http://lexikos.journals.ac.za

Lexicographic Treatment of Kinship Terms in an English/Sepedi–Setswana–Sesotho Dictionary with an Amalgamated Lemmalist

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

Abstract: This article describes the lemmatisation and treatment of kinship terms in a proposed English–Sotho, Sotho–English dictionary with an amalgamated lemmalist. The first requirement is to build a list of kinship terminology for the Sotho languages. Secondly, it is necessary in terms of space restriction to determine the most frequently used forms to be lemmatised in such a dictionary. Thirdly, the macrostructure and microstructure of the dictionary should be planned in terms of an amalgamated approach. A short explanation of the amalgamated model will be presented and a schematic illustration of the paternal family tree structure in the Sotho languages is given in the appendix. Specific attention is given to the compilation of the amalgamated lemmalist focusing on absolute cognates and absolute cognates with a difference in form. Finally, where the reduction of huge quantities of terms, e.g. all derived forms of a specific term in all three Sotho languages are at stake, a lexicographic convention will be suggested to sensibly reduce the number of lemmas and to combat redundancy.

Keywords: AMALGAMATED DICTIONARIES, KINSHIP TERMS, SOTHO LANGUAGES, LEXICOGRAPHIC CONVENTION, CLOSELY RELATED LANGUAGES

Opsomming: Die leksikografiese hantering van verwantskapsterme in 'n Engels/Sepedi–Setswana–Sesotho-woordeboek met 'n gealgameerde lemmalys. Hierdie artikel beskryf die lemmatisering en bewerking van verwantskapsterme in 'n voorgestelde Engels–Sotho, Sotho–Engels woordeboek met 'n geamalgameerde lemmalys. Die eerste vereiste is die samestelling van 'n lys van verwantskapsterminologie vir die Sothotale. Tweedens is dit nodig om ter wille van ruimtebesparing die mees gebruiklike vorme te bepaal wat in so 'n woordeboek gelemmatiseer moet word. Derdens moet die makro- en mikrostruktuur van die woordeboek beplan word in terme van 'n geamalgameerde benadering. 'n Kort verduideliking van die geamalgameerde model sal aangebied word en 'n skematiese voorstelling van die paterne stamboomstruktuur in die Sothotale word in die bylaag aangegee. Spesifieke aandag word gegee aan die samestelling van die geamalgameerde lemmalys met die fokus op absolute kognate en absolute kognate met 'n vormverskil. Ten slotte, waar die vermindering van groot hoeveelhede van die terme, byvoorbeeld alle afgeleide vorme van 'n spesifieke term in al drie Sothotale ter sake is,

sal 'n leksikografiese konvensie voorgestel word om die aantal lemmas sinvol te verminder en redundansie te bestry.

Sleutelwoorde: GEAMALGAMEERDE WOORDEBOEKE, VERWANTSKAPSTERME, SOTHOTALE, LEKSIKOGRAFIESE KONVENSIE, NOUVERWANTE TALE

1. Introduction

The aim of this article is to describe the treatment of kinship terms in an English–Sotho, Sotho–English dictionary with an amalgamated lemmalist. The kinship system in the Sotho languages is complicated, (see appendix), and was selected as an object of study in order to test the viability of the amalgamated approach for such complex structures. Prinsloo (2012) distinguishes three categories of kinship terms for Sepedi, i.e. underived single words such as *malome* 'uncle', *rakgadi* 'aunt' and *tate* 'father', derived words such as *malomeagwe* 'his uncle', *morwediake* 'my daughter' and *bomalomeago* 'your uncles' and phrases such as possessive constructions *mogatša wa mokgotse wa ka* 'my brother in law's wife'. He suggested specific lemmatisation strategies to cater for the large number of kinship terms in these categories in Sepedi, including a specific dictionary convention. An attempt to handle kinship terminology for three languages simultaneously is an even greater challenge since quantity wise the number of kinship terms to be lemmatised is threefold and new challenges on macrostructural as well as on microstructural levels come to the fore. The question is whether it is possible to do justice to all three languages in terms of similarities versus differences, following an amalgamated approach.

The first requirement for the lexicographer is to build a list of kinship terminology for the Sotho languages. It is also necessary in terms of space restriction to determine the most frequently used forms to be lemmatised in such a dictionary. In this article an attempt will be made to collect a number of kinship terms for the Sotho languages, i.e. Sepedi, Setswana and Sesotho. Secondly, the frequency of use of Sotho kinship terms in corpora for these languages will be determined. Thirdly, the treatment of Sotho kinship terms in separate English Sepedi/Setswana/Sesotho dictionaries will be studied in order to establish the viability of such an amalgamated approach and in order to suggest model dictionary articles.

The collection of Sepedi kinship terminology is mainly based on Prinsloo and Van Wyk (1992), Setswana on Van Wyk and Haasbroek (1990) and Sesotho on Molalapata (2004) supplemented by terms found in dictionaries and corpora of the Sotho languages. By way of introduction, a short explanation of the amalgamated model will be presented, followed by a schematic illustration of the paternal family tree structure in the Sotho languages. Finally, the formulation of model entries with an amalgamated approach will be presented.

2. The amalgamated model

The design of amalgamated dictionaries is credited to Martin and Gouws (2000) for introducing the concept and also for compiling the first amalgamated dictionary for Afrikaans and Dutch, *Groot Woordeboek Afrikaans en Nederlands* (ANNA).

