the dictionary and what to exclude. (Tomaszczyk 1983: 51)
>
> Selection is guided by usefulness, and usefulness is determined by the degree to which terms most likely to be looked for are included. (Gove 1961[3]: 4a)

> Lexicographers constantly have to make pragmatic decisions on what to include in a dictionary to conform to the dictates of space available. (Walter 1996: 640)
>
> The decision what to include in the dictionary still has to be made by the lexicographer himself, however, and this depends in turn upon the nature and size of the dictionary and its intended users. In this respect lemmatised frequency-lists can be a further help, … we have reached a stage where co-operation between man and machine is useful and perhaps indispensable in making better dictionaries. (Martin *et al.* 1983: 81-82, 87)

Formulated differently, in order to decide what to put in and what to exclude from a *useful paper dictionary*, lemmatised frequency lists may be advanced as guidance.

2. Some Typical Macrostructural Inconsistencies

Corpus-orientated lexicographers are quick to point out and elaborate on the many *inconsistencies* in the macrostructural compilation of dictionaries that were not compiled with the use of corpora. Quite a number of typical macrostructural inconsistencies can indeed be cited:

1. inconsistencies when it comes to the relative length of alphabetical stretches, by treating certain sections of the lemma-sign list more exhaustively than others;
2. inconsistencies regarding the creation of the lemma-sign list (mostly as a result of an enter-them-as-they-cross-my-way approach to dictionary compilation) such as:
   - 2.1. the omission of *words most likely to be looked for*, while words less likely to be looked for are included,
   - 2.2. the partial treatment of lexical items belonging to a *closed set* (currencies, letters of the alphabet, digits, seasons, etc.),
   - 2.3. the unequal treatment of *various prefixes* (i.e. mostly ‘inflection’ in Bantu),
   - 2.4. the absence of a policy to deal with *productive* versus *non-productive suffixes* (i.e. mostly ‘derivation’ in Bantu),
   - 2.5. the blind running of each stem through *all possible verbal and nominal derivations*, simply concatenating affixes, which results in serious doubts among mother tongue speakers whether many of these derivations do exist,
   - 2.6. the ad hoc handling of *transparent* versus *non-transparent derivations*;
3. inconsistencies in terms of the choice of canonical forms.

Since space restriction does not allow us to treat all these types of macrostructural inconsistencies here, we suggest to briefly consider three of them.

The first, ‘1. inconsistencies when it comes to the relative length of alphabetical stretches, by treating certain sections of the lemma-sign list more exhaustively than others’, is for instance found in the *Pukuntšu woordeboek* (Kriel 1983^3), a bilingual Sepedi – Afrikaans learners’ dictionary. Thumbing through this dictionary, one realises that Kriel treated the first few alphabetical stretches exhaustively but seemed to ‘get tired’ as he moved through the alphabet. This is illustrated visually in Figure 1 with two random sections: one from the beginning and one from the end of Kriel’s dictionary.

Figure 1a. Random section from the beginning of the *Pukuntšu woordeboek* (Kriel 1983^3).

aka, *a.ka*. (*-ile*, *-etše*), lieg, leuens vertel, jok, onwaarheid spreek (dial. kyk: *aketša*).

aka, *a.ka*, inhaak, vashaak, haak, aanhaak, soen, omarm, lieg, liefkoos; *akwa*, gehaak/ingehaak word; *akêla*, haak vir; *akelana*, mekaar liefkoos, vriendskaplik verkeer; *akelwa*, ingehaak word vir; *akiwa*, ingehaak word; *ake, ga, sa*, nie (in)haak nie; *akê*, mag/moet haak of inhaak; *moaki*, haker; *baaki*, hakers.

akalala, *a ka la.la*, sweef, hang oor, oorhang; *akalalêla*, sweef vir/oor; *akalatša*, laat sweef, vlerke oopsprei om te sweef; *akaladitše*, het laat sweef; *se bone nong go -*, *go wa fase ke ga lona*, hoogmoed kom tot 'n val; *akalatšwa*, genoodsaak om te sweef; *akalalwa* gesweef word; *akalêla*, hang/sweef oor, wydsbeen staan oor; *akaletše*, het gesweef oor; *moakaladi*, persoon wat sweef.

akama, *a ka.ma*, verwonder/verbaas wees; *akamela*, inlaat (bemoei) met; *akametša*, (laat) verbaas, verbasing wek, aangaap, toeroep; *akametšwa*, verbaas/aangegaap word, toegeroep word.

akere, *'a kê.'rê*, akker.

aketša, *a ke.tša*, leuen vertel, lieg, jok; *akeditše*, het (gelieg) 'n leuen vertel; *sa aketše*, nie lieg nie.

akga, *a.kga*, werp, gooi, slinger, swaai, beweeg; *akgaakga*, heen en weer beweeg (soos branders), slinger, skommel; *akgaakgwa*, heen en weer geslinger word; - *diatla*, arms swaai, met leë hande loop; - *dinao*, voet in die wind slaan; *akgwa*, beweeg/geslinger word; - *akgêga*, skommel, swaai; - *akgêla*, slinger, swaai, werp; *akgêla*, slinger na/vir, tou om die horings gooi, met 'n vangtou vang, uitkrap, soos kole uit 'n vuur; *akgelwa*, geslinger word, gevang word met 'n tou; - *dikobo*, klere uitpluk.

Figure 1b. Random section from the end of the *Pukuntšu woordeboek* (Kriel 1983^3).

tsirikana, *'tsi'ri ka.na*, klink.

tsirima, *'tsi'ri.ma*, klink, lui, uitspuit, vorentoe spring.

tsirimetša, 'tsi'ri me.tša, laat klink, vasbyt, laat lui, styf vasbind.

tsirinya, *'tsi'ri.nya*, laat klink, lui.

tširoga, *'tši ro.ga*, wakker skrik, senuweeagtig word, opskrik, moedeloos word.

tširogo *'tši ro.gô*, impuls.

tširoša *'tši ro.ša*, wek, skrikmaak.

From Figure 1a one can see that Kriel started off with great enthusiasm, trying to include all verbal and nominal (both singular and plural!) derivations from a particular stem in the article of that stem, extensively covering expressions and collocations, and even giving grammatical guidance. This lumping resulted in numerous column-lines per article. By the time Kriel reached the end of the alphabet, he had not only changed his lemmatisation policy from lumping to splitting, but also limited the treatment per article to an absolute minimum. The latter is obvious from Figure 1b. Expressed in number of articles per page, Table 1 illustrates the same inconsistency numerically.

Table 1. Number of articles in the *Pukuntšu woordeboek* (Kriel 1983^{3}) in different alphabetical stretches.

Alphabetical stretch	Random page number	Number of articles
A	2	22
O	241	52
S	281	75

Table 1 clearly indicates that towards the end of the alphabet more than thrice the number of lemma signs were treated per page compared to the first alphabetical category, suggesting that Kriel changed his lemmatisation strategy (and thus created a huge inconsistency regarding *lumping* and *splitting* in the same dictionary), but also that he got tired.[1]

As a second example of typical macrostructural inconsistencies in Sepedi dictionaries, we may look at '2.2. the partial treatment of lexical items belonging to a *closed set*'. The most comprehensive dictionary currently available for Sepedi is the *Pukuntšu ye kgolo* (Ziervogel & Mokgokong 1975). Figure 2 shows all the lexical items from the closed set 'days of the week' that were entered in this dictionary.

Figure 2. The days of the week in *Pukuntšu ye kgolo* (Ziervogel & Mokgokong 1975).

