Southern African Linguistics and Applied Language Studies 2016, 34(2): 169–185

SOUTHERN AFRICAN LINGUISTICS
AND APPLIED LANGUAGE STUDIES
ISSN 1607-3614 EISSN 1727-9461
http://dx.doi.org/10.2989/16073614.2016.1206478

A corpus-driven account of the noun classes and genders in Northern Sotho

Elsabé Taljard[1]* and Gilles-Maurice de Schryver[1,2]

[1]Department of African Languages, University of Pretoria, Pretoria, South Africa

[2] BantUGent – UGent Centre for Bantu Studies, Department of Languages and Cultures, Ghent University, Belgium

**Corresponding author email: elsabe.taljard@up.ac.za*

Abstract: This article offers a distributional corpus analysis of the Northern Sotho noun and gender system. The aim is twofold: first, to assess whether the existing descriptions of the noun class system in Northern Sotho are corroborated by information provided by the analysis of a large electronic corpus for this language, with specific reference to singular-plural pairings, and second, to present a number of novel visualisation aids to characterise a noun class system (in a radar diagram) and a noun gender system (using a two-directional weighted representation) for Northern Sotho in particular, and for any Bantu language in general. The findings include the discovery of two new genders in Northern Sotho (i.e. class pairs 1/6 and 3/10), and also indicate that the Northern Sotho noun class system, and by extension any one for Bantu, should be seen as dynamic.

Introduction

The present undertaking adds to a relatively small, but growing number of corpus-based grammatical descriptions of Bantu language features, several of which regard South African languages. These include the first corpus-based diachronic investigation of a linguistic phenomenon in a Bantu language, i.e. of the Zulu locative prefix *ku-* (de Schryver & Gauton 2002), a study of the semantic import of the Zulu nominal suffix *–kazi* (Gauton et al. 2004), a detailed analysis of the semantics and combinatorial properties of the higher-order locative *n*-grams in Northern Sotho (de Schryver & Taljard 2006), and an investigation into the historical relationship between members of the class 'adjective' and other word classes, particularly enumeratives and nominal relatives, in Northern Sotho (Taljard 2006). Corpus-based linguistic work having other Bantu languages as foci includes a study of the phonetics of Cilubà (de Schryver 1999), an attempt to discover the actual patterns that govern the use of the class 16, 17 and 18 *amba-* locative relatives in Swahili (Toscano & Sewangi 2005), a quantitative analysis of various aspects of the Lusoga noun (de Schryver & Nabirye 2010), a synchronic exploration into the expression of possibility in Kirundi (Bostoen et al. 2012), a diachronic examination of the semantic evolution of the modal verb *-sóból-* in Luganda (Kawalya et al. 2014), a look at the antipassive/associative polysemy in Cilubà (Dom et al. 2015), and most recently, the reconstruction of the actuation and transmission of a phonological innovation known as prefix reduction in Kikongo (Bostoen & de Schryver 2015).

These studies have inter alia shown that traditional introspection-based grammatical descriptions can be fruitfully supplemented by corpus-based analyses, since the latter provide the researcher with access to large amounts of real-life language data which can be analysed computationally. In each of these studies, aspects which had hitherto been overlooked by grammarians were uncovered. Over the past fifteen years, the field of Bantu corpus linguistics has also gradually moved from corpus-based to corpus-driven studies (for the distinction, see Tognini-Bonelli 2001).

In order to assess the extent to which existing descriptions of the noun class system in Northern Sotho correspond to data mined from a large electronic corpus for this language, a succinct account of the corpus data is presented below, from which a number of didactic and theoretical implications are drawn.

Noun classes in traditional grammars of Northern Sotho

When the discussion of the noun class system of Northern Sotho in the various standard grammars is perused, it quickly becomes apparent that it is always presented as a straightforward, static and somewhat one-dimensional state of affairs. Table 1 summarises the system as found in four standard grammars, as well as in the grammar included in the front matter of the *Comprehensive Northern Sotho Dictionary* by Ziervogel and Mokgokong (1975).

It comes as no surprise that there is general consensus among Northern Sotho grammarians as to the noun classes which are distinguished in this language. Differences between grammatical descriptions relate mostly to the non-recognition of some classes. Ziervogel et al. (1969) do not distinguish an infinitive class (class 15), nor do they mention the existence of the *ga-* locative class. Ziervogel and Mokgokong (1975) make no provision for either the *ga-* or the *N-* locative classes. They do, however, distinguish sub-classes for classes 9 and 10, i.e. 9a and 10a, to which loan words are assigned, the reason being that these words do not follow the normal phonological rule of plosivation of the initial consonant (e.g. *rôkô* 'dress' *N-* + r- does not become *N-* + th- > th-; *galase*

Table 1: Noun classes in traditional grammars of Northern Sotho

Cl. #	CP	Example	Ziervogel et al. (1969)	Ziervogel and Mokgokong (1975)	Lombard et al. (1985)	van Wyk et al. (1992)	Poulos and Louwrens (1994)
1	mo-	*mo*sadi 'woman'	✓	✓	✓	✓	✓
1a	Ø-	malome 'uncle'	✓	✓	✓	✓	✓
2	ba-	*ba*sadi 'women'	✓	✓	✓	✓	✓
2b	bô-	*bô*malome 'uncle & Co.'	✓	✓	✓	✓	✓
3	mo-	*mo*nwana 'finger'	✓	✓	✓	✓	✓
4	me-	*me*nwana 'fingers'	✓	✓	✓	✓	✓
5	le-	*le*bone 'light' *le*baka,	✓	✓	✓	✓	✓
	Ø-	baka 'reason; time'	×	×	×	×	✓
6	ma-	*ma*bone 'lights' *ma*baka 'reasons; times' *ma*di 'blood' *ma*dulo 'kinds of residences'	✓	✓	✓	✓	✓
7	se-	*se*lepe 'axe' *se*hlare,	✓	✓	✓	✓	✓
	Ø-	hlare 'medicine; tree'	×	×	×	×	×
8	di-	*di*lepe 'axes' *di*hlare 'medicines; trees'	✓	✓	✓	✓	✓
9	*N-*	*m*pša 'dog'	✓	✓	✓	✓	✓
	Ø-	hlogo 'head'	✓	✓	✓	✓	✓
10	di*N-*	*dim*pša 'dogs'	✓	✓	✓	✓	✓
	di-	*di*hlogo 'heads'	✓	✓	✓	✓	✓
14	bo-	*bo*hlokwa 'importance' *bo*dulo 'residence'	✓	✓	✓	✓	✓
15	go-	*go* ruta 'to learn'	×	✓	✓	✓	✓
16	fa-	*fa*se 'below'	✓	✓	✓	✓	✓
17	go-	*go*dimo 'above'	✓	✓	✓	✓	✓
18	mo-	*mo*rago 'behind'	✓	✓	✓	✓	✓
*N-**	*N-*	*n*tle 'outside'	✓	×	✓	✓	✓
	Ø-	pele 'in front'	✓	×	✓	✓	✓
24*	ga-	*ga*re 'middle'	×	×	✓	✓	✓

N: homorganic nasal

*The *N-* and *ga-* locative classes are generally not numbered in Northern Sotho grammars, but following Gauton (2000) we refer to this class as the *N-* class, respectively class 24. In Northern Sotho, these two additional classes are, just as is the case for the locative classes 16, 17 and 18, not productive.