The ANNA-approach is to provide treatment for what the amalgamated languages have in common first (A|N, A=Afrikaans, N=Nederlands (Dutch)) followed by the treatment of aspects applicable to the specific languages. Consider the article of *ouderwets* 'old fashioned' from ANNA. The article consists of three sections viz. A|N, N and A. Similarities and differences are indicated throughout by the symbols "=" 'equal' and "≠" 'differ' respectively.

ouderwets bnw., ***ouderwets*** b.nw.
A|N **(v. vroeger) *ouderwets* =** ouderwetse kleren *ouderwetse klere*; een ouderwetse stoomtrein *'n ouderwetse stoomtrein*; ouderwetse opvattingen *ouderwetse opvattings*; hopeloos ouderwets *hopeloos ouderwets* **≠** *stewige ouderwetse meubels* oerdegelijk meubilair
N **(net als vroeger) *outyds, ouwêrelds* =** ouderwetse degelijkheid *outydse deeglikheid*; een ouderwetse winter *'n outydse winter* **≠** het was weer ouderwets gezellig *dit was weer gesellig soos in die ou tyd*
A **(*oulik; slim*) bijdehand =** *'n ouderwetse kind* een bijdehand kind

Detailed discussions of the amalgamated approach and of ANNA in particular can be found in Martin (2012a and 2012b), Martin and Gouws (2000), Marais (2011), Bosman (2013) and in the user's guide of ANNA. Martin's intention with the amalgamated model was also to pave the way for other closely related languages:

> the aim was not only to produce a contrastive dictionary Afrikaans-Dutch, but also to lay the foundation for an *exportable model*, one that could be used for other closely related languages, such as the 'black' languages in South Africa: Xhosa and Zulu, and North-Sotho, South Sotho and Tswana etc. (Martin 2012b: 413)

Amalgamated dictionaries would employ a single lemmalist for closely related languages such as Afrikaans/Dutch, Sepedi/Setswana/Sesotho, isiZulu/isiXhosa/Siswati/isiNdebele and have a unique microstructural architecture in their treatment of the languages in question.

The first requirement for an amalgamated approach is that the languages to be treated should be closely related, i.e. that they should have a substantial number of words in common.

> it can only be applied to closely related languages … both the 'form' of the words (spelling) needs to be the 'same' and at least one of the meanings. … there has to be a sufficient critical mass. (Martin 2012b: 414)

Martin (2012b: 415) puts the overlap between Afrikaans and Dutch as 2/3, i.e. 66.7%.

For the Sotho languages Prinsloo compared the 10,000 most frequently used words in Sesotho, Setswana and Sepedi corpora and came to the conclusion that the vocabulary of these languages overlap to a large extent. The three languages have 19,4% words in common, Sepedi and Setswana share 32,7%, Sepedi and Sesotho 26,9% and Setswana and Sesotho 34,4%.

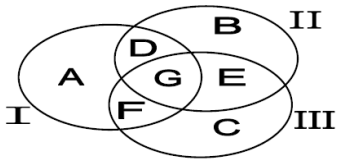


Figure 1: Comparison of the Sotho languages
I = Sepedi; II = Setswana; III = Sesotho
A = 5 978; B = 5 226; C = 5 813; D = 1 333; E = 1 498; F = 746; G = 1 943
(Prinsloo 2005)

This degree of overlap would result in a single amalgamated lemmalist of 22,537 in contrast to a list of 30,000 lemmas (10,000 each for Sepedi, Setswana and Sesotho) if three separate dictionaries were compiled, thus a saving of almost 30%. For the English–Sesotho section the space saving stands at 67% compared to three English sections in three separate dictionaries, i.e. English–Sepedi, English–Setswana and English–Sesotho.

Martin (ANNA: 25) distinguishes five different types of words relevant to an amalgamated approach, i.e. (a) absolute cognates: words in the related languages which are identical in form and meaning, (b) absolute cognates with difference in form, (c) partial cognates: words that differ in at least one sense, (d) non-cognates: words with the same meaning but clear difference in form and (e) false friends: words identical in form but which differ in meaning.

In this article the focus will be on absolute cognates, and cognates with a difference in form.

Absolute cognates are most beneficial to an amalgamated approach because a single lemma represents all of the related languages. For example, the translation equivalents for *woman*, *love* and *neck* in Sepedi, Setswana and Sesotho are identical.

woman *n.* mosadi
love *v.* rata
neck *n.* molala

Likewise, *efe* in all three Sotho languages can be translated with a single equivalent:

efe *enum.* which (one)?

Absolute cognates are discussed in more detail in paragraph 4 below.

Absolute cognates with difference in form, e.g. *kgaet*šedi (Sepedi), *kgaitsadi* (Setswana) and *kgaitsedi* (Sesotho), 'sister/brother' also fit within an amalgamated approach but with some consequences for user-friendliness, cross-referencing and redundancy which will be discussed in more detail below. Partial cognates find their place in an amalgamated approach but require separate treatment for senses where they differ. Non-cognates do not bring much gain in an amalgamated approach since they have to be lemmatised and treated separately. See Prinsloo (2013) for a detailed discussion.

3. Kinship terms in the Sotho languages

An attempt was made to capture single-word kinship terms for Sepedi, Setswana and Sesotho from Prinsloo and Van Wyk (1992), Van Wyk and Haasbroek (1990) and Molalapata (2004) respectively. Their occurrence in the respective corpora was subsequently determined. Finally, a randomly selected number of dictionaries for each of these languages were studied in terms of their lemmatisation and treatment of kinship terms occurring more than once in these corpora.

3.1 Sepedi kinship terms

Single-word kinship terms from Prinsloo and Van Wyk (1992) that occur in the Pretoria Sepedi Corpus (PSC) and in one or more of five randomly selected Sepedi dictionaries are given in table 1 with their frequency counts and inclusion versus omission from the dictionaries marked as "√" and "x" respectively.