LÁBÓBEDÍ (< **tšatši la bobêdi**) (**Labobêdi**) Dinsdag // Tuesday; (< **lentšu la bobedi) (labobêdi)** altstem // alto (voice)

LÁBÓRÁRO (< **letšatši la boraro**) (**Laboraro**) Woensdag // Wednesday; (< **lentšu la boraro**) tenoor(stem) // tenor (voice)

LÁBÓNE (< **tšatši la bone**) (**Labonê**) Donderdag // Thursday; (< **lentšu la bone**) (**labonê**) bas(stem) // bass (voice)

LÁBÓHLÁNO (< **tšatši la bohlano**) Vrydag // Friday

SÓN'TAGA, (se-)/di- (Sôntaga) (< Afr.), cf. LÁMODÍMO, Sondag // Sunday

LÁMORENA (**Lamorêna**) (< **letšatši la Morêna**) Sondag // Sunday

One can see from Figure 2 that Ziervogel & Mokgokong only included five of the seven days of the week, totally neglecting the existence of *Mošupologo* 'Monday' and *Mokibelo* 'Saturday'. Ironically, these two days belong to the top-three of the most-frequently used days. Moreover, these two missing days

[1] Inconsistencies regarding the relative length of alphabetical stretches are discussed in great detail in Prinsloo & De Schryver (*forthcoming*).

belong to the top-2000 word-band of the Sepedi lexicon. Note also that a cross-reference is given from *Sontaga* to *Lamodimo*. The latter, however, is nowhere to be found in the macrostructure. For a stem-based dictionary, this is particularly bad, as, upon realising that *Lamodimo* has not been included as such, the user will try to find *Lamodimo* under *-modimo*, then under *-dimo*, and finally under *-mo*. All to no avail.[2]

Finally, as a third example, we can look into '3. inconsistencies in terms of the choice of canonical forms'. More specifically, we may study the treatment of adjectives in various Sepedi dictionaries. Since Bantu adjectives take the nominal prefix of the noun they are modifying, there are basically two ways in which one can enter adjectives in a Bantu dictionary. In a so-called 'stem-based dictionary' only the stem will be entered (preferably preceded by '-' to indicate that a prefix should be attached to the stem). This stem then functions as the 'canonical form'. Yet, in a so-called 'word-based dictionary' all the possible forms of the adjective are entered, at which point there is no need to enter the stem. One could however deviate from the latter and only include the most frequent forms, or one could deviate from the former and include, besides the canonical form, also the forms with 'sound strengthening'. Table 2 shows the treatment of the adjective *-golo* 'big' (which has as sound-strengthened form *kgolo* in classes 8 to 10) in five different Sepedi dictionaries.

Table 2. The adjective *-golo* 'big' in five different Sepedi dictionaries.

Class	'big'	Freq. (PSC 5.8M)	*New English* (Kriel 1976[4])	*Popular* (Kriel 1988[3])	*Pukuntšu* (Kriel & Van Wyk 1989[4])	*Sediba* (Lombard *et al.* 1992)	*New Sepedi* (Prinsloo & Sathekge 1996)
1 & 3	*mogolo*	2018	✓	—	✓	✓	✓
2	*bagolo*	1040	—	—	✓	✓	✓
4	*megolo*	274	—	✓	—	✓	✓
5	*legolo*	667	—	—	✓	✓	✓
6	*magolo*	509	✓	✓	✓	✓	✓
7	*segolo*	504	✓	✓	✓	✓	✓
8 – 10	*kgolo*	2242	✓	✓	✓	✓	✓
14	*bogolo*	921	✓	✓	✓	✓	✓
15 – 18	*gogolo*	35	✓	—	✓	✓	—
(stem)	*-golo*		*golo*	*golo*	*-golo*	—	—

From Table 2 it is clear that *Sediba* is a word-based dictionary: the stem was not entered while all forms were. *New Sepedi* is also a word-based dictionary, and frequency considerations were used in selecting the forms to be entered in the dictionary. This can be seen through a comparison of the last column with the third, where frequency counts derived from the current 5.8-million-word

[2] A thorough exposition of the lexicographic treatment of days in Sepedi can be found in De Schryver & Lepota (2001).

Pretoria Sepedi Corpus (PSC) are shown. *New English*, *Popular* and *Pukuntšu*, however, display a mixed system, entering both the stem and *some* of the forms. Especially the first two, where the stem is not differentiated in any way from the full forms (as the stem is not preceded by any marker like '-'), are totally inconsistent as far as the choice of canonical forms is concerned. In addition, one cannot but deplore the fact that in dictionaries with such a haphazard approach, precious space is allocated to forms which are unlikely to be looked up by the target users (e.g. *gogolo* in *New English*) whilst highly used forms (e.g. *bagolo* in *New English*) have been omitted.[3]

3. A Perspective on Corpus-based Activities versus Intuition-based Compilations by Lexicographers

It is clear that if the lexicographers had based their lemma-sign lists on frequency counts, the inconsistencies observed in § 2 could have been avoided. It should therefore not come as a surprise that we firmly believe in the advantages of utilising a corpus for lemma-sign list creation.[4] Nonetheless, one should not overreact and *assume that alternative methods* for the creation of the lemma-sign list of dictionaries have *no virtues*, or that alternative methods are in principle *per definition* marred by inconsistencies. What is called for is a macrostructural perspective on *corpus*-based activities versus *intuition*-based compilations by lexicographers.[5]

As true corpus disciples, our main claim in this article is indeed rather bold: "While the supremacy of a corpus remains undisputed in compiling a lemma-sign list, a well-planned *combination of a variety of existing lists* that were assembled manually, results in a lemma-sign list with a remarkable internal consistency."

3.1 The Corpus as Arbiter

Before expounding on this claim we must first devote some lines to the arbiter used to monitor the outcome: the *Pretoria Sepedi Corpus* (PSC). As noted above, PSC currently stands at 5.8 million running words (tokens). Some sections of the research reported on below were checked against an earlier PSC, when it stood at 4.0 million tokens. Indeed, PSC being an 'organic corpus' (cf.

[3] The lemmatisation of adjectives in Sepedi is discussed more extensively in Gouws & Prinsloo (1997).

[4] See in this respect for instance De Schryver & Prinsloo (2000b) which deals with the creation of a dictionary's macrostructure taking an electronic corpus as point of departure.

[5] For an example of a microstructural perspective, see Prinsloo & Gouws (2000) where corpus-based examples of use are contrasted with made-up examples.

De Schryver & Prinsloo 2000a: 92) the size and composition of it is in constant evolution. What is important however is that the data derived from PSC are independent from the size and composition of the corpus. In the context of this article this simply means that core vocabulary versus peripheral vocabulary must be constant, or thus that the organic PSC is 'stable'. A comprehensive investigation of the necessary conditions led us to the following conclusion:

> In the case of a Bantu language with the same degree of conjunctiveness / disjunctiveness as Sepedi, it can be expected that well-designed "general corpora" of 2 million running words can be considered to be "stable" for both frequent and less frequent items. Formulated differently, doubling the size of such well-designed corpora will not substantially alter the stability of the "growing organic corpus." (Prinsloo & De Schryver 2001: 101)

In other words, 'stability tests' were carried out on presumably highly used items on the one hand and seldom used items on the other. The outcome, when expressed relative to this article's premises, is: "In corpora of at least two million running words the ratio of peripheral vocabulary to basic vocabulary is constant." All PSC data in this article are derived from corpora at least twice that size.

3.2 Five-Step Collation: On Combs and Missing Teeth

As a point of departure, comparisons were made between the lemma-sign lists of existing Sepedi dictionaries and lemma-sign lists with the same number of items derived from PSC. This quickly revealed that *all* existing manually compiled dictionaries failed in selecting basic vocabulary at the expense of lemma signs with extremely low or even zero counts. (It is appropriate to note that all Sepedi dictionaries published to date, except for the *Pukuntšu ye kgolo* (Ziervogel & Mokgokong 1975), were conceived as user-friendly learners' dictionaries. This means (i) that they are word-based, and also (ii) that, since Sepedi is written disjunctively, corpus types (i.e. the unique corpus items) can be *directly* equated with dictionary canonical (or citation) forms.[6])

This can be illustrated by means of queries performed on the 4.0-million-word PSC. A corpus-orientated lexicographer might for example wish to consider for inclusion in the dictionary, *all* items which occur at least once in a million words, or thus at least 4 times in 4 million words. In the 4.0-million-word PSC there are roughly 30,000 different items with a frequency of at least 4, so the lexicographer would compile a dictionary containing 30,000 articles. To see how one would go about it, we can focus on one random letter, **R**. In the category **R** of PSC there are roughly 900 items with a frequency of 4 or higher.