'glass' *N-* + g- does not become *N-* + kg- > kg-). This distinction is, however, not generally recognised in Northern Sotho grammars. Furthermore, Poulos and Louwrens (1994: 37) claim that certain nouns in class 9, particularly those expressing kinship, can, in addition to their regular plural in class 10, also take a plural in class 2b, and provide the following examples: *kgaetšedi* 'sister', pl. *dikgaetšedi* and *bôkgaetšedi* 'sisters', and *ngwetši* 'bride', pl. *dingwetši* and *bôngwetši* 'brides'. However, a corpus query reveals that both singular nouns, i.e. *kgaetšedi* 'sister' and *ngwetši* 'bride', actually have double class membership: judging by the concords with which they appear in sentences, these nouns belong to both classes 1a and 9, thus the plurals in class 2b would represent the regular plural of the class 1a singular forms and the class 10 plurals the regular plural of the class 9 singular forms. Furthermore, it needs to be noted that the forms in class 2b express collective plurality.

The underlying assumption in all Northern Sotho grammars is that there generally is a one-to-one correspondence between any singular class and its plural counterpart, as may be seen from Table 1, classes 1 to 10. The only 'irregularities' that are pointed out, are the following:

- Class 5 nouns with plural forms in both classes 6 and 10, e.g. *lenaka* 'horn', pl. *manaka* and *dinaka* 'horns'.
- Class 9 nouns with plurals in classes 10 and 6, e.g. *nku* 'sheep', pl. *dinku* 'sheep' and *manku* 'flocks of sheep'.
- Class 14 nouns which take their plurals in class 6: *bogobe* 'porridge', pl. *magobe* 'kinds of porridge'.
- Grammarians agree that in the case of the class pairings 9/6 and 14/6, the plural forms in class 6 have a collective meaning (as may also be derived from the glosses presented at the examples in the previous two points).
- Class 15, as the infinitive class, has no plural, nor have nouns in any of the locative classes.

As will be argued below, such a one-dimensional description is an oversimplification of the linguistic reality as evidenced by the corpus data.

A corpus-driven account of the noun classes in Northern Sotho

A Northern Sotho corpus of 6.9 million running words (i.e. tokens) was queried, consisting of roughly 155 000 different orthographic words (i.e. types). Each type with a frequency of at least 69 (or thus an occurrence of 'once in every one hundred thousand words') was extracted, of which there were 4 980. From that list, all nouns were selected, which resulted in a noun list containing 1 769 singular nouns and 795 plural nouns.

The overall picture of the different type categories (still in the top-frequent section of the 6.9 million word Northern Sotho corpus, i.e. using a threshold of 69) is as shown in Table 2.

Table 2: Type categories in the top-section of the Northern Sotho corpus used

Type category	*N*	%
noun	2 564	51.49
verb	1 853	37.21
adjective	131	2.63
adverb	76	1.53
pronoun	52	1.04
demonstrative	47	0.94
demonstrative copulative	13	0.26
concord	42	0.84
particle	34	0.68
TAM (tense, aspect, mood marker)	16	0.32
conjunction	31	0.62
interjection	26	0.52
enumerative	12	0.24
ideophone	4	0.08
other	79	1.59
SUM	4 980	100.00

These categories correspond roughly to the parts of speech that are distinguished for Northern Sotho. Note that since Northern Sotho is a disjunctively written Bantu language, categories are assigned to orthographic units, and not to linguistic units.

Focusing on the nouns, each class is as shown in Table 3. From Table 3 it may be seen that as many as one quarter (25.35%) of all the noun types in the top-section of the Northern Sotho corpus, are nouns in class 9 (650 nouns, out of a total of 2 564 nouns).

Also in Table 3, the exact number of occurrences of each noun type has been specified. For example, the 650 noun types in class 9 have a combined frequency of over a quarter million (265 869 tokens). The summed overall frequency for all the nouns considered in this study is 1 087 219, which thus means that expressed in terms of noun tokens, class 9 nouns also represent as many as a quarter (24.45%) of all noun tokens.

Actually, there is a surprisingly good correlation between the noun type distribution across all the classes on the one hand, and the noun token distribution on the other: the Pearson product moment correlation coefficient *r* between the two series is as high as 0.969. On the whole, this may be interpreted as follows: frequency of usage of a noun is not dependent on the particular noun class that noun is in; on the contrary, each noun has on average the same chance to occur (even though some nouns are of course used much more frequently than others, while others are used far less frequently).

The noun type distribution across the various noun classes of Northern Sotho is shown graphically in the radar diagram of Figure 1; that for the noun token distribution is shown in Figure 2.

To the best of our knowledge, a representation such as the one shown in Figures 1 and 2 for the distribution of the various Bantu noun classes has never been suggested in the scientific literature before. It is, however, highly didactic, as it instantly gives an indication of the relative size of each noun class, as realised in natural language usage.

The prominence of class 9 is rather surprising, both in terms of intuitive language knowledge and in terms of the topicality hierarchy, formulated by Givón (1976: 152). He presents this hierarchy as a set of distinct, but interacting hierarchical relations:

a. HUMAN > NON-HUMAN
b. DEFINITE > INDEFINITE
c. MORE INVOLVED PARTICIPANT > LESS INVOLVED PARTICIPANT
d. 1ST PERSON > 2ND PERSON > 3RD PERSON

Items that are higher in the hierarchy are more likely to be topics within discourse and are more likely to be talked about, reflecting the anthropocentric nature of discourse, or in the words of Hawkinson and Hyman (1974: 161), the fact that what people usually talk about are other people. Morolong and Hyman (1977: 215) state that 'human beings necessarily have greater prominence over non-humans, since they typically bring about, receive and are the beneficiaries of actions'. These different hierarchies are furthermore interrelated: human items are more likely to represent definite information and tend to be more involved participants than non-human items. In terms of this hierarchy, it is to be expected that class 1 nouns would be most frequent, which clearly is not the case—not on type level, and not even on token level. In order to rule out the possibility that this may be an idiosyncratic trait of Northern Sotho, and since comparable data is available for Lusoga (an eastern interlacustrine Bantu language spoken in Uganda), a near-direct comparison may be undertaken. The data provided in de Schryver and Nabirye (2010:

Table 3: Distribution of the noun classes in the top-section of the Northern Sotho corpus used

Class	1	2	1a	2b	3	4	5	6	7	8	9	10	14	16	17	18	N-	24	SUM
Types (*N*)	176	106	42	9	182	110	251	288	210	106	650	245	162	4	3	5	13	2	2 564
Type %	6.86	4.13	1.64	0.35	7.10	4.29	9.79	11.23	8.19	4.13	25.35	9.56	6.32	0.16	0.12	0.20	0.51	0.08	100.00
Tokens (Freq.)	101 426	62 570	26 993	1 245	88 459	36 626	89 425	110 568	64 118	33 338	265 869	67 851	65 024	6 451	6 749	20 208	29 217	11 082	1 087 219
Token %	9.33	5.76	2.48	0.11	8.14	3.37	8.23	10.17	5.90	3.07	24.45	6.24	5.98	0.59	0.62	1.86	2.69	1.02	100.00