Table 1: Sepedi kinship terms

Groot Noord-Sotho-woordeboek, (GNSW), *Sesotho sa Leboa/English Pukuntšu Dictionary* (SLEPD), *Oxford Bilingual School Dictionary: Northern Sotho and English* (ONSD), *Pukuntšutlhaloši ya Sesotho sa Leboa* (PTLH), *Pukuntšu Dictionary* (PUKU1) and *Popular Northern Sotho Dictionary* (POP).

Term	Translation	PSC	GNSW	SLEPD	ONSD	PTLH	PUKU 1	POP
morwa	son	3803	√	√	√	√	√	√
tate	father	1521	√	X	√	√	√	√
mma	mother	1060	√	X	√	√	√	√
morwedi	daughter	1000	√	√	√	√	√	√
ngwetši	daughter in law	637*	√	√	√	√	√	√
kgaetšedi	brother/sister	635	√	√	√	√	√	√
malome	uncle	558	√	√	√	√	√	√
rangwane	father's younger brother	354	√	√	√	√	√	√
rakgadi	father's sister	348	√	√	√	√	√	√
mokgonyana	son in law	260	√	√	√	√	√	√

koko	grandmother	223	√	√	√	√	√	√
rakgolo	grandfather	190	√	√	√	√	√	√
mokgotse	brother/sister in law	189	√	√	√	√	√	√
mogatšaka	my wife/husband	188	X	X	√	X	X	√
motlogolo	grandchild, my sister's child	177	√	√	√	√	√	√
tata	father	169	√	√	X	X	√	√
morwarre	brother	127	√	√	√	√	√	√
motswala	cousin	120	√	√	√	√	√	√
moratho	younger brother or a sister	114	√	√	√	√	√	√
ramogolo	father's older brother	114	√	√	√	√	√	√
mme	mother	5524*	√	√	√	X	√	√
molamo	brother in law	315*	√	√	√	√	√	√
mogolle	older brother/sister	90	√	X	√	X	√	√
mogadibo	brother's wife	82	√	√	√	√	√	√
mmatswale	mother in law	75	√	√	√	√	X	√
mmane	mother's younger sister	61	√	√	√	√	√	√
mogaditšong	co-wife	61	√	√	X	√	√	√
mmamogolo	mother's older sister	57	√	√	X	√	√	√
rra	father	55	X	√	X	X	√	√
morwake	my son	45	X	√	X	√	X	√
samma	younger brother or sister	40	√	√	X	√	√	√
mmangwane	aunt	39	√	√	X	√	√	√
kgaitšedi	brother/sister	38	√	X	X	X	√	√
ngwanangwanaka	my grandchild	35	X	√	X	√	X	X
morwediake	my daughter	34	X	√	X	√	X	X
monyanana	younger brother or sister	28	√	X	X	√	√	√
ratswale	father in law	26	√	√	X	√	√	√
matswale	mother in law	14	√	X	X	X	X	√
motswalake	my cousin	14	X	X	X	X	X	√
mogadikane	co-wife	12	√	√	X	√	√	√
mmamalome	uncle's mother	11	X	X	X	X	X	X
tatemogolo	grandfather	9	X	X	X	X	X	X
nnake	my younger brother/sister	8	√	√	X	√	√	X
rakgolokhukhu	great grandfather	8	√	√	X	√	√	X
morwaka	my son	7	X	X	X	X	X	√
morwediaka	my daughter	7	X	X	X	X	X	X
ngwanangwanake	my grandchild	5	X	X	X	X	X	X
mmakgolo	grandmother	3	X	X	X	X	√	X
mogadikana	co-wife	3	X	√	X	√	X	X

* frequency counts include homonyms which are not kinship terms

3.2 Setswana kinship terms

Single-word kinship terms from Van Wyk and Haasbroek (1990) that occur in the

Pretoria Setswana Corpus (PSETC) and in one or more of five randomly selected Setswana dictionaries are given in table 2.

Table 2: Setswana kinship terms

SESD = *Setswana English Setswana Dictionary*
TYS = *Thanodi ya Setswana: Sefala kgobokgobo*
CSD = *Compact Setswana Dictionary*
TYMYS = *Tlhalosi ya medi ya Setswana*
DSEA = *Dikišinare ya Setswana–English–Afrikaans Dictionary/Woordeboek*

		PSETC	SESD	TYMYS	TYS	DSEA	CSD
mme	mother	33211*	√	√	√	√	√
nna	I	24685	√	√	X	√	√
rre	father	3677	√	√	√	√	X
mma	mother	3175	√	√	√	√	x
rra	father	2760	√	√	√	√	√
morwa	son	2210	√	√	√	√	√
ngwanaka	my child	1128	√	X	√	√	X
koko	grandmother	714	X	√	√	X	X
malome	uncle	551	√	√	√	√	√
ntate	father	365	√	√	√	√	√
ngwanake	my child	351	√	X	√	√	X
mohumagadi	chief's wife	326*	√	√	√	X	X
mogolole	my elder sibling	272	√	√	√	√	X
nkoko	grandmother	262	√	√	√	√	X
mogatsa	spouse	253	√	√	√	√	√
ngwetsi	daughter-in-law	229	√	√	√	√	X
rremogolo	grandfather	220	√	√	√	√	√
morwadi	daughter	219	√	√	√	X	√
nkgonne	elder brother/sister	217	√	√	√	√	X
rangwane	uncle	217	√	√	√	√	√
rakgadi	aunt	176	√	√	√	√	√
mogatsaka	my spouse	169	√	√	√	√	X
mogatsake	my spouse	169	√	√	√	√	X
ntsala	cousin	168	√	√	√	X	X
ausi	older sister	153	√	√	√	X	X
mmangwane	aunt	153	√	√	√	√	√
morwaaka	my son	139	√	√	√	√	X
morwaake	my son	118	√	√	√	√	X
motlogolo	nephew	109	√	√	√	√	X
setlogolo	grandchild	103	√	√	√	x	√
kgaitsadi	brother/sister	90	√	√	√	√	√