[6] For a contrast with the situation for English and Afrikaans, as compared to Sepedi, see Prinsloo & De Schryver (*forthcoming*).

R in the dictionary should thus roughly contain 900 items (which represents c. 3% of the dictionary).

As noted at the outset of this article, regardless of size, any general dictionary and certainly any learners' dictionary should at least cover the basic or core vocabulary. We can assume that, in order to provide for this basic or core vocabulary, the top-5000 items from PSC should at least be included. As item number 5000 has a frequency of 44 in PSC, all word-initial **R** items with a frequency of at least 44 should be entered in the dictionary. There are 126 such items. If we compare the inclusion or omission of those items under **R** in currently available dictionaries such as Kriel's *New English* (Kriel 1976^{4}) or Van Wyk's *Pukuntšu* (Kriel & Van Wyk 1989^{4}), we come to the astonishing conclusion that Kriel only included 56% of the top-5000 **R** items. Van Wyk's dictionary is even worse, as only 46% of the top-5000 **R** items were entered in his dictionary. The data for the latter two claims can be verified in Appendix 1.

From this, one would assume that a corpus-based approach is the only sound one. Yet, we experimented with the idea to *carefully combine a variety of existing lists* that were compiled manually – both published and unpublished ones. It is important to stress that this selection must be done with great care, as there is no point, for instance, to include several editions of the same dictionary. One should rather try to use sources by compilers with backgrounds as varied as possible, such as endeavours by anthropologists on the one hand and by linguists on the other.

From the moment we started to experiment with the collation of the macrostructures of different manually compiled dictionaries, we noted that the percentage of basic vocabulary versus peripheral vocabulary increases *substantially* in the combined list. This observation eventually led us to bring together the following five sources:

- **Step 1**: The Northern Sotho – English section of the *Popular Northern Sotho Dictionary* (Kriel, Prinsloo & Sathekge 1997^{4}) || Prinsloo = part-time lexicographer
- **Step 2**: Some 15,000 cards prepared during the past decade by a *Dictionary Committee* at the University of Pretoria || Dictionary Committee = mother-tongue speakers with minimal academic background
- **Step 3**: The Northern Sotho – English section of *The New English – Northern Sotho Dictionary* (Kriel 1976^{4}) + Kriel's own unpublished revision notes for this dictionary || Kriel = amateur dictionary compiler
- **Step 4**: The Noord-Sotho – Afrikaans section of the *Pukuntšu* (Kriel & Van Wyk 1989^{4}) || Van Wyk = linguist
- **Step 5**: The third version of some 50,000 unpublished cards brought together by Van Warmelo in the first half of the 20th century || Van Warmelo = anthropologist

As these five sources were arranged from small to big, we expected the ‘combined lemma-sign list’ to grow for every alphabetical category. What we did not expect however, was to *see a pattern emerge*. In order to illustrate what happens, we can focus on a random section from the alphabetical stretch **R**, shown in Table 3.

Table 3. Five-step collation for a random section from the alphabetical stretch **R**.

F.	PSC	STEP 1	STEP 2	STEP 3	STEP 4	STEP 5	STEPS
	4.0M	Prinsloo	Dictionary	Kriel	Van Wyk	Van	1+2+3+4+5
			Committee			Warmelo	
...	...	...	...	...	...	...	...
25	*rara*	*rara*	*rara*	*rara*	*rara*	*rara*	*Rara*
	—	*rarabana*	—	*rarabana*	*rarabana*	—	*rarabana*
	—	—	—	*rarabane*	*rarabane*	—	*rarabane*
	—	*rarabolla*	*rarabolla*	—	—	—	*rarabolla*
	—	*rarabologa*	*rarabologa*	*rarabologa*	*rarabologa*	—	*rarabologa*
13	*raragana*	—	—	—	—	*raragana*	*raragana*
12	*raragane*	*raragane (go)*	—	—	—	—	*raragane (go)*
	—	—	*rarakana*	—	—	—	*rarakana*
4	*rarakane*	—	—	—	—	—	—
	—	—	—	—	*rarakantšha*	—	*rarakantšha*
5	*rarakanya*	—	—	*rarakanya*	—	—	*rarakanya*
	—	—	—	*rarakanye-tša*	—	—	*rarakanye-tša*
	—	—	—	—	—	*rarama*	*rarama*
	—	—	*raramolla*	—	*raramolla*	*raramolla*	*raramolla*
	—	*raramologa*	*raramologa*	*raramologa*	*raramologa*	*raramologa*	*raramologa*
6	*rarana*	*rarana*	*rarana*	*rarana*	*rarana*	*rarana*	*Rarana*
	—	—	—	*rarane*	*rarane*	—	*rarane*
5	*raranego*	—	—	—	—	—	—
	—	—	*rarankana*	*rarankana*	—	—	*rarankana*
	—	—	—	*rarankane*	—	—	*rarankane*
	—	—	—	—	*rarankga*	—	*rarankga*
5	*rarankgana*	—	—	—	—	—	—
	—	—	—	—	—	*raranolla*	*raranolla*
	—	—	—	*rarantšha*	—	—	*rarantšha*
	—	—	—	*rarantšwe*	—	—	*rarantšwe*
10	*raranya*	—	*raranya*	*raranya*	*raranya*	*raranya*	*Raranya*
	—	—	—	—	—	*raranyetša*	*raranyetša*
43	*rare*	*rarê*	—	*rarê*	*rarê*	*rarê*	*Rare*
9	*rarega*	*rarêga*	*rarêga*	*rarêga*	*rarêga*	*rarêga*	*Rarêga*
4	*raregile*	—	—	*raregilê*	—	—	*raregilê*
74	*rarela*	*rarêla*	*rarêla*	*rarêla*	*rarêla*	*rarêla*	*Rarêla*
...	...	...	...	...	...	...	...

In the column ‘PSC 4.0M’ all the lemma signs that should be considered for inclusion according to the corpus for this section of **R** are enumerated. The actual items included in the lists of the manual compilations of the five steps are shown next to it, each in a separate column. One can successfully make the following analogy. Each step can be seen as an imperfect comb – imperfect, as quite a number of teeth are missing, while too many peripheral teeth have been added. Yet, when the different combs are brought together, some teeth overlap, while missing teeth in one comb are filled by teeth from another comb. The resulting 'combined comb' following the addition of all five steps is shown in the last column. One sees that we arrive at a comb without too many missing teeth. But how good is the combined comb?

If we study the entire letter **R**, we see that, although a huge number of teeth are missing in the different steps, the combination of all imperfect combs results in a near-perfect comb. We saw that Kriel, Step 3, only included 56% and Van Wyk, Step 4, only 46%, yet together with all the lemma signs from the other steps, the resulting list contains an astonishing 97% of the basic vocabulary!

3.3 THE SEPEDI DICTIONARY PROJECT (SEDIPRO)

The observed patterns discussed above were the impetus for a large-scale project, the *Sepedi Dictionary Project* (SeDiPro), in which not one letter, **R**, was collated, but in which the *complete* macrostructure of *all* five manually compiled sources were joined. In the process, the quintessence of the microstructures was also brought together, and as a result, a very different dictionary database emerged, since the outcome combines input from amateurs as well as professionals, linguists as well as anthropologists, and mother-tongue speakers as well as second-language speakers and learners. To have a rough feeling of the overall macrostructural representativeness of the SeDiPro database, we can first compare the alphabetical breakdown in the latter with the corresponding breakdown of the types in the current corpus. As observed above, 'user-friendliness' and 'disjunctiveness' imply that corpus type counts and dictionary citation forms are directly comparable. This comparison is shown in Table 4.