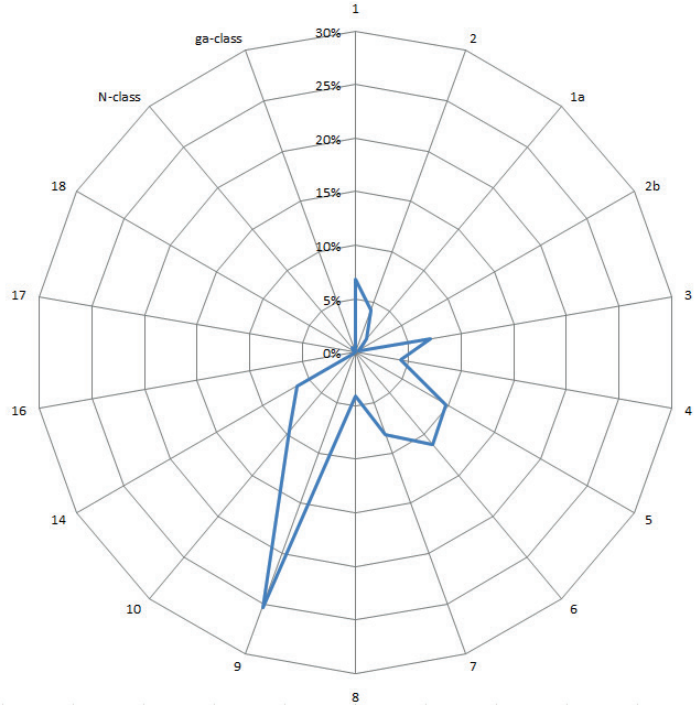

Figure 1: Noun type distribution (expressed in %) in the top-section of the Northern Sotho corpus used

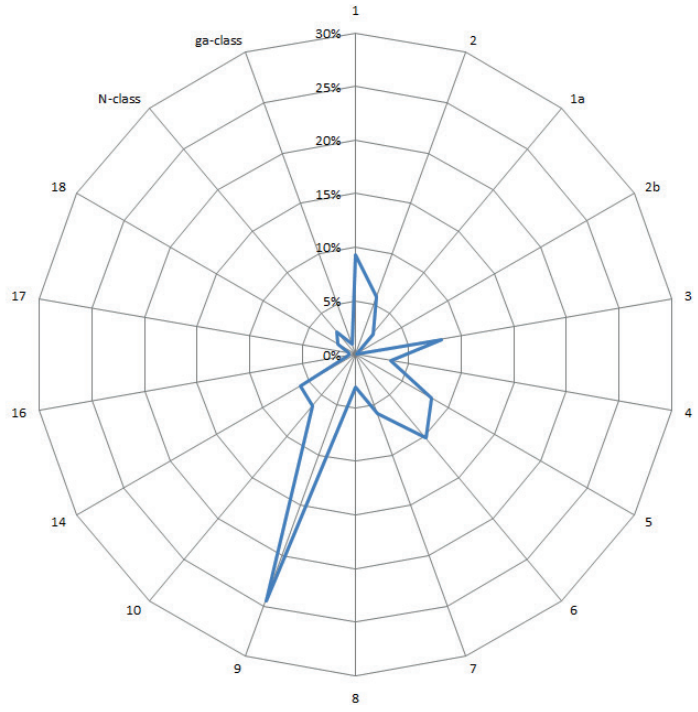

Figure 2: Noun token distribution (expressed in %) in the top-section of the Northern Sotho corpus used

104–105) may be reprocessed and represented in a similar way, at which point Figures 3 and 4 are obtained.

In Figure 3, which represents the type level, the similarly outsized class 9 is also immediately apparent. However, when considering the token level as seen in Figure 4, whereby one thus takes the actual occurrences of all the nouns into account, one notices that Lusoga *does* conform to expectation; the sum of the occurrences in the Lusoga classes 1, 2, 1a and 2a even make up 31.38%, as compared to 15.91 % for the sum of the occurrences in the Lusoga classes 9 and 10. For Northern Sotho, the corresponding values are 17.68% for the sum of the occurrences in classes 1, 2, 1a and 2b, vs. 30.69% for the sum of the occurrences in classes 9 and 10. By and large, the Northern Sotho token data (17.68% vs. 30.69%) is the reverse of the Lusoga token data (31.38% vs. 15.91%). This is all the more surprising as the correlation coefficient *r* between the Northern Sotho and Lusoga type distributions for the corresponding classes (i.e. classes 1 to 10, and 14 and 16) is still a respectable 0.771.

Arguing within the framework of the topicality hierarchy and taking into consideration that human nouns are not restricted to class 1, a possible explanation for this perceived anomaly in Northern Sotho could be that the feature [+HUMAN] is obscured by the class 9 membership of nouns, and that a sizeable number of these nouns could actually be human nouns, which would then explain their high frequency. An analysis of class 9 nouns in terms of the feature [+HUMAN] vs. [-HUMAN] reveals that the feature [+HUMAN] is found in 22.3 out of the 650 noun types in class 9 (i.e. in 9/10, 9/- and 9/6), or thus for only 3.43% of class 9 nouns.[1] Taking the occurrence of each of those class 9 noun types into account, or thus their actual frequency in the corpus, the [+HUMAN] feature is still only apparent in 19 409 out of the 265 869 class 9 tokens, or thus in 7.30% of the cases. Clearly, then, the high frequency of class 9 nouns cannot be accounted for in terms of Givón's topicality hierarchy and no other clear reason currently presents itself for this state of affairs.

A corpus-driven account of the noun genders in Northern Sotho

The term 'noun gender' which is used in Bantu linguistics is generally not used in the description of the South African Bantu languages; the term 'noun class' being the preferred one. For the purpose of this study, the term 'gender' is used to refer to what is otherwise known as a class pair. To illustrate: the word *monwana* 'finger' belongs to class 3. Its plural is *menwana* 'fingers'; therefore the words *monwana* and *menwana* belong to gender 3/4. A further refinement is made in cases where a noun has—in terms of its morphology—either only a plural or a singular form: a noun such as *marega* 'winter' which, based on its morphological features is assigned to class 6, is consequently assigned to a single-class gender, i.e. gender 6.

The distribution of the actual number of nouns (in terms of types) for each gender in the corpus is summarised in Table 4, and visualised in Figure 5.

In traditional descriptions of the Northern Sotho noun class system, the element of directionality does not feature. It is tacitly assumed that there is a corresponding plural for every singular noun, and vice versa, with no attention being paid to the actual relationship between singulars and plurals, and between plurals and singulars. This view is reflected in the words of Louwrens (1994: 126) who states as follows: 'Generally speaking, noun classes can be coupled in pairs of which one member of the pair denotes a *singular* meaning, and the other member a corresponding *plural* meaning...'. The only mention that is made of singular forms that have no corresponding plural forms are the obvious candidates, i.e. the infinitive class 15, the locative classes and a number of nouns in class 14. Also mentioned are the so-called *pluralia tantum*, which are found in class 6 only and include liquids (*maswi* 'milk'), abstract concepts (*maatla* 'strength'), and temporal nouns (*maabane* 'yesterday').

When comparing Table 1 (together with the information in the bulleted list above, at the end of the section on the noun-class treatment in traditional grammars of Northern Sotho) with Figure 5, it is immediately apparent that two genders are missing in existing descriptions, i.e. gender 1/6 and gender 3/10. Although the percentage of nouns in class 1 that have a corresponding plural in class 6 and vice versa seems insignificant (1% in both directions), it needs to be taken into consideration that only the top-frequency types were included in the study. The same is valid for gender 3/10. Compare the corpus lines shown in (1) and (2), for nouns that belong to these two genders.