nnake	my younger brother	90	√	√	√	√	√
mmemogolo	grandmother	72	√	√	√	√	X
kgaitsadiake	my brother/sister	71	X	√	√	√	X
mogadibo	my husband's sister/my brother's wife	68	X	√	√	√	X
mogwe	son-in-law, brother-in-law	66	√	√	√	X	√
ntatemogolo	grandfather	62	X	√	√	√	√
morwadiake	my daughter	43	√	√	√	√	X
sesi	sister	39	x	√	X	X	√
aubuti	older brother	38	√	√	√	X	√
morwadiaka	my daughter	36	X	√	√	√	X
ntsalake	my cousin	35	√	√	√	√	X
khumagadi	wife of the king	32	√	√	√	√	X
matsale	mother-in-law	22	√	√	√	X	√
leitibolo	firstborn	21	X	√	√	X	√
mmane	mother's younger sister	20	√	√	√	X	X
mmamalome	uncle's mother	14	X	√	√	X	X
mokgonyana	son in law	13	X	X	√	X	X
motswala	cousin	12	X	X	X	X	X
mogadikane	co-wife	10	√	√	X	X	X
mokgwenyana	son in law	8	X	√	X	X	X
mmamogolo	aunt	7	√	√	X	X	√
mogwagadi	man's father/mother-in-law	7	√	√	X	X	√
ntataago	your father	7	X	√	X	√	X
ratswale	father	7	X	X	X	X	√
rramogolo	grandfather	7	√	X	X	X	√
kgantsadi	brother/sister	4	√	√	X	X	X
mogwagwadi	man's father/mother-in-law	4	X	√	X	X	X
sebare	brother-in-law	4	x	√	X	X	√
ratsale	father	2	x	x	X	X	√

* frequency counts include homonyms which are not kinship terms

3.3 Sesotho kinship terms

Single-word kinship terms from Molalapata (2004) that occur in the Pretoria Sesotho Corpus (PSSC) and in one or more of five randomly selected Sesotho dictionaries are given in table 3.

Table 3: Sesotho kinship terms

NSSD = New South Sotho English Dictionary SSED = Southern Sotho English Dictionary
LSS = Longman Sethantšo sa Sesotho BUKAN = Bukantswe http://bukantswe.sesotho.org/
FREEL = http://www.freelang.net/dictionary/sesotho.php

		PSSC	NSSD	SSED	LSS	BUKAN	FREEL
mme	mother/(and)	19560*	√	X	X	X	√
mora	son	3463	√	√	X	X	√
monna	husband/(man)	3036	√	√	X	√	√
ntate	father	2007	√	√	√	√	√
moradi	daughter	732	√	√	X	√	√
kgaitsedi	brother/sister	244	√	√	X	√	√
ngwetsi	sister-in-law/daughter-in-law	201	√	X	X	√	√
rangwane	uncle	192	X	X	X	√	√
malome	uncle	188	√	√	√	√	√
mmangwane	aunt	141	√	X	X	√	√
moholwane	older sister/brother	133	X	X	X	√	√
nnake	my little sister	90	X	√	X	X	X
rakgadi	aunt	88	X	X	X	√	√
motswala	cousin	67	X	√	X	√	√
moena	younger brother/sister	48	X	√	X	√	√
setloholo	grandchild	47	√	X	X	√	√
kgorula	youngest child	46	X	X	X	√	√
motjhana	nephew/niece	43	X	√	X	√	√
ntatemoholo	grandfather	43	√	√	√	√	√
mohwehadi	mother-in-law	33	X	√	X	√	√
mohwe	father-in-law	27	√	√	X	√	√
rangoane	uncle	15	X	X	√	X	X
matsale	mother-in-law	14	√	√	√	√	√
molamo	brother	4	X	X	√	X	X

* frequency counts include homonyms which are not kinship terms

From tables 1 to 3, although they do not reflect the full scope of kinship terms and their derivations, it is clear that the lexicographer has to deal with quite a number of kinship terms in each of the Sotho languages. The question is whether an amalgamated approach will be able to do justice at least to the frequently used terms on both macrostructural and microstructural levels.

4. Macrostructural considerations

On the English side of the proposed bi-directional dictionary, a single lemma-

list is given instead of three English lemmalists for three separate English–Sepedi, English–Setswana and English–Sesotho dictionaries. Once again a 67% reduction is possible because the English lemmalist of kinship terms is presented only once. The challenge, however, is the compilation of the lemmalist on the Sotho side where the model requires amalgamation of the three separate lemmalists for Sepedi, Setswana and Sesotho into a single lemmalist. The five types of cognates identified in ANNA have different implications for the model and the first two are considered here. As briefly stated above, a*bsolute cognates* are the most beneficial because a single lemma can represent all three languages. Consider also the following frequently used and identical absolute cognates in the three Sotho languages.

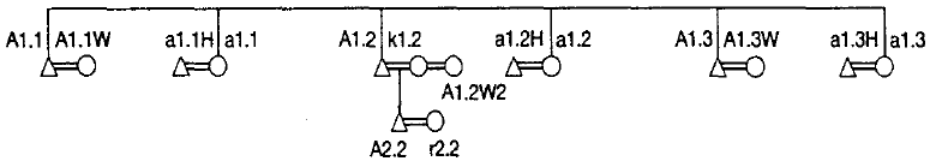
malome mother's younger brother
ntate (my) father
rangwane father's younger brother

In all three Sotho languages *ntate*, *rangwane* and *malome*, respectively, have the same meanings, e.g. *ntate* 'father' in all three languages. Thus in the Sotho to English section space is saved in comparison to three separate dictionaries. *Ramogolo*, *rangwane* and *malome* are lemmatised once instead of three times, each in a Sepedi–English, Setswana–English and Sesotho–English dictionary.