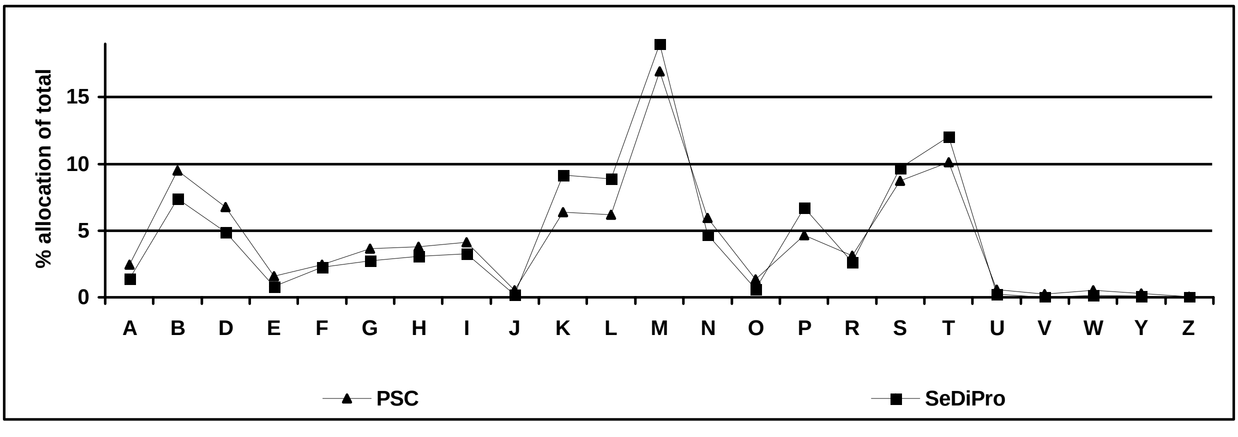
Table 4. Comparing the number of lemma signs in SeDiPro with the types in PSC.

1	2	3	4	5	6	7	8
	PSC 5.8M		DIFFERENCE		SeDiPro		
	# types	% types	abs. %	rel. %	% lemma signs	# lemma signs	
A	3638	2.47	-1.07	**-43.46**	1.40	459	***A***
B	13984	9.49	-2.10	-22.09	7.39	2431	**B**
D	9964	6.76	-1.88	-27.81	4.88	1605	**D**
E	2338	1.59	-0.75	**-47.48**	0.83	274	**E**
F	3645	2.47	-0.24	-9.51	2.24	736	**F**
G	5397	3.66	-0.94	-25.60	2.72	896	**G**
H	5549	3.77	-0.72	-19.08	3.05	1002	**H**

I	6074	4.12	-0.88	-21.42	3.24	1065	**I**
J	798	0.54	-0.34	-63.50	0.20	65	**J**
K	9404	6.38	+2.78	**+43.54**	9.16	3012	**K**
L	9137	6.20	+2.67	**+43.12**	8.87	2918	**L**
M	24937	16.92	+2.06	+12.18	18.98	6242	**M**
N	8742	5.93	-1.23	-20.80	4.70	1545	**N**
O	1995	1.35	-0.71	**-52.38**	0.64	212	**O**
P	6854	4.65	+2.08	**+44.63**	6.73	2212	**P**
R	4566	3.10	-0.45	-14.61	2.65	870	**R**
S	12887	8.74	+0.97	+11.07	9.71	3194	**S**
T	14907	10.12	+1.94	+19.17	12.05	3964	**T**
U	826	0.56	-0.30	-53.88	0.26	85	**U**
V	346	0.23	-0.22	-92.23	0.02	6	**V**
W	814	0.55	-0.39	-70.82	0.16	53	**W**
Y	459	0.31	-0.21	-68.76	0.10	32	**Y**
Z	108	0.07	-0.06	-75.10	0.02	6	**Z**
	147369	99.98			100.00	32884	

From Table 4 one sees that, e.g., PSC allocates 9.49% to **B** whilst SeDiPro allocates 7.39% to **B**, or for **M** 16.92% versus 18.98% respectively, etc. A more explicit comparison between the breakdowns of PSC (Column 3) and SeDiPro (Column 6) is shown in Columns 4 and 5. Column 4 is the difference in absolute terms, Column 5 the difference in relative terms. The data indicate that the categories **A**, **E**, and **O** are under-treated in SeDiPro, and that the categories **K**, **L**, and **P** are over-treated in SeDiPro. (Note that the high ‘rel. %’ values are not that significant for the smaller alphabetical categories.) Although the corpus data could now be used to adjust those alphabetical stretches in the SeDiPro database that are under- or over-treated, the fact of the matter is that the outcome of this bold experiment is rather stunning indeed, since the correlation coefficient *r* between the PSC breakdown and the SeDiPro breakdown is as high as 0.96. This is illustrated graphically in Figure 3.

Figure 3. Relative sizes (in %) of the alphabetical stretches in PSC versus SeDiPro (*r* = 0.96).



3.4 INFORMAL TESTS: RANDOM ONOMASIOLOGICAL FIELDS

Following the completion of the SeDiPro database, and thus after we had observed the excellent correlation between the *overall* macrostructure and the corpus suggestion (§ 3.3), and thus also after we had observed that the 'five-step methodology' ensures that the largest percentage of the *basic* vocabulary ends up in the dictionary (§ 3.2), we decided to conduct yet another set of experiments. For those experiments we simulated the work done in SeDiPro, yet keeping a tight grip on all possible variables. The SeDiPro database can be considered as having the characteristics of a 'general-purpose dictionary', and we wondered how a series of clear-cut *onomasiological fields* – mini terminology lexica in a way – would fare if the five-step collation were applied in their creation. Six different onomasiological fields were chosen: 1. *lenyalo* 'marriage'; 2. *dienywa tša nageng* 'fruit'; 3. *thuto* 'education'; 4. *mebala* 'colours'; 5. *dithaloko* '(traditional / cultural) games'; and 6. *koma* 'initiation'. For each of those fields all the types with a frequency of 'over 5' in the 5.8M PSC were excerpted. The resulting six groups would be used as arbiters in the test. Then five mother-tongue speakers with very different backgrounds (urban vs. rural, young vs. aged, highly schooled vs. little schooling, male vs. female, etc.) were chosen. Each of them was asked to independently jot down all the terms they could come up with in connection with each of the six fields. We then analysed the data much in the same way as for the random section from the alphabetical stretch **R** (Table 3).

As the outcome for all six fields is very similar, we will limit the present discussion to just one of them, namely *koma* 'initiation'. The data of this experiment can be found in Appendix 2, and we will summarise the facts here (with reference to Appendix 2). It is best to start with Column 4. In this column the ticks represent all the 'initiation' terms in PSC that occur at least six times. All these terms are thus serious candidates for inclusion in a dictionary compiled with PSC as arbiter. The terms themselves can be found in Column 3, with an approximate translation in Column 2. Further, Column 1 shows the frequencies, with the format 'singular frequency / plural frequency' for nouns. This left side of Appendix 2 is the 'arbiter'. For the second half of the experiment, it is best to start with Columns 10 down to 6. In those columns all the terms suggested by the five informants have been listed, again using ticks that correspond with Columns 3 and 2. The informants' data are listed from the smallest suggestion (Column 10, 7 terms) to the largest (Column 6, 48 terms). The sum of all these teeth is shown in the comb, Column 5.