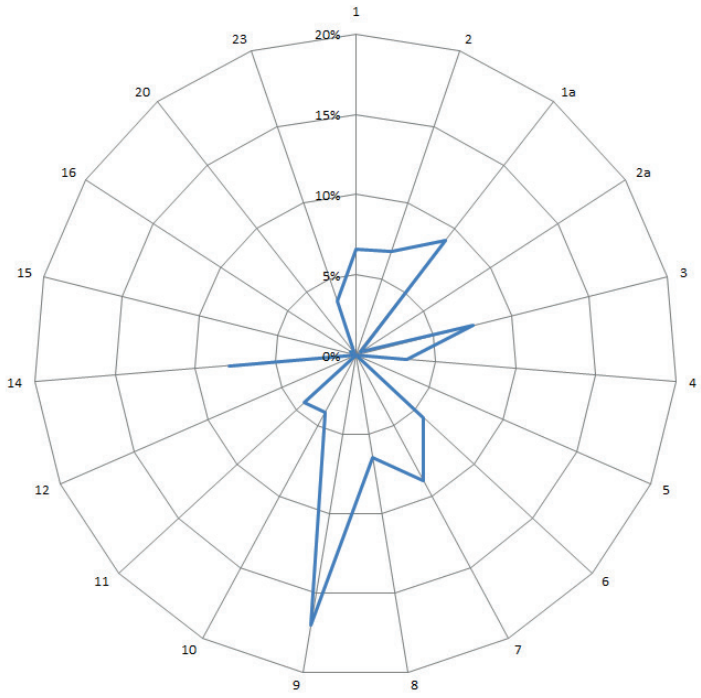

Figure 3: Noun type distribution (expressed in %) in the top-section of a Lusoga corpus

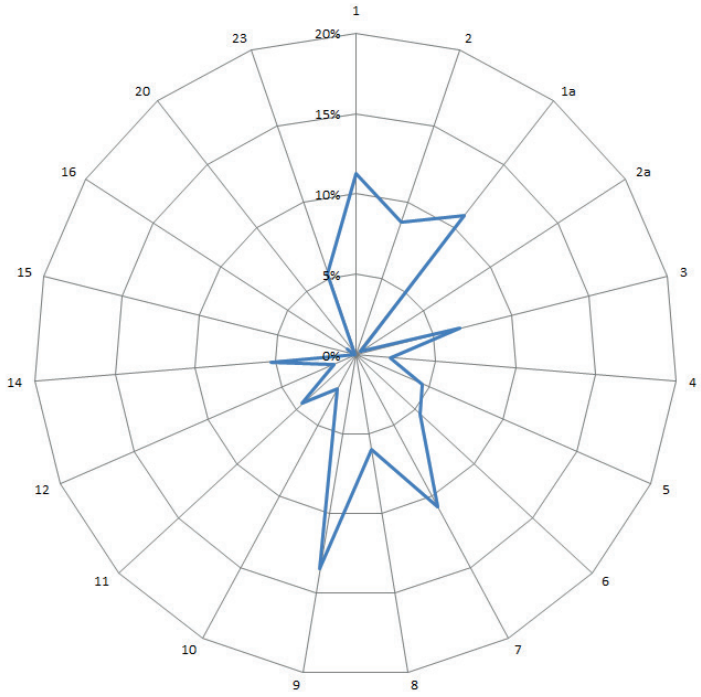

Figure 4: Noun token distribution (expressed in %) in the top-section of a Lusoga corpus

Table 4: Distribution of the genders (in terms of types) in the top-section of the Northern Sotho corpus used

Noun gender	*N*	%
1/2	167	95
1/-	7	4
1/6	2	1
1/**2** *pl.*	105	99
-/**2** *pl.*	1	1
1a/2b	40	95
1a/-	2	5
1a/**2b** *pl.*	9	100
3/4	161	89
3/-	19	10
3/10	2	1
3/**4** *pl.*	104	95
-/**4** *pl.*	6	5
5/6	222	88
5/-	27	11
5/10	2	1
5/**6** *pl.*	167	58
6	77	27
14/**6** *pl.*	36	12
9/**6** *pl.*	6	2
1/**6** *pl.*	2	1
7/8	160	76
7/-	50	24
7/**8** *pl.*	102	96
-/**8** *pl.*	4	4
9/10	557	86
9/-	88	13
9/6	5	1
9/**10** *pl.*	235	96
-/**10** *pl.*	8	3
5/**10** *pl.*	2	1
14	125	77
14/6	37	23
16	4	100
17	3	100
18	5	100
N-	13	100
24	2	100
SUM	2 564	

N- = homorganic nasal; *pl.* = plural

(1) Corpus lines for gender 1/6

Ba duma go ba le ngwetši ya	*Moafrika*	*ka gobane ke Maafrika.*
ye e fetilego. Kgopolo ye e thekgwa ke	*Maafrika*	*ka moka. Mabapi le dinyepo bjalo ka*
“Re ka thabela go bona morwa wa rena,	*morena*	*Matseba”. Ao ke mantšu a morena*
le tše di sa bonwego; ditulo tša	*marena*	*le bogoši bjohle, le mebušo le ba*

(2) Corpus lines for gender 3/10

Diponagalo tša tlhago bjalo ka letšatši,	*ngwedi,*	*dinaledi, legadima le mollo le tšona e ba*
tša bolepi. E ekiša dipolanete tše botse	*dingwedi,*	*le dihlopha tše dingwe tše dinyenyane tše*
Bula faele ye ngwe ya kheisi ya	*ngwaga*	*wo mongwe le wo mongwe wa ditšhelete*
khwalithi ya maphelo a bona. Ka	*dingwaga*	*tša bo1980, batho ba ganane tshepetšo*