In the English–Sotho section *uncle* is lemmatised only once instead of three times in three separate dictionaries (English–Sepedi, English–Setswana and English–Sesotho).

uncle ramogolo/rremogolo (father's older brother), rangwane (father's younger brother), malome (mother's brother)

Ideally a single term for *uncle* in all three of the Sotho languages would have resulted in additional space saving as in the case of *ntate* '(my) father'. In the case of *uncle* semantic divergence does not lie on the level of differences between the Sotho languages — *ramogolo* 'father's older brother' *rangwane* 'father's younger brother' and *malome* 'mother's brother' have the same meanings respectively in all three Sotho languages. They, however, refer to different relations in terms of the age of the related person and his position in the family tree, consider extracts from the family tree from Prinsloo and Van Wyk (1992) given in the appendix:



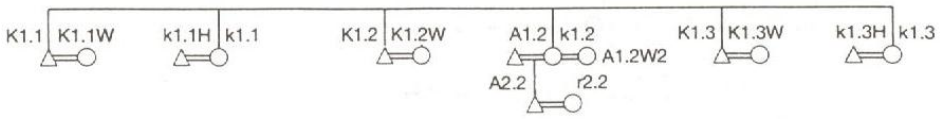


Figure 2: Relationship between a man (A2.2) and his uncles and aunts

Uncle, referring to a man's father's elder brother (A1.1) is *ramogolo*, father's younger brother (A1.3) is *rangwane*. *Uncle*, in reference to mother's younger brother (K1.3) is *malome*. *Ramogolo*, *rangwane* and *malome* are lexicalised terms that could, for lack of equivalents, at best in English be described by means of a paraphrase "a man's father's elder brother", "father's younger brother" and "mother's younger brother".

In the case of *absolute cognates with difference in form*, the first consideration is the presumed knowledge of the target users. The more knowledgeable they are of one or more of the Sotho languages, (a) the more user-friendly an amalgamated lemmalist will be to them, (b) the less problematic it will be for the lexicographer to compile such a lemmalist and (c) the compilation of the lemmalist is less reliant on cross-referencing as lexicographic device to combat decontextualisation brought about by strict alphabetical ordering.

The key consideration, however, for the compilation of an amalgamated lemmalist for this type of cognates is the degree/extent of the difference in form. Martin (2012b: 14) categorises such words as items with a small, systematic spelling or morphological difference or items with a bigger, non-systematic difference but which are still recognizably similar in form. He gives Dutch *pompoen* 'pumpkin' and *pinguïn* 'penguin' versus Afrikaans *pampoen* and *pikkewyn* respectively as examples. There is only a minor difference between *pompoen* and *pampoen* but a substantial difference between *pinguïn* and *pikkewyn*. For the Sotho languages a closer look at the degrees of similarity/difference in spelling is required and will be attempted in a hierarchical order from "very similar" to "more substantial" differences.

The first instance pertains to words which differ only in terms of a diacritic sign, e.g. *s* versus *š*. Setswana and Sesotho use the same word *ngwetsi* 'daughter-in-law' versus Sepedi *ngwetši*. Here the same letter (*s*) occurs — the only difference is *s* with or without the inverted circumflex "ˇ" and there is no need to lemmatise *ngwetsi* and *ngwetši* as separate lemmas with *ngwetši* as a main lemma directly following *ngwetsi* in the vertical layout of alphabetical ordering. This paradigm could simply consist of a presentation indicating the names of all three languages, **ngwetsi**$_{[Set,\ Ses]}$**, ngwetši**$_{[Sep]}$ or an unmarked *ngwetsi* followed by a marked occurrence for *ngwetši*, i.e. **ngwetsi, ngwetši**$_{[Sep]}$. There is also no need for cross-referencing.

The second type of typical examples are words which differ in terms of a single letter. This single letter could be (a) different, or (b) added/omitted. Sepedi and Setswana has *ntatemogolo* 'grandfather' compared to Sesotho *ntatemoholo*

with "*g*" versus "*h*" as the only difference. As for *mogatsaka* versus *mogatšaka* such examples are less problematic — a single lemma paradigm will suffice e.g. **ntatemogolo**[Sep, Set], **ntatemoholo**[Ses]. As for (b), consider Sepedi and Setswana *morwa* versus Sesotho *mora* 'son'. The lexicographer, with the abilities of the target users in mind, must decide whether a user looking for *mora* will find it under *morwa*. In the case of 'daughter' in the Sotho languages both (a) and (b) apply, i.e. Sepedi *morwedi* differs from Setswana *morwadi* in respect of one letter and Sesotho *moradi* in respect of one omitted letter to Setswana. The difference between Sepedi and Sesotho however comprises both (a) and (b) and the question is whether the user looking for the lemma **moradi** will find it under **morwedi**. Given the alphabetical remoteness of "*a*" in *moradi* from "*w*" in *morwedi* (almost at opposite ends in an alphabetical stretch in the dictionary), it could be argued that they should both be lemmatised with cross-reference from the untreated lemma to the treated lemma(s).

Since an alphabetical ordering is followed, the degree of similarity or likelihood of recognition as cognates is influenced by the position inside the word where the differences occur, i.e. at the beginning, middle or end of the word. Spelling differences at the end or even in the middle of words are less problematic, e.g. *ntatemogolo* versus *ntatemoholo* but differences in the first few letters pose a greater risk of the user not finding the lemma e.g. *moradi* versus *morwedi* where the difference lies within the first four letters. Sepedi *ramogolo* and Setswana *rremogolo* also only differ in one instance of (a) and of (b) but although *ramogolo* and *rremogolo* is relatively easily recognisable as cognates when seen together, the user who wants to look up *rremogolo* will probably not see *ramogolo* because *ra-* is alphabetically remote from *rre-*.