Again, the outcome is truly surprising. Although *none* of the informants even comes close to the 102 'initiation' terms suggested by PSC, with Informant 2 coming up with as little as 7 terms, Informant 3 with 10, Informant 5 with 24, Informant 4 with 25, and Informant 1 with 48, the combined comb (SUM) contains as many as 85 terms. Of those 85, 71 occur at least six times in PSC. In other words, the five-step collation brought together 71 of the 102, or 70%, of the *entire* coverage of this one onomasiological field. Focusing on the top-5000 items (where the threshold is minimum 64 in the 5.8M PSC) one sees that a total of 49 items out of 59 were included, or thus 83%.

3.5 The Miraculous Consistency Ratio ‘(x 1.25)^4 = x 2.4’

So far we have seen that carefully monitored collations of five ‘manual/ introspective’ lists result in lemma-sign lists that are sound from the point of view of a random alphabetical stretch, the entire alphabetical breakdown, and random onomasiological fields. We therefore have all good reasons to believe that the entire end product is sound. Yet, the SeDiPro experiment, as bold as it might have seemed initially, revealed even more – much more.

The ultimate observed consistency can be formulated as follows: “For many an alphabetical category, roughly 25% lemma signs were added when moving from one step to the next. Going through the five steps for each alphabetical category meant that the number of lemma signs between Step 1 and Step 5 was multiplied with (1.25)^4 or thus 2.4.” This is shown schematically in Figure 4.

Figure 4. The miraculous consistency ratio ‘(x 1.25)^4 = x 2.4’ (schematically).

$$\underbrace{\text{STEP 1} \;\underset{\text{x 1.25}}{\rightarrow}\; \text{STEP 2} \;\underset{\text{x 1.25}}{\rightarrow}\; \text{STEP 3} \;\underset{\text{x 1.25}}{\rightarrow}\; \text{STEP 4} \;\underset{\text{x 1.25}}{\rightarrow}\; \text{STEP 5}}_{(\text{x } 1.25)^4 = \text{ x } 2.4}$$

More surprisingly, whenever an increase between two steps deviated from 25% within a certain alphabetical category, this deviation was annihilated in the subsequent step(s) of that very alphabetical category. In other words, on average *every* alphabetical category (and hence also the lemma-sign list as a whole) was multiplied with 2.4 between Steps 1 and 5. We came to dub this ‘x 2.4’ the ‘miraculous consistency ratio’. A detailed breakdown of all the increases for the entire alphabet can be seen in Appendix 3. (Note that the sequence of the letters represents the sequence in which the SeDiPro data were assembled.)

For **R**, the increases between the different steps are + 28%, + 27%, +19% and + 24% respectively, so in total x 2.4. The last line shows that this ‘miraculous consistency ratio’ viewed over the sum of all the letters nicely stays within the range 2.4 to 2.5, to end at 2.4.[7]

[7] More accurately, (1.25)^4 = 2.44, so it is logical that the average varies between 2.4 and 2.5.

4. Concluding Remarks

In conclusion we wish to emphasise the main findings. Firstly, it remains true, as expressed by Walter, that "many lexicographers have become used to treating the corpus as the ultimate arbiter on inclusion" (1996: 640). As far as such corpora are concerned, we have pointed out a first consistency: "In corpora of at least two million running words the ratio of peripheral vocabulary to basic vocabulary is constant." We subsequently used such corpora, not as the *ultimate* arbiter, but as instruments to evaluate a non-corpus approach.

Secondly, it remains truly surprising that a variety of manually compiled lists, each of which poorly represents the basic vocabulary, can show so much consistency when combined with one another. The most stunning fact of all is that the end result is actually a fairly good representation of both the basic and the peripheral vocabulary. It seems as if the lacunae of one compiler were accounted for by the other compilers, and so on, and vice versa. The second observed consistency can therefore be formulated as follows: "There is a remarkable consistency per alphabetical category (and hence also for the lemma-sign list as a whole) between 'a combination of various intuitively compiled macrostructures' and 'a corpus-based lemma-sign list'."

We therefore wish to suggest that, in the absence of an electronic corpus – which is the case for all but a few of the Bantu languages – a well-planned combination of a variety of lemma-sign lists of existing dictionaries and unpublished manuscripts, is reasonably representative of a language's basic (and peripheral) vocabulary. We trust that seriously considering the two observed consistencies can truly benefit prospective dictionary compilers.

References

Bussmann, Hadumod. 1996.

Routledge Dictionary of Language and Linguistics (Translated and edited by Gregory Trauth and Kerstin Kazzazi). London: Routledge.

De Schryver, Gilles-Maurice and B. Lepota. 2001.

The Lexicographic Treatment of Days in Sepedi, or When Mother-Tongue Intuition Fails. **Lexikos** 11 (AFRILEX-reeks/series 11: 2001): 1-37.

De Schryver, Gilles-Maurice and D.J. Prinsloo. 2000a.

The compilation of electronic corpora, with special reference to the African languages. **Southern African Linguistics and Applied Language Studies** 18(1-4): 89-106.

2000b *Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The* macrostructure. **South African Journal of African Languages** 20(4): 291-309.

Gouws, Rufus H. and D.J. Prinsloo. 1997.

Lemmatisation of Adjectives in Sepedi. **Lexikos** 7 (AFRILEX-reeks/series 7: 2001): 45-57.

Gove, Philip B. (ed.) 1961[3].
Webster's Third New International Dictionary of the English Language. Springfield: Merriam-Webster.

*Hartmann, Reinhard R.K. (ed.) 1983.
Lexicography: Principles and Practice (Applied Language Studies 5). London: Academic Press.

Hartmann, Reinhard R.K. and Gregory James. 1998.
Dictionary of Lexicography. London: Routledge.

Kriel, Theunis J. 1976[4].
The New English – Northern Sotho Dictionary, English – Northern Sotho, Northern Sotho – English. Johannesburg: Educum Publishers.

1983[3] *Pukuntšu woordeboek, Noord-Sotho – Afrikaans, Afrikaans – Noord-Sotho*. Pretoria: J.L. van Schaik.

Kriel, Theunis J., D.J. Prinsloo and Bethuel P. Sathekge. 1997[4].
Popular Northern Sotho Dictionary, Northern Sotho – English, English – Northern Sotho. Cape Town: Pharos.

Kriel, Theunis J. and Egidius B. van Wyk, 1989[4].
Pukuntšu woordeboek, Noord-Sotho – Afrikaans, Afrikaans – Noord-Sotho. Pretoria: J.L. van Schaik.

Landau, Sidney I. 2001.
Dictionaries: The Art and Craft of Lexicography (2nd edition). Cambridge: Cambridge University Press.

Lombard, Daniel P., Rietta Barnard and Gerhardus M.M. Grobler. 1992.
Sediba, Practical List of Words and Expressions in Northern Sotho, Northern Sotho – Afrikaans – English, English – Northern Sotho / Praktiese lys van woorde en uitdrukkings in Noord-Sotho, Noord-Sotho – Afrikaans – Engels, Afrikaans – Noord-Sotho. Pretoria: Via Afrika.

Martin, Willy J.R., Bernard P.F. Al and Piet J.G. van Sterkenburg. 1983.
On the Processing of a Text Corpus, From textual data to lexicographical information. In Reinhard R.K. Hartmann (ed.), pp. 77-87.

McArthur, Tom. 1989.
The Background and Nature of ELT Learners' Dictionaries. In *Learners' Dictionaries: State of the Art* (Anthology Series 23), Makhan L. Tickoo (ed.), pp. 52-64. Singapore: SEAMEO Regional Language Centre.

Prinsloo, D.J. and Gilles-Maurice de Schryver (eds.) 2000.
SeDiPro 1.0, First Parallel Dictionary Sepêdi–English. Pretoria: University of Pretoria.

2001 *Monitoring the Stability of a Growing Organic Corpus, with special reference to Sepedi and Xitsonga*. **Dictionaries: Journal of The Dictionary Society of North America** 22: 85-129.

forthc. *Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries.*

Prinsloo, D.J. and Rufus H. Gouws. 2000.
The Use of Examples in Polyfunctional Dictionaries. **Lexikos** 10 (AFRILEX-reeks/series 10: 2000): 138-156.