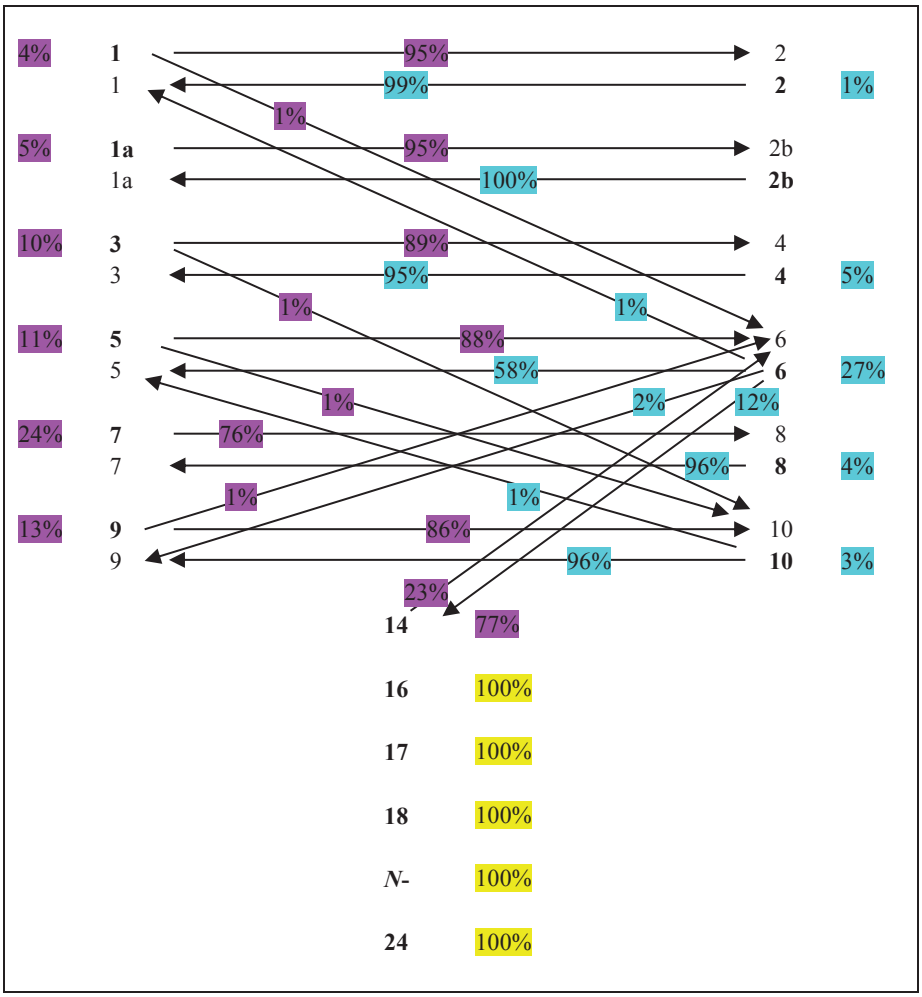

Figure 5: The Northern Sotho gender system visualised and quantified

The existence of a gender 3/10 can be explained in terms of a reinterpretation process according to which speakers intuitively reassign nouns to other noun classes. The morphophonology underlying these forms is for some reason not accessible to speakers, leading to an analogous reinterpretation of these nouns to new classes, thus generating new genders. The nouns *ngwedi* 'moon' and *ngwaga* 'year' both belong to class 3, the class prefix *ngw-* being an allomorph of the prefix *mo-* in cases where this prefix is affixed to a vowel-initial noun stem, thus **mo-edi* > *ngwedi*, and **mo-aga* > *ngwaga*. Plural formation in these cases follows an additive strategy according to which the plural prefix *me-* is affixed to the full noun stem, i.e. with retention of the singular prefix, resulting in the forms *mengwedi* (< me-mo-edi) and *mengwaga* (< *me-mo-aga*). However, as attested by the corpus data, the alternative plural forms *dingwedi* and *dingwaga* are also found. The surface realisation *ng(w)-* of the class prefix *mo-* is seemingly reinterpreted by speakers as being that of class 9, i.e. *N-*. Consequently, the plural formation strategy, i.e. the addition of the class 10 plural prefix *di-* which normally applies to class 9 nouns is utilised in these cases as well, resulting in a gender 3/10.

The information provided in Figure 5 can now be explained as follows, using gender 3/4 as an example: for 89% of nouns in class 3, a corresponding plural in class 4 is attested, while 95% of plural

nouns in class 4 have a singular form in class 3. Those without corresponding forms are found only in class 3 (in 10% of the occurrences of class 3 nouns) and class 4 (in 5% of the occurrences of class 4 nouns), thus constituting single-class genders. In addition, there is thus also a gender 3/10, which takes care of the last percentage of nouns that occur in class 3 (and here their plural is thus in class 10). Note that such an approach makes provision for the notion of bi-directionality, not only quantifying the relationship between singular and plural forms, but also between plural and singular forms. Viewed from a different angle, and using class 6 as an example, it can be stated that for any noun in class 6, there is a 58% chance that it belongs to gender 5/6 (*le-/ma-*), a 27% chance that it belongs to the single-class gender 6 (*ma-*), a 12% chance that it belongs to gender 14/6 (*bo-/ma-*), a 2% chance that it belongs to gender 9/6 (*N-*, *Ø-*/*ma-*), and finally a 1% chance that it belongs to gender 1/6 (*mo-/ma-*). Conversely, and with class 6 at the receiving end, it is the plural of class 5 nouns for 88% of the class 5 nouns, the plural of class 14 nouns for 23% of the class 14 nouns, the plural of class 1 nouns for 1% of the class 1 nouns, and the plural of class 9 nouns for 1% of the class 9 nouns.

One could rightly ask what the significance of this information could be. For language learners, these statistics could prove useful when confronted with an unknown noun in, for example, class 6. Without statistical guidance, learners have no real basis on which they can rely to identify the singular form of such a noun. This would have direct consequences for practical language skills, such as dictionary use, since Northern Sotho dictionaries generally lemmatise only singular forms.[2] Also, in natural language processing where machine-learning tools are used in, for instance, noun guessing, these statistics could assist with the selection of probable singular forms for plural nouns, and vice versa.

The existence of single-class genders—apart from the more obvious class 6 *pluralia tantum*, the class 14 abstracts, and the locatives—have only implicitly been part of the knowledge base of the grammar of Northern Sotho. It is argued that it is useful to also formalise this aspect. Among others, this would help to dispel the notion that all noun classes come in regular singular-plural pairs, with the exception of the cases mentioned above. Formalising single-class genders is furthermore valuable from a language learning perspective. From a semantic point of view, there often is no reason why a particular plural form can, for example, not have a corresponding singular form, and vice versa. Since the singular/plural pairing is morphologically largely regular, a language learner cannot be faulted for assuming that the plural form of, say, *selemo* ‘summer’ is **dilemo* ‘summers’, or that the singular form for *baratani* ‘lovers’ is **moratani* ‘lover’. It is therefore important to enumerate the nouns, or at least the most frequent ones, that belong to these single-class genders. Consequently, the most frequent ones, as extracted from the corpus, are listed below.

For this task, single-class genders were identified by searching the top-frequent section of the corpus for possible corresponding singular ↔ plural forms; if none were found, it was assumed that the noun studied belongs to a single-class gender. To illustrate: based on its morphological features (class prefix *mo-* and agreement morphemes as revealed in the corpus), the noun *mosegare* ‘noon, midday’ is classified as a class 3 noun. A corpus search for the corresponding plural form **mesegare* returns no results, therefore it can be assumed that *mosegare* belongs to a single class gender 3/-. In the case of nouns which are morphologically plural, the same procedure was followed in order to establish whether a corresponding singular could be found. It can, however, not be assumed that if a corresponding member is not found in the corpus, it does not exist, but the implication would be that the frequency of such an item would be extremely low (i.e. with an occurrence of less than once in seven million running words).

(3) Gender 1/- (*mo-*/-)

Mma.	‘Mrs, Ms’	*mongake*	‘traditional healer’
Mna.	‘Mr’	*Morena*	‘Lord, God’
Modimo	‘Lord, God’	*mothaka*	‘fellow’
mogatšaka	‘spouse’		

Note that although the plural forms *badimo* ‘ancestral spirits’ and *barena, marena* ‘gentlemen’ do occur, these cannot be regarded as the plurals of the forms *Modimo* and *Morena* respectively, mainly because extra-linguistic factors cause a difference in meaning.

(4) Gender -/2 (*-/ba-*)

baratani	‘lovers’

The noun *baratani* ‘lovers’ is a deverbative noun, derived from the verb root *rat-* ‘like, love’ to which a reciprocal suffix *-an-* had been added. The semantic implication of reciprocality logically presupposes plurality, which would explain the non-occurrence of a singular equivalent for *baratani* ‘lovers’.

(5) Gender 3/- (*mo-/-*)

mekhuri	‘mercury’	*monola*	‘humidity’
mmalwa	‘several, many, much’	*mosa*	‘pity, mercy’
mmalwanyana	‘few, little’	*moše*	‘other side’
modirišohlaodi	‘situative mood’	*mosegare*	‘noon, midday’
modirišopego	‘indicative mood’	*moseo*	‘inside area opposite a door/entrance’
modirišotaelo	‘imperative mood’		
moelamoya	‘air flow’	*mošola*	‘other side’
moko	‘(bone) marrow’	*Moya o Mokgethwa*	‘Holy Spirit’
momagano	‘coalescence’	*ngwagola*	‘last year’
mona	‘jealousy’	*ngwedi*	‘moon’

The nouns *moše* and *mošola* ‘other side’ are usually categorised as belonging to the locative class 18, but corpus evidence indicates they are (also) in gender 3/-. This will be expounded on further below in the section on the dynamism of the Northern Sotho noun class system.