The lemma paradigm to be considered often consists of more than one cognate for each of the languages. Consider for example the paradigm for *brother* or *sister* where Sepedi has *kgaetšedi*, *dikgaetšedi*, *bokgaetšedi*, *kgaitšadi*, *bokgaitšadi*, *dikgaitšadi*, *kgaitšedi*, *bokgaitš*edi, *dikgaitšedi* versus Setswana *kgaitsadi*, *bokgaitsadi*, *dikgaitsadi*, *kgantsadi*, *bokgantsadi*, *dikgantsadi* versus Sesotho *kgaitsedi*, *bokgaitsedi*, *dikgaitsedi*. Assuming that the target users have a reasonable knowledge of at least one of the Sotho languages, and know the basic formation rule for singular/plural, these 18 forms could be reduced to *kgaetšedi, kgaitšadi, kgaitšedi, kgaitsadi, kgantsadi* and *kgaitsedi*. This lemma paradigm could sensibly be further reduced on the basis of frequency: *kgaetšedi* (635), *kgaitšedi* (39), *kgaitšadi* (4), *kgaitsadi* (90), *kgantsadi* (4), *kgaitsedi* (244) to *kgaetšedi*, *kgaitšedi* [Sep]; *kgaitsadi* [Set]; *kgaitsedi* [Ses].

For the lexicographer the extent of utilisation of cross-references is in the first place a measurable one. The norm followed in ANNA (Martin 2012b: 419) is that only members of a specific lemma paradigm which are alphabetically more than seven positions away from the lemma paradigm where treatment is given, must be cross-referred to the lemma paradigm. The number of such cross-references represents a redundancy factor against the success of the amalgamated approach because additional dictionary space is utilised for such

lemmas. Formulated differently, the more lemmas required to be entered separately from their lemma paradigms, the less successful the amalgamated approach will be because the ideal is to have a single lemma paradigm for each term for all three languages.

Cross-referencing is, however, intuitive in the sense of the presumed user's ability to find the lemma. In the case of *ntatemogolo* versus *ntatemoholo* it can be assumed that even the less knowledgeable user will be able to find the lemma but in cases such as *ramogolo* versus *rremogolo* the less sophisticated user should be assisted by including *rremogolo* as a lemma with cross reference to *ramogolo*.

5. Microstructural considerations

For short articles, e.g. consisting of little more than translation equivalents, such as for *mme*, *ntate* and *malome* given above, the success of an amalgamated approach is obvious. The question, however, is whether an amalgamated approach is still viable for longer articles.

Returning to the paradigm for *kgaetšedi* compiled above, consider the articles given for Sepedi (GNSW), Setswana (SESD) and Sesotho (SSED).

(GNSW)

KGÁETŠÉDI, (n–)/di– (kgaetšêdi), cf. **KGÁITŠÁDI, KGÁITŠÉDI,** suster [so genoem deur broer], broer [so genoem deur suster] // sister [so called by brother], brother [so called by sister]; *lesogana lê sa êtego le nyala* ~ 'n mens verbreed jou kennis deur te reis // travel broadens the mind
KGÁETŠÉDI, –/bó– (kgaetšêdi) v. **KGÁETŠÉDI, (n–)/di–**

(SESD)

kgaitsadi N. CLS. 1Aø- AND 9N-, same as *kgantsadi*, SING. OF *bokgaitsadi* and *dikgaitsadi*, sibling of the opposite sex; a sister; a brother.

(SSED)

kgaitsëdi (*bö·*) n.cl.1a, *kgaitsëdi* (*di·*) n.cl.5, brother of a woman, sister of a man; my brother, my sister; *kgaitsedie, kgaitsedi a hae*, his sister, her brother; *kgaitsedio*, your sister, your brother. /*kgàìtsédí*/.
bokgaitsëdi, n., state of being a brother or sister.

Consider the following attempts at the compilation of amalgamated articles for kinship terms:

kgaetšedi, kgaitšedi [Sep]**; kgaitsadi** [Set]**; kgaitsedi** [Ses] (**1a**/2b/10) brother (so-called by sister), sister (so-called by brother) ♦ ***lesogana le sa sepelego le nyala ~*** *a young man who does not travel marries his sister*: travel broadens the mind

The lemma paradigm in this example as well as its relatively short article has a high information density which can be paraphrased as follows. First, in terms of comment on form *kgaetšedi*, *kgaitšedi* [Sep]; *kgaitsadi* [Set]; *kgaitsedi* [Ses] account for, compare and contrast, the most frequently used terms for all three of the Sesotho languages. This is indicated by the clear, functional and space saving convention [Sep], [Set] and [Ses] in subscript. Secondly, noun class indication is given by a compact but clear convention. The boldfaced number indicates the class to which the lemmas belong, 2b and 10 the classes in which the plural forms occur. As for comment on semantics, the fact that the lemma can refer to a brother or a sister depending on the gender of the speaker is important and it is neatly explained by the brief contextual guidance given in brackets. Finally, the proverb given as an example of usage is well-selected because it is used in all three languages. The fact that the example is given in only one of the Sotho languages will not be problematic to the target user in this case because the forms are very similar in the other two languages. Thus no need to indicate the languages nor to attempt giving an example for each of the languages. Thus it saves dictionary space, also in terms of examples.