Prinsloo, D.J. and Bethuel P. Sathekge. 1996.
New Sepedi Dictionary, English – Sepedi (Northern Sotho), Sepedi (Northern Sotho) – English. Pietermaritzburg: Shuter & Shooter.

Tomaszczyk, Jerzy. 1983.
On Bilingual Dictionaries, The case for bilingual dictionaries for foreign language learners. In Reinhard R.K. Hartmann (ed.), pp. 41-51.

Walter, Elizabeth. 1996.
Parallel Development of Monolingual and Bilingual Dictionaries for Learners of English. In *Euralex '96 Proceedings I-II, Papers submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg, Sweden,* Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström and Catarina R. Papmehl (eds.), pp. 635-641. Gothenburg: Department of Swedish, Göteborg University.

Whitcut, Janet. 1988.
Lexicography in Simple Language. **International Journal of Lexicography** 1(1): 49-55.

Ziervogel, Dirk and Pothinus C.M. Mokgokong. 1975.
Pukuntšu ye kgolo ya Sesotho sa Leboa, Sesotho sa Leboa – Seburu/Seisimane / Groot Noord-Sotho-woordeboek, Noord-Sotho – Afrikaans/Engels / Comprehensive Northern Sotho Dictionary, Northern Sotho – Afrikaans/English. Pretoria: J.L. van Schaik.

Appendix 1. Inclusion / omission of the top-5000 items under **R** in two Sepedi dictionaries.

Freq. (count) (4.0M PSC)	Freq. (%) (4.0M PSC)	Item (4.0M PSC)	*New English* (Kriel 1976[4])	*Pukuntšu* (Kriel & Van Wyk 1989[4])
		126 items	**71 items (56%)**	**58 items (46%)**
4641	0.11	*ra*	*ra*	*ra*
82	—	*radio*	*Radio*	*radio*
131	—	*raga*	*Raga*	*raga*
49	—	*ragaraga*	*Ragaraga*	*ragaraga*
48	—	*rage*	—	—
48	—	*rago*	—	—
49	—	*ragoga*	*ragoga*	*ragoga*
253	—	*raka*	*raka*	*raka*
135	—	*rakgadi*	*rakgadi*	*rakgadi*

47	—	*rakgadiagwe*	—	—
126	—	*rakgolo*	*rakgolo*	*rakgolo*
44	—	*rakile*	—	—
66	—	*rakwa*	—	—
48	—	*ralala*	*ralala*	*ralala*
138	—	*raloka*	*raloka*	*raloka*
95	—	*ramogolo*	*Ramogolo*	*ramogolo*
287	—	*rangwane*	*Rangwane*	*rangwane*
85	—	*rapa*	*rapa*	*rapa*
69	—	*rapaletše*	*rapalêtše*	—
482	0.01	*rapela*	*rapêla*	*rapêla*
54	—	*rapele*	—	—
66	—	*rapelela*	—	—
71	—	*rapeletša*	*rapêlêtša*	—
74	—	*rarela*	*rarêla*	*rarêla*
76	—	*rarolla*	*rarolla*	*rarolla*
3013	0.07	*rata*	*rata*	*rata*
876	0.02	*ratago*	—	—
197	—	*ratana*	—	—
127	—	*ratau*	*ratau*	—
946	0.02	*rate*	—	—
94	—	*ratega*	*ratêga*	—
51	—	*rategago*	*ratêgago*	—
116	—	*ratego*	—	—
58	—	*ratha*	*Ratha*	*ratha*
77	—	*ratharatha*	*ratharatha*	*ratharatha*
162	—	*ratile*	—	—
48	—	*ratilego*	—	—
182	—	*rato*	—	—
137	—	*ratwa*	—	—
64253	1.59	*re*	*Re*	*re*
93	—	*rea*	*Rêa*	*rêa*
2764	0.07	*realo*	*Realô*	*realô*
70	—	*reela*	*Rêêla*	*rêêla*
2196	0.05	*rego*	—	—
568	0.01	*reka*	*Rêka*	*rêka*
69	—	*reke*	—	—
90	—	*rekela*	*rêkêla*	—
90	—	*rekile*	—	—
216	—	*rekiša*	*rêkiša*	*rêkiša*
128	—	*rekwa*	*rêkwa*	—
242	—	*rema*	*rêma*	*rêma*
5768	0.14	*rena*	*rena*	*rena*
2890	0.07	*reng*	*reng?*	*reng*
363	—	*rera*	*Rêra*	*rêra*

261	—	*rereša*	*Rêrêša*	*rêrêša*
85	—	*rerešitše*	—	—
93	—	*rerile*	—	—
60	—	*rerišana*	*rêrišana*	*rêrišana*
69	—	*rerwa*	—	—
441	0.01	*reta*	*Rêta*	*rêta*
60	—	*rete*	—	—
139	—	*retologa*	*rêtologa*	*rêtologa*
83	—	*retologela*	*rêtologêla*	—
192	—	*retwa*	—	—
62	—	*retwe*	—	—
86	—	*rialo*	—	—
2389	0.06	*rile*	*Rile*	—
94	—	*rilego*	—	—
83	—	*ripa*	*ripa*	*ripa*
163	—	*roba*	*rôba*	*rôba*
764	0.02	*robala*	*rôbala*	*rôbala*
77	—	*robalago*	—	—
151	—	*robale*	—	—
49	—	*robatša*	*rôbatša*	*rôbatša*
84	—	*robega*	*rôbêga*	*rôbêga*
452	0.01	*robetše*	—	—
54	—	*robetšego*	—	—
67	—	*robile*	—	—
49	—	*robja*	—	—
174	—	*roga*	*roga*	*roga*
54	—	*rogana*	*rogana*	*rogana*
92	—	*rola*	*rola*	*rola*
48	—	*rolela*	—	—
585	0.01	*roma*	*roma*	*roma*
53	—	*rome*	—	—
336	—	*romela*	*romêla*	—
50	—	*romele*	—	—
53	—	*romelwa*	—	—
76	—	*rometše*	—	—
132	—	*romile*	—	*romilê*
44	—	*romilego*	—	—
197	—	*romilwe*	—	—
59	—	*rona*	*rona*	*rona*
47	—	*rone*	—	—
112	—	*rongwa*	*rongwa*	—
128	—	*roromela*	—	—
107	—	*rotha*	*rôtha*	*rôtha*
85	—	*rothiša*	*rôthiša*	—
46	—	*roto*	*rôtô*	*rôtô*

389	—	*rotoga*	*rotoga*	*rotoga*
128	—	*rotogela*	—	—
45	—	*rotoša*	*rotoša*	*rotoša*
47	—	*rra*	*Rra*	*rra*
80	—	*rrago*	*rrago*	—
232	—	*rragwe*	—	—
94	—	*rua*	*rua*	*rua*
89	—	*ruile*	—	—
180	—	*ruma*	*ruma*	*ruma*
45	—	*rumo*	—	—
52	—	*rumola*	*rumola*	*rumola*
2453	0.06	*ruri*	*ruri*	*ruri*
137	—	*ruriruri*	*ruriruri*	—
543	0.01	*ruta*	*ruta*	*ruta*
50	—	*rutago*	—	—
85	—	*rute*	—	—
71	—	*rutha*	*rutha*	*rutha*
44	—	*rutile*	—	—
46	—	*rutilwe*	—	—
151	—	*rutwa*	—	—
502	0.01	*rwala*	*rwala*	*rwala*
76	—	*rwale*	—	—
84	—	*rwalela*	*rwalêla*	—
59	—	*rwalwa*	—	—
559	0.01	*Rwele*	—	*rwêle*
92	—	*Rwelego*	—	—
75	—	*Rweša*	*rwêša*	*rwêša*

Appendix 2. Five informants versus PSC for the field *koma* ‘initiation’.