(6) Gender -/4 (*-/me-*)

mebalabala	‘different colours’	*mehutahuta*	‘different kinds’
megabaru	‘greed’	*meokgo*	‘tears’
mehlamu	‘small talk’	*mešogofela*	‘always, by all means’

The reduplication of the noun stem in *mebalabala* and *mehutahuta* logically precludes the possibility of a corresponding singular form, as such reduplication implies plurality.

(7) Gender 5/- (*le-/-*)

leago	‘neighbourliness’	*lenyora*	‘thirst’
lebelwana	‘low speed’	*lephera*	‘leprosy’
lebese	‘fresh milk’	*lesedi*	‘(ray of) light’
lebjale	‘present tense’	*lesomenne*	‘fourteen’
leboa	‘north’	*lesomepedi*	‘twelve’
Lebowa	‘Lebowa’ (place name)	*lesomesenyane*	‘nineteen’
lefaufau	‘atmosphere’	*lesomeseswai*	‘eighteen’
lefetile	‘past tense’	*lesomešupa*	‘seventeen’
lehlabula	‘autumn’	*lesometee*	‘eleven’
lehlwa	‘snow’	*lesometharo*	‘thirteen’
lehono	‘today’	*lesometlhano*	‘fifteen’
lenyaga	‘this year’	*lesometshela*	‘sixteen’
lenyatšo	‘contempt’	*letago*	‘brightness’
		letl.	‘p.’

Immediately apparent is the presence of nine numerals. Although forms which are morphologically speaking the plural counterparts of these numerals do exist, it needs to be pointed out that there is a distinct semantic difference. From the perspective of a language learner it is important to note that the form *masomenne*, which is the morphological plural of *lesomenne* ‘fourteen’, has a different meaning, i.e. ‘forty’ and not, as could be expected ‘fourteens’. The same principle applies to all of the other numerals.

(8) Gender 6 (*ma*-)

maatla	‘strength’	*madira*	‘troops’
mabapi	‘concerning’	*mahlatse*	‘good luck’
madi	‘blood’	*mantšiboa*	‘evening’
mathomo	‘beginning’	*manyami*	‘pity’
meetse	‘water’	*maswi*	‘milk’
maabane	‘yesterday’	*matseno*	‘introduction’
madimabe	‘misfortune’		

A total of 77 nouns in the top-section of the corpus belong to this gender. Since this is the only single-class gender that has received substantial treatment in Northern Sotho grammars, with Poulos and Louwrens (1994: 25–26) covering a total of 13 instances, only the top-frequent ones are listed in (8).

(9) Gender 7/- (*se-/-*)

segagešo	‘our language/culture’	*seedi*	‘light’
selemo	‘summer’	*seruthwane*	‘spring’
seswai	‘eight’	*senyane*	‘nine’
setu	‘silence’		

For this single-class gender, 50 nouns have been identified as members. This gender typically contains names of languages and cultures, cf. *Sekgowa* ‘English’, *Sepedi* ‘Northern Sotho’, *Sesotho* ‘Sotho’, etc., which do not distinguish corresponding plural forms. Other examples are listed in (9).

(10) Gender -/8 (*-/di-*)

difokeng	‘at/in the royal palace’	*diphatlalatši*	‘media’
diketekete	‘several thousands’	*ditšhabatšhaba*	‘different nations’

As was the case with gender -/4, the reduplication of the stems in the listed examples excludes the possibility of a corresponding singular form.

(11) Gender 9/- (*N-/-*)

kadijela	‘veteran’	*ngangego*	‘tension’
kgethologanyo	‘discrimination’	*palomoka*	‘total’
kgole	‘far (away)’	*petrolo*	‘petrol’
koporo	‘copper’	*phološo*	‘salvation’
korong	‘wheat’	*tekano*	‘limit’
kwelobohloko	‘sympathy’	*tirišano*	‘cooperation’
nepagalo	‘precision’	*tshenyo*	‘damage’

A total of 88 of the top-frequency nouns belong to this gender. Nouns denoting localities such as the names of countries, provinces, cities and towns, the names of the months and other personal names are found in this gender. Other examples include the ones listed in (11).

(12) Gender -/10 (*-/diN-*)

dikgadima	‘thunderstorms’	*dipalopalo*	‘statistics’
dingwalo	‘literature’	*ditlalemeso*	‘morning news’
dinose	‘honey’	*ditlamorago*	‘consequences’
dipalo	‘mathematics’	*ditšiebadimo*	‘nonsense’

One may rightfully ask how one can assign these nouns to class 10 since there is no corresponding singular. In other words, one could argue that there is actually no way of knowing whether these are

class 8 or class 10 nouns, given that class 8 and class 10 have identical noun prefixes and trigger identical concords. However, in the case of some of the deverbatives, such as *dikgadima*, *dipalo* and *dipalopalo*, the influence of the underlying nasal prefix of class 9 is still evident. In other cases, such as *ditšiebadimo*, the first part of the compound noun is a class 9 noun, which implies that the plural can only be in class 10.

(13) Gender 14 (*bo*-)

Although Northern Sotho grammars all mention the fact that many nouns in class 14 do not have a corresponding plural form due to their abstract semantic nature, such cases are not regarded as belonging to a gender which is separate from the gender 14/6. In the top-section of the corpus, 125 nouns belong to this single-class gender. These do indeed include mostly abstract nouns.

Northern Sotho noun classes as a dynamic system

As mentioned above, in standard Northern Sotho grammars, the noun class system is usually presented as a static system in which the different noun classes represent watertight categories. The notion that nouns can be and are being reassigned to different classes is rarely mentioned. This being said, the diachronic process during which Bantu speakers reinterpreted the noun universe resulting inter alia in the establishment of a class solely reserved for nouns with the feature [+HUMAN] is well-described in the literature (cf. Givón 1971; Mould 1971; Louwrens 2000). According to this hypothesis, at some stage speakers began to reinterpret the noun universe, a process which resulted in the noun class system changing from a non-hierarchised, multi-gender system to an anthropocentric one, in which nouns with the feature [+HUMAN] appeared at the top of the hierarchical tree. In the initial system, nouns were categorised according to their semantic features, which implies that a one-to-one correlation existed between a noun's semantic features and its morphological features, viz. its class prefix. In this non-anthropocentric system, nouns with the feature [+HUMAN] were categorised together with nouns referring to animals in one class, based on the shared semantic feature [+animate]. The reinterpretation of human nouns led to the dismantling of the correlation between form and meaning, which resulted in nouns starting 'to "migrate" all over the noun class system giving rise to the present day chaotic and largely language specific situation' (Givón 1971: 41). However, this process is implicitly presented as a historical one which is now fully completed and no attention is given to the possibility that reinterpretation of the noun classes is an ongoing and dynamic process. An important difference between the historical process described above and the reinterpretation process which can currently be observed is the underlying motivation. Whereas the historical reinterpretation was seemingly triggered by a change in the cognition of the noun universe and the position of humans within the universe, leading to the emergence of an anthropocentric cosmology, the current process seems to be largely driven by morphophonological considerations, which are illustrated below.