Consider also the suggested articles for *great grandfather* and *rakgolokhukhu*:

great grandfather rakgolokhukhu [Sep], rremogolo/ntatemogolo[Set], ntatemoholo[Ses]

The treatment of the lemma **great grandfather** indicates that Sepedi uses the term *rakgolokhukhu* while Setswana uses *rremogolo* and *ntatemogolo* and that Sesotho also has the latter term with minor spelling variation, i.e. *ntatemoholo*.

rakgolokhukhu [Sep] **1a**/2b great grandfather. Also **mmelega rakgolo** who gave birth to grandfather. → **rremogolo/ntatemogolo**[Set], **ntatemoholo**[Ses]

This is an example where one of the three Sotho languages employs a unique term for a specific relationship while the other two use different terms. The user wants to find the meaning of *rakgolokhukhu* and looks it up under R in the dictionary. The treatment indicates that it is a Sepedi word in class 1a with plural form in class 2b and that the English translation equivalent is *great grandfather*. It also informs him/her of the alternative *mmelega rakgolo* and its literal meaning. Finally in the spirit of the amalgamated approach, i.e. to highlight similarities and differences, an explicit cross-reference by means of the reference marker "→" is given to the reference addresses for the Setswana and Sesotho terms **rremogolo/ntatemogolo**, **ntatemoholo** in the dictionary where more information can be found.

6. Using the convention for lemmatisation of kinship terms in an amalgamated approach

Prinsloo (2012) adapted the original *ga/sa/se* convention (Prinsloo and Gouws 1996) for the reduction of lemma paradigms for kinship terms in Sepedi. He indicated how a complicated set of derivations of *malome* 'uncle' such as *malomeago* 'your uncle', *malomeagwe* 'his/her uncle', *bomalomeabona* 'their uncles', etc. as well as a set of phrases involving *malome* could be reduced to a single lexicographic convention, i.e. **bo/mma/mogatša ~ ago/agwe**. In an amalgamated approach for the Sotho languages the question is whether *three* sets of complex derivations totalling more than 50 options could still be handled by a single convention taking the equivalents for *brother/sister* as a case in point.

Words related to *kgaetšedi* in Sepedi:

bodikgaetšedi
bokgaetšedia
bokgaetšediabo
bokgaetšediagwe
bokgaetšediaka
dikgaetšedi
kgaetšedi
kgaetšedia
kgaetšediabona
kgaetšediago
kgaetšediagogoba
kgaetšediagwe
kgaetšediaka
kgaetšediake
kgaetšediarena
kgaetšeditsebegokwa
ngwanakgaetšediago
ngwanakgaetšediake
polaokgaetšedi
ngwanakgaetšediago
sekgaetšedi

Words related to *kgaitsadi* in Setswana:

bokgaitsadi
bokgaitsadia
bokgaitsadiabone
bokgaitsadiagwe
kgaitsadiabone
kgaitsadiabona
kgaitsadiaka
bokgaitsadiaka
bokgaitsadiake
bokgaitsadie
bokgaitsadio
kgaitsadiake
kgaitsadialona
kgaitsadie
dikgaitsadi
kgaitsadi
kgaitsadia
kgaitsadiarona
kgaitsadio
ngwanaakgaitsadiake

Words related to *kgaitsedi* in Sesotho:

bokgaitsedie
dikgaitsedi
kgaitsedi
kgaitsedia
kgaitsediao
kgaitsedie
kgaitsedinyana
kgaitsedio
kgaitsediyaka
bokgaitsedi
bokgaitsediae
dikgaitsedie

It is simply not possible to lemmatise and treat all derivations of every single kinship term in the dictionary. It could be argued that, depending on the knowledge of the target user, it is possible to sensibly reduce these entire paradigms to a single convention for text reception purposes:

bo/di/se/ kgaetšedi$_{[Sep]}$ **kgaitsadi**$_{[Set]}$**, kgaitsedi**$_{[Ses]}$ **/a/abone/abo/agwe/aka/ago/ arena/arona**

Such a convention requires detailed explanation in the users guide of the dictionary as has been done for the original *ga/sa/se* convention in POP.

7. Conclusion

The lemmatisation and treatment of kinship terms for a bi-directional dictionary bridging English and the Sotho languages in an amalgamated approach poses great challenges to the lexicographer on both the macro and microstructural levels.

On macrostructural level the first step will be to gather all single word basic terms, derived terms and phrases expressing kinship relations for all three Sotho languages and for English. The aim should be to compile a user friendly amalgamated lemmalist and that requires among other, insight and consideration of the presumed knowledge and dictionary using skills of the target user. Against this background of the user perspective the lexicographer should find a sound balance between the compilation of a lemma paradigm covering all three the Sotho languages versus separate lemmas, and utilisation of the medio structure. It is a matter of combating redundancy, i.e. to use less dictionary space for the lemmalist as long as user-friendliness in terms of the skills of the target user is not compromised. It has been argued in detail that the key consideration for the compilation of an amalgamated lemmalist is the degree/extent of the difference in form.

On the microstructural level the aim should be to achieve high text density which is still user-friendly and that clearly brings out differences and similarities between the amalgamated languages. Depending on the size of the dictionary more, or less comment on form and semantics could be given, i.e. longer or shorter articles as long as the information is well-balanced between the languages.

Where the reduction of huge quantities of terms, e.g. all derived forms of a specific term in all of the three Sotho languages is at stake, a lexicographic convention such as the adapted *ga/sa/se* convention could be used to combat redundancy and resolve the impossibility to lemmatise all the relevant forms. Care should, however, be taken that the compilation convention remains user-friendly, i.e. not attempting to include too many derivations.

The compilation of amalgamated dictionaries has great potential for African languages and the foundation laid by Martin's design and the publication

of ANNA is a source of inspiration to apply the model to closely related languages such as the Sotho and Nguni languages.

Acknowledgement

This research is (a) conducted within the SeLA project (Scientific e-Lexicography for Africa), supported by a grant from the German Ministry for Education and Research, administered by the DAAD and (b) supported in part by the National Research Foundation of South Africa (grant specific unique reference number (UID) 85763).