1	2	3	4	5	6	7	8	9	10
Freq. PSC 5.8M	Approximate translation equivalent	Term	P S C	S U M	Inf. 1	Inf. 4	Inf. 5	Inf. 3	Inf. 2
(↓*nouns = sg/pl*)		(*number of terms* →)	102	85	48	25	24	10	7
78	come out of a heathen ceremony as an initiate	*Aloga*	✓	✓	—	✓	—	—	—
48/24	first stage(s) of the boys’ circumcision school	*bodika / madika*	✓	✓	✓	✓	✓	—	✓
37/1	first stage(s) of the girls’ initiation school	*bodikane / madikane*	✓	✓	✓	—	—	✓	—

69	second circumcision ceremony of boys	*Bogwera*	✓	✓	✓	✓	✓	—	—
125	get circumcised	*Bolla*	✓	✓	✓	—	—	—	—
38	circumcise	*Bolotša*	✓	—	—	—	—	—	—
47	circumcised; went to initiation school	*Bolotše*	✓	—	—	—	—	—	—
8	circumcised	*Bolotšwe*	✓	—	—	—	—	—	—
138	manhood	*Bonna*	✓	✓	✓	—	—	—	—
62	womanhood	*Bosadi*	✓	✓	✓	—	—	—	—
12	teach at an initiation school	*Dita*	✓	—	—	—	—	—	—
7	simple single bangles of twisted **modula** (= kind of grass) (worn by uninitiated girls)	*Ditsheka*	✓	—	—	—	—	—	—
314/8	snuff	*fola / difola*	✓	✓	✓	—	—	—	—
1724 (Note[8])	sheepskin dress of girls from the initiation school	*Hlaba*	✓	—	—	—	—	—	—
0	become mad	*hlakanahlogo*	—	✓	✓	—	—	—	—
36	exhale	*Huetša*	✓	✓	✓	—	—	—	—
0	fire (at initiation school)	*Kgalatswi*	—	✓	—	—	✓	—	—
297/60	stick(s), cane(s)	*kgati / dikgati*	✓	✓	—	—	—	✓	—
120	~ **ye ntsho** = victim of the initiation ritual	*Kgokong*	✓	—	—	—	—	—	—
50/11	snail(s)	*kgopa / dikgopa*	✓	✓	✓	—	—	—	—
1752/223	entrance(s)	*kgoro / dikgoro*	✓	✓	✓	—	—	—	—
7835/422	king(s)	*kgoši / dikgoši*	✓	✓	✓	—	—	—	—
75	hide	*Khuta*	✓	✓	✓	—	—	—	—
533/298	clothing	*kobo / dikobo*	✓	✓	—	—	✓	—	—
3/2	(on the) scruff(s) of the neck	*kodung / dikodung*	—	✓	—	—	✓	—	—
841/148	initiation school(s)	*koma / dikoma*	✓	✓	✓	✓	✓	✓	✓
0	a process of hunting (by circumcision school boys)	*koma e ya go fula*	—	✓	—	—	—	✓	—
152/30	at the initiation school(s)	*komeng / dikomeng*	✓	—	—	—	—	—	—
670/317	song(s)	*koša / dikoša*	✓	✓	✓	—	—	—	—
38/16	apron(s) of beads worn by girls at the initiation school	*lebole / mabole*	✓	✓	—	—	✓	—	—
46	circumcision rites	*Lebollo*	✓	—	—	—	—	—	—
34	chant (as of the initiation school)	*Leepo*	✓	—	—	—	—	—	—
12/5	boy(s) who passed through the **bodika** (= first circumcision school) and will enter the **bogwera** (= second circumcision school) the following season	*legaola / magaola*	✓	✓	—	—	✓	—	—
7/0	sleeping mat(s)	*legoga / magoga*	✓	✓	—	—	✓	—	—

[8] The high frequency count belongs to *hlaba* ‘stab’.

10/5	abusive and obscene song(s) sung by **baditi** (= teachers at the initiation school) at women who badly cooked their porridge	*legwete / magwete*	✓	—	—	—	—	—	—
17/18	girl(s) shortly before initiation	*leisa / maisa*	✓	—	—	—	—	—	—
10/39	native convert(s)	*lejakane / majakane*	✓	✓	—	—	✓	—	—
26/11	stick(s) carried by graduate from initiation school	*lekgai / makgai*	✓	✓	—	✓	✓	—	—
129/31	crupper(s)	*lekgeswa / makgeswa*	✓	✓	—	✓	✓	—	—
34/31	girl(s) dressed in reeds in the initiation school	*lepono / mapono*	✓	—	—	—	—	—	—
60/95	uncircumcised boy(s)	*lešoboro / mašoboro*	✓	✓	✓	—	—	—	✓
302/4	uninitiated girl(s)	*lesoka / masoka*	✓	—	—	—	—	—	—
46/97	uninitiated young girl(s)	*lethumaša / mathumaša*	✓	✓	✓	—	—	—	✓
2/0	circumcision school(s) for boys	*letshelapše / matshelapše*	—	✓	—	✓	—	—	—
94/15	red clay	*letsoku / matsoku*	✓	✓	✓	✓	✓	✓	✓
145	suitable / entitled to marry a woman	*Lokela*	✓	✓	✓	—	—	—	—
1740	blood	*Madi*	✓	✓	✓	—	—	—	—
13	ceremonies (at the initiation school)	*madingwana*	✓	—	—	—	—	—	—
41	(*interjection of astonishment used by circumcised men*)	*Mafefo*	✓	—	—	—	—	—	—
16	several circumcision lodges	*Magwera*	✓	—	—	—	—	—	—
0	grass woven like a chain	*malepeletšane*	—	✓	—	—	✓	—	—
376	winter	*Marega*	✓	✓	✓	—	—	—	—
32	drum used at the initiation ceremony for girls	*Mašupšane*	✓	—	—	—	—	—	—
202	head of initiation school	*Matlala*	✓	—	—	—	—	—	—
30/3	boy(s) at the circumcision school	*modika / badika*	✓	✓	—	✓	—	—	—
18/6	initiate(s); boy(s) at the circumcision school	*modikana / badikana*	✓	—	—	—	—	—	—
140/176	teacher(s) at the initiation school	*modiši / badiši*	✓	—	—	—	—	—	—
81/81&49	initiated young man/men who serve(s) as teacher(s) at the initiation school	*moditi / baditi & mediti*	✓	✓	—	✓	✓	✓	—
1/66	bowl(s)	*mogopo / megopo*	✓	✓	✓	—	—	—	—
1295/971	boy(s) in the second stage of the circumcision school	*mogwera / bagwera*	✓	✓	—	✓	—	—	—
1	boy at the circumcision school who has already been circumcised	*mogweramogolo*	—	✓	—	✓	—	—	—
16/841	kind of thorn tree(s)	*mokga / mekga*	✓	✓	✓	—	—	—	—
561/48	open space(s)	*molaleng / melaleng*	✓	✓	✓	—	—	—	—
2203/885	law(s)	*molao / melao*	✓	✓	✓	—	—	—	—
90/51	valley(s), river bed(s)	*molapo / melapo*	✓	✓	—	✓	—	—	—
1589/188	fire	*mollo / mello*	✓	✓	✓	—	—	—	—