Corpus evidence indicates that a number of nouns have double class membership (i.e. nouns with the exact same meaning which are found in different noun classes), suggesting that these nouns are being reinterpreted as belonging to two different noun classes and that the double class membership is an indication of an incomplete reinterpretation process. Noun classes which are particularly affected by this reinterpretation process are the locative classes, specifically class 18 and the so-called *N*-locative class. It would seem that a number of nouns in these two classes are being reinterpreted as belonging to classes 3 and 9 respectively, the reason being the morphological similarity between the class prefixes of classes 18 and 3 on the one hand, and *N*- and 9 on the other. Locative classes in Northern Sotho are unproductive, i.e. the class prefixes can no longer be affixed to nouns or noun stems to impart a locative meaning to non-locative nouns, as is the case for other Bantu languages, such as Chichewa (Bresnan & Mchombo 1989) and Cilubà (Kabuta 1998). These classes have furthermore been subjected to semantic bleaching in that specific locative meanings can no longer be attached to specific locative classes, and a reduction in agreement morphemes has also taken place—in Northern Sotho locative classes mainly use the agreement morphemes of class 17. It would therefore seem that nouns belonging to the locative classes are good candidates for reassignment to

other noun classes, possibly eventually resulting in the demise of these classes in Northern Sotho—a process similar to what has happened to classes 12 and 13, the so-called diminutive classes. Strong evidence of such a process of reinterpretation is presented in the corpus data illustrated below.

Class membership is morphologically signalled or marked by means of a class prefix, and in cases where a prefix is orthographically similar for two or more noun classes, e.g. classes 1, 3 and 18 which all have the class prefix *mo-*, identification of the noun class to which a particular noun belongs depends on the agreement morphemes generated by the particular noun. It needs to be pointed out that due to the process of semantic bleaching, all locative classes make use of the agreement morphemes of class 17 (cf. Louwrens 1991: 116). Generating class 17 agreement therefore serves as confirmation of membership of any of the locative classes; specification of the particular locative class can then be done based on the nominal prefix displayed by the noun. A corpus search for three nouns *moše* 'other side', *mošola* 'other side' and *mošono* 'this side' which are traditionally classified as belonging to class 18 (cf. Lombard et al. 1985: 50) is carried out in order to determine their class membership, as revealed by the agreement morphemes which they generate when used in authentic texts. A concordance search for *moše* 'other side' results in 134 hits. In 68 of the corpus lines, the noun *moše* generates agreement, and in 20 of these lines the agreement morphemes are those of class 17, thus confirming membership of class 18, in the remaining 48 lines the agreement morphemes are those of class 3. For *mošola* 'other side' the total number of hits is 203, of which 132 have evidence of agreement. In only 8 of these, the agreement morphemes are those of the locative classes (here, class 18), the rest (124) are those of class 3. A search for *mošono* 'this side' throws up 61 hits, of which 34 contain evidence of agreement, all of which are the agreement morphemes of class 3; no examples are found in which this noun generates the agreement morphemes of the locative classes. Figure 6 summarises these data. Compare also the examples in (14) to (16), culled from the corpus, where the lines presented in (a) illustrate class membership in class 3, while the lines in (b) illustrate class membership in class 18, albeit with the agreement morphemes of class 17.

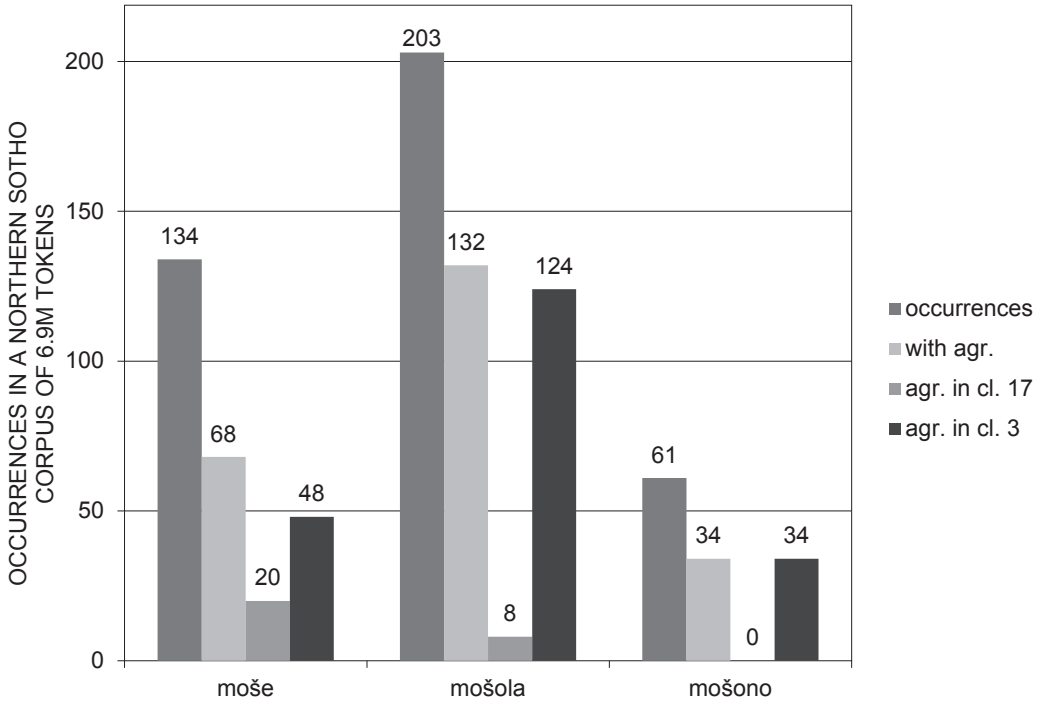

Figure 6: Double class membership in Northern Sotho

(14a) *Go be go direga gantši gore ba moše* ***wola wa*** *noka … ba tshelele moše o mongwe*.
'It often happened that the ones from that side of the river…had to cross to the other side.'
(*wola* = demonstrative class 3, *wa* = possessive concord class 3)
(14b) *Tša ka moše* ***ga*** *lebitla ga re di tsebe*.
'What is on the other side of the grave, we don't know.'
(*ga* = possessive concord class 17)

(15a) *Nelson Mandela o mo file monyetla wa go ithuta mošola* ***wa*** *mawatle*.
'Nelson Mandela gave her the opportunity to study overseas (lit. on the other side of the sea).'
(*wa* = possessive concord class 3)
(15b) *Mo Tubatse, ka mošola* ***ga*** *noka ya Tubatse ge o etšwa ka thoko yela ya gaMahlakwena, go na le sekolo se se phagamego.*
'Here in Tubatse, on the other side of the Tubatse river, if you come from that direction of gaMahlakwena, there is a high school.'
(*ga* = possessive concord class 17)

(16a) *Yena ke monna yo a tšwago ka mošono* ***wa*** *Udi*.
'He is a man who comes from this side of the Udi river.'
(*wa* = possessive concord class 3)

The same phenomenon is observed for some nouns belonging to the so-called *N*- locative class. Compare the examples in (17) and (18), in which the two nouns *kgole* 'far' and *kgauswi*/*kgaufsi* 'close (by), near' display agreement morphemes of class 17 (here, the *N*- locative class) and class 9, in (a) and (b) respectively.