The Grantholder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research are that of the author, and that the NRF accepts no liability whatsoever in this regard.

8. References

Dictionaries

ANNA = Martin, Willy (Ed.-in-chief). 2011. *Pharos Groot Woordeboek. Afrikaans en Nederlands (Prisma Groot Woordenboek Afrikaans en Nederlands).* Cape Town: Pharos.

BUKAN = Bukantswe http://bukantswe.sesotho.org/

CSD = Dent, G.R. and C.L.S. Nyembezi. 1994. *Compact Setswana Dictionary English–Setswana, Setswana–English.* Pietermaritzburg: Shuter & Shooter.

DSEA = Snyman, J.W., J.S. Shole and J.C. le Roux. 1990. *Dikišinare ya Setswana–English–Afrikaans Dictionary/Woordeboek.* Pretoria: Via Afrika.

FREEL = http://www.freelang.net/dictionary/sesotho.php.

GNSW = Ziervogel, D. and P.C. 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.

LSS = Hlalele, J.B. 2005. *Longman Sethantšo sa Sesotho.* Maseru: Longman.

NSSD = Chapole, S.R. 1997. *New South Sotho Dictionary*. Pietermaritzburg: Shuter & Shooter.

ONSD = De Schryver, G.-M. et al. (Eds.). 2008. *Pukuntšu ya Polelopedi ya Sekolo. Sesotho sa Leboa le Seisimane. E gatišitšwe ke Oxford/Oxford Bilingual School Dictionary. Northern Sotho and English.* Cape Town: Oxford University Press.

POP = Kriel, T.J., D.J. Prinsloo and B.P. Sathekge. 1997. *Popular Northern Sotho Dictionary, Northern Sotho–English, English–Northern Sotho*. Cape Town: Pharos.

PTLH = Mojela, M.V. (Ed.). 2007. *Pukuntšutlhaloši ya Sesotho sa Leboa*. Pietermaritzburg: Nutrend.

PUKU 1 = Kriel, T.J. 1983. *Pukuntšu woordeboek.* Pretoria: J.L. van Schaik.

SESD = Matumo, Z.I. 1993 *Setswana–English–Setswana Dictionary.* Gaborone: Botswana Book Centre and Macmillan Botswana Publishers.

SLEPD = Mojela, M.V., M.C. Mphahlele, M.P. Mogodi en M.R. Selokela. 2006. *Sesotho sa Leboa/ English Pukuntšu Dictionary.* Cape Town: Phumelela.

SSED = Mabille, A. and H. Dieterlen. 1988. *Southern Sotho–English Dictionary*. Revised by R.A. Paroz. Morija: Morija Sesotho Book Depot.

TYMYS = Otlogetswe, T.J. 2012. *Tlhalosi ya medi ya Setswana.* Gaborone: Medi Publishing.

TYS = Mareme, G.B. (Ed.). 2008. *Thanodi ya Setswana: Sefala kgobokgobo.* Pietermaritzburg: Nutrend.

Other sources

Bosman, Nerina. 2013. Die gebruik van ANNA in 'n Nederlandse taalverwerwingskursus — toegangsgemak en inligtingskoste. Botha, W., P. Mavoungou en D. Nkomo. 2013. *Festschrift RUFUS H. GOUWS*: 39-54. Stellenbosch: SUN PReSS.

Marais, R. 2011. Een mooie dikke dame. "ANNA", het Groot Woordenboek Afrikaans en Nederlands. *Ons Erfdeel* 1: 190-192.

Martin, W. 2012a. Amalgamated Bilingual Dictionaries. Genis, René et al. (Eds.). 2012. *Between East and West. Festschrift for Wim Honselaar on the Occasion of his 65th Birthday*. *Pegasus Oost-Europese Studies* 20: 437-449. Amsterdam: Uitgeverij Pegasus.

Martin, W. 2012b. ANNA: A Dictionary with a Name (and what Lies Behind it). *Lexikos* 22: 406-426.

Martin, Willy and Rufus Gouws. 2000. A New Dictionary Model for Closely Related Languages: The Dutch–Afrikaans Dictionary Project as a Case-in-point. Heid, Ulrich, Stefan Evert, Egbert Lehmann en Christian Rohrer (Eds.). 2000. *Proceedings of the Ninth EURALEX International Congress*. *EURALEX 2000. Stuttgart, Germany, August 8th–12th, 2000*: 783-792. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Stuttgart University.

Molalapata, B.T. 2004. *The Treatment of Kinship Terminology in Sotho Dictionaries, with Special Reference to Setswana.* Unpublished M.A. Dissertation. Pretoria: University of Pretoria.

Prinsloo, D.J. 2005. *Compiling a Sotho–English/English–Sotho Dictionary: A Viability Study*. Unpublished paper presented at the Tenth International Conference of the African Association for Lexicography, held at the University of the Free State, Bloemfontein, South Africa, 13–15 July 2005.

Prinsloo, D.J. 2012 Die leksikografiese bewerking van verwantskapsterme in Sepedi. *Lexikos* 22: 272-289.

Prinsloo, D.J. and R.H. Gouws. 1996. Formulating a New Dictionary Convention for the Lemmatization of Verbs in Northern Sotho. *South African Journal of African Languages* 16(3): 100-107.

Prinsloo, D.J. and J.J. van Wyk. 1992. Verwantskapsterminologie van die Noord-Sotho. *South African Journal of Ethnology* 15(2): 43-58.

Van Wyk, J.J. and F.T. Haasbroek. 1990. Verwantskapsterminologie van die Batswana. *South African Journal of Ethnology* 13(4): 159-179.

Appendix