114	initiation school; **~ o swele** = the die is cast	*Moloto*	✓	—	—	—	—	—	—
3	leg of a locust	*Monoto*	—	✓	—	✓	—	—	—
10/1	king’s child(ren) who is/are the first one(s) to get circumcised at the initiation school	*morobe / merobe*	✓	✓	—	✓	—	—	—
48/13	place(s) where male initiates stay	*moroto / meroto*	✓	✓	✓	—	✓	—	—
0/0	boy(s) at the circumcision school	*morwabahwibitšana / barwabahwibitšana*	—	✓	—	✓	—	—	—
2066/5	capital(s)	*mošate / mešate*	✓	✓	✓	—	—	—	—
2/1	a roaming about	*mosebetho / mesebetho*	—	✓	—	✓	—	—	—
764/577	boy(s)	*mošemane / bašemane*	✓	✓	✓	—	—	—	—
389/508	girl(s)	*mosetsana / basetsana*	✓	✓	✓	—	—	—	—
3452/756	village(s)	*motse / metse*	✓	✓	✓	—	—	—	—
434/52	native regiment(s) bearing the distinctive name of its/their initiation group(s)	*mphato / mephato*	✓	✓	—	—	✓	—	—
2119/317	meat	*nama / dinama*	✓	✓	✓	—	—	—	—
1569/449	witchdoctor(s)	*ngaka / dingaka*	✓	✓	✓	—	—	—	—
308/9208 (Note[9])	girl(s) undergoing initiation rites	*ngwale / bjale*	✓	✓	✓	✓	—	—	—
194/34	taboo(s)	*ntepa / dintepa*	✓	✓	—	—	✓	—	—
17/10	leader(s)	*ntona / mantona*	✓	✓	✓	—	—	—	—
2199 (Note[10])	organisation of initiation ceremony	*Ntšha*	✓	✓	✓	—	—	—	—
448	**kgokong ye ~** = victim of the initiation ritual	*Ntsho*	✓	—	—	—	—	—	—
28	strike one another	*Otlana*	✓	✓	—	—	—	✓	—
7/2	cairn(s) (= mound(s) of rough stones) erected by caretakers in the initiation school	*phišana / diphišana*	✓	—	—	—	—	—	—
475/915	animal(s)	*phoofolo / diphoofolo*	✓	✓	✓	—	—	—	—
35	head of a circumcision lodge (not the operator); expert, initiation master	*Rabadia*	✓	✓	—	✓	✓	✓	✓
107	cut	*Ripa*	✓	✓	✓	—	—	—	—
15	go through the initiation ceremonies	*Rupa*	✓	—	—	—	—	—	—
29	become swollen	*Ruruga*	✓	✓	✓	—	—	—	—
12/29	pupil(s) who return(s) from the initiation school	*Sealoga / dialoga*	✓	✓	—	—	✓	✓	—
43/9	small bowl(s)	*segwana / digwana*	✓	✓	✓	—	—	—	—

[9] The high frequency count belongs to *bjale* ‘now’.

[10] The high frequency count belongs to *ntšha* ‘take out’.

276	band of boys circumcised together	*Segwera*	✓	—	—	—	—	—	—
465/571	medication	*sehlare / dihlare*	✓	✓	✓	—	—	—	—
399/222	secret(s)	*sephiri / diphiri*	✓	✓	✓	—	—	—	—
7/7	second initiation lodge(s)	*serotha / dirotha*	✓	✓	—	—	✓	—	—
58/2	initiation school(s) for girls	*sešane / dišane*	✓	✓	—	✓	—	—	—
0/2	pupil(s) at the initiation school	*setleetlee / ditleetlee*	—	✓	—	✓	—	—	✓
4042/752	community/ies	*setšhaba / ditšhaba*	✓	✓	✓	—	—	—	—
245/13	crupper(s) worn by men	*setsiba / ditsiba*	✓	✓	—	✓	—	—	—
42/440	knuckle bone(s); divination	*taola / ditaola*	✓	✓	✓	—	—	—	—
604/294	at the mountain(s)	*Thabeng / dithabeng*	✓	✓	✓	—	—	—	—
186/40	string skirt(s) of an initiate	*thapo / dithapo*	✓	—	—	—	—	—	—
275	use of traditional medication to prevent witches	*Thekga*	✓	✓	✓	—	—	—	—
0	person who does the operations at the initiation school	*Thipane*	—	✓	—	✓	✓	✓	—
6	initiated girl	*Thojane*	✓	—	—	—	—	—	—
103/23	stick(s), cane(s)	*thupa / dithupa*	✓	✓	✓	—	—	—	—
14	awning / sheet under which the group under circumcision sleeps	*thupantlo*	✓	✓	—	✓	—	—	—
14/13	one of the two leather flaps of a circumcised male's **lekgeswa** (= crupper) for the wearer to sit on; sheepskin dress of girls from the initiation school	*tlhaba / ditlhaba*	✓	—	—	—	—	—	—
2	dove feathers used by women to adorn themselves	*tlhapetsane*	—	✓	—	—	✓	—	—
6/3	grass used to make clothes for boys at the circumcision school	*Tlhokwa / ditlhokwa*	—	✓	—	—	✓	—	—
682	hunt	*Tsoma*	✓	✓	✓	—	—	—	—
21	has been through the initiation school	*Weditše*	✓	—	—	—	—	—	—
1551	get circumcised	*Wela*	✓	✓	—	✓	—	—	—
63	send to the initiation school	*Wetša*	✓	—	—	—	—	—	—

Appendix 3. The miraculous consistency ratio ‘(x 1.25)^4 = x 2.4’.

Category	**R**	**S**	**H**	**G**	**F**	**E**	**O**	**A**	**I**	**N**	**P**	**U**	**V**	**W**	**Y**	**Z**	**T**	**J**	**L**	**K**	**B**	**D**	**M**
# words (*left side*)	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
% (*right side*)	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**
# words in STEP 1	361	1263	432	354	278	127	97	182	548	600	789	31	2	33	23	2	1599	32	1012	1125	1040	929	2818
% increase	**28**	**29**	**27**	**27**	**19**	**20**	**61**	**37**	**33**	**16**	**91**	**55**	**150**	**33**	**26**	**100**	**80**	**53**	**47**	**58**	**39**	**35**	**47**
# words in STEP 2	462	1629	549	449	330	153	156	250	728	693	1508	48	5	44	29	4	2877	49	1492	1772	1448	1256	4148
% increase	**27**	**17**	**22**	**33**	**40**	**39**	**15**	**26**	**35**	**43**	**15**	**38**	**0**	**9**	**3**	**50**	**11**	**22**	**23**	**5**	**27**	**2**	**16**
# words in STEP 3	586	1908	671	597	462	213	180	314	981	991	1727	66	5	48	30	6	3194	60	1837	1869	1836	1277	4824
% increase	**19**	**32**	**22**	**26**	**25**	**9**	**7**	**20**	**4**	**22**	**8**	**12**	**20**	**6**	**3**	**0**	**4**	**0**	**21**	**32**	**10**	**17**	**20**
# words in STEP 4	700	2520	819	750	579	234	193	376	1020	1209	1859	74	6	51	31	6	3309	60	2230	2458	2022	1499	5792
% increase	**24**	**27**	**22**	**19**	**27**	**17**	**10**	**22**	**4**	**28**	**19**	**15**	**0**	**4**	**3**	**0**	**20**	**8**	**31**	**23**	**20**	**7**	**7**
# words in STEP 5	870	3194	1002	896	736	274	212	459	1065	1545	2212	85	6	53	32	6	3964	65	2918	3012	2431	1605	6242
Ratio 1 → 5	**x2.4**	**x2.5**	**x2.3**	**x2.5**	**x2.6**	**x2.2**	**x2.2**	**x2.5**	**x1.9**	**x2.6**	**x2.8**	**x2.7**	**x3.0**	**x1.6**	**x1.4**	**x3.0**	**x2.5**	**x2.0**	**x2.9**	**x2.7**	**x2.3**	**x1.7**	**x2.2**
Total # words STEP 1	361	1624	2056	2410	2688	2815	2912	3094	3642	4242	5031	5062	5064	5097	5120	5122	6721	6753	7765	8890	9930	10 859	13 677
Total # words STEP 5	870	4064	5066	5962	6698	6972	7184	7643	8708	10 253	12 465	12 550	12 556	12 609	12 641	12 647	16 611	16 676	19 594	22 606	25 037	26 642	32 884
Total ratio 1 → 5	**x2.4**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.4**	**x2.4**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.5**	**x2.4**