(17a) *Sefolo, o kgole* ***ga*** *kgaufsi*.
'Sefolo, you are far from (being) near/close.'
(*ga* = possessive concord class 17)
(17b) *Kgauswi ya ba kgole, kgole* ***ya*** *ba kguaswi.*
'Near became far, far became near.'
(*ya* = possessive concord class 9)

(18a) *E topa maswikana a mararo ya a bea kgauswi* ***ga*** *pitša*.
'She picked up three pebbles and put them close to the pot.'
(*ga* = possessive concord class 17)
(18b) *Kgauswi* ***ya*** *batho ba bangwe ke kgole ya batho ba bangwe*.
'Near for some people is far for other people.'
(*ya* = possessive concord class 9)

In view of the corpus data presented above it can therefore be concluded that the noun class system of Northern Sotho is indeed a dynamic one, subject to reassignment and reinterpretation of nouns. This reinterpretation process is mostly morphophonologically driven, but the fact that the noun class system is not (or rather no longer) a semantic one is probably a contributing factor.

Conclusion

It is a truism in corpus linguistics that access to large amounts of data reveals insights which surpass those made possible through introspective analysis. This study has once more provided evidence in this regard, even though Northern Sotho belongs to the better described Bantu languages. In some cases, an investigation of corpus data even provides the researcher with results which are counter-intuitive. The novel radar diagrams which represent the frequency and distribution of noun classes in Northern Sotho, and which were compared to those for Lusoga, are a case in point; the Northern Sotho data seemingly presenting evidence which is contrary not only to speakers' intuition, but which

also seems to be in contrast with the hypothesis regarding topicality in language and the position in which [+HUMAN] nouns appear in this hierarchy. It was also indicated that the noun gender system of Northern Sotho is not a one-directional, singular-plural one, but a two-directional one for which the frequency of occurrence of each gender 'direction' may be weighted or quantified, and which also includes several single-class genders. Evidence was also provided to the effect that the noun class system of Northern Sotho, and by extension the gender system, is non-static and that reinterpretation and reassignment of nouns to different noun classes is an ongoing process. It is argued that the results of this corpus-driven investigation have implications for language teaching, computational linguistics, better informed use of reference works such as dictionaries, and also last but not least for purely fundamental linguistic research.

Acknowledgements — The research by the first author was funded by the Special Research Fund of Ghent University.

Notes

1. Senses of noun types were taken into account, which is why the total number of noun types exhibiting the feature [+HUMAN] is not a round figure.
2. The *Oxford Bilingual School Dictionary: Northern Sotho and English* (de Schryver 2007) being a notable exception.

References

Bostoen K, de Schryver G-M. 2015. Linguistic innovation, political centralization and economic integration in the Kongo Kingdom: Reconstructing the spread of prefix reduction. *Diachronica* 32(2): 139–185 (plus 13 pages of supplementary material online).

Bostoen K, Mberamihigo F, de Schryver G-M. 2012. Grammaticalization and subjectification in the semantic domain of possibility in Kirundi (Bantu, JD62). *Africana Linguistica* 18: 5–40.

Bresnan J, Mchombo SA. 1989. *On the syntax of Bantu noun class prefixes* (manuscript). Berkeley: Stanford University and University of California.

de Schryver G-M. 1999. *Cilubà phonetics, proposals for a 'corpus-based phonetics from below'-approach*. Ghent: Recall.

de Schryver G-M. 2007. *Oxford Bilingual School Dictionary: Northern Sotho and English/Pukuntšu Ya Polelopedi Ya Sekolo: Sesotho Sa Leboa Le Seisimane. E Gatišitšwe Ke Oxford*. Cape Town: Oxford University Press Southern Africa.

de Schryver G-M, Gauton R. 2002. The Zulu locative prefix ku- revisited: A corpus-based approach. *Southern African Linguistics and Applied Language Studies* 20(4): 201–220.

de Schryver G-M, Nabirye M. 2010. A quantitative analysis of the morphology, morphophonology and semantic import of the Lusoga noun. *Africana Linguistica* 16: 97–153.

de Schryver G-M, Taljard E. 2006. Locative trigrams in Northern Sotho, preceded by analyses of formative bigrams. *Linguistics, An Interdisciplinary Journal of the Language Sciences* 44(1): 135–193.

Dom S, Segerer G, Bostoen K. 2015. Antipassive/associative polysemy in Cilubà (Bantu, L31a): A plurality of relations analysis. *Studies in Language* 39(2): 354–385.

Gauton R. 2000. Locative noun classes in Bantu: The case for recognizing two additional locative noun class prefixes. In: Wolff HE, Gensler OD (eds), *Proceedings of the 2nd World Congress of African Linguistics, Leipzig 1997.* Cologne: Rüdiger Köppe. pp. 525–542.

Gauton R, de Schryver G-M, Mohlala L. 2004. A corpus-based investigation of the Zulu nominal suffix *-kazi*: A preliminary study. In: Akinlabi A, Adesola O (eds), *Proceedings of the 4th World Congress of African Linguistics, New Brunswick 2003.* Cologne: Rüdiger Köppe Verlag. pp. 373–380.

Givón T. 1971. Some historical changes in the noun-class system of Bantu, their possible causes and wider implications. In: Kim C-W, Stahlke H (eds), *Papers in African Linguistics.* Edmonton: Linguistic Research Inc. pp. 33–54.

Givón T. 1976. Topic, pronoun, and grammatical agreement. In: Li CN (ed.), *Subject and Topic.* New York: Academic Press. pp. 149–188.

Hawkinson AK, Hyman LM. 1974. Hierarchies of natural topic in Shona. *Studies in African Linguistics* 5(2): 147–170.

Kabuta NS. 1998. Loanwords in Cilubà. *Lexikos* 8: 37–64.

Kawalya D, Bostoen K, de Schryver G-M. 2014. Diachronic semantics of the modal verb -sóból- in Luganda: A corpus-driven approach. *International Journal of Corpus Linguistics* 19(1): 60–93.

Lombard DP, van Wyk EB, Mokgokong PC (eds). 1985. *Introduction to the Grammar of Northern Sotho*. Pretoria: J.L. van Schaik.

Louwrens LJ. 1991. *Aspects of Northern Sotho Grammar.* Pretoria: Via Afrika.

Louwrens LJ. 1994. *Dictionary of Northern Sotho Grammatical Terms*. Pretoria: Via Afrika.

Louwrens LJ. 2000. Anthropocentrism, utilitarianism and supernaturalism in African world view: Some linguistic evidence. *South African Journal of Ethnology* 23(2–3): 91–101.

Morolong M, Hyman LM. 1977. Animacy, objects and clitics in Sesotho. *Studies in African Linguistics* 8(3): 199–218.

Mould MJ. 1971. The agreement of nominal predicates in Luganda. *Studies in African Linguistics* 2(1): 25–36.

Poulos G, Louwrens LJ. 1994. *A linguistic analysis of Northern Sotho*. Pretoria: Via Afrika.

Taljard E. 2006. Corpus-based linguistic investigation for the South African Bantu languages: A Northern Sotho case study. *South African Journal of African Languages* 26(4): 165–183.

Tognini-Bonelli E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Toscano M, Sewangi S. 2005. Discovering usage patterns for the Swahili amba- relative forms Cl. 16, 17, 18: Using corpus data to support autonomous learning of Kiswahili by Italian speakers. *Nordic Journal of African Studies* 14(3): 274–317.

Van Wyk EB, Groenewald PS, Prinsloo DJ, Kock JHM, Taljard E. 1992. *Northern Sotho for first-years*. Pretoria: J.L. van Schaik.

Ziervogel D, Lombard DP, Mokgokong PC. 1969. *A Handbook of the Northern Sotho Language*. Pretoria: J.L. van Schaik.

Ziervogel D, Mokgokong PC. 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 & UNISA.