A quantitative analysis of the morphology, morphophonology and semantic import of the Lusoga noun

Gilles-Maurice de Schryver
and Minah Nabirye

Abstract

In this article it is shown how distributional corpus analysis may be used to start the description of a (mostly) undocumented language. The approach is illustrated for Lusoga (JE16), an eastern interlacustrine Bantu language spoken in and around Jinja, Uganda. The topic is the noun in Lusoga, with three levels receiving particular attention: the morphological, morphophonological and semantic.

In a first section we show that a relative distribution of the type and token counts for each noun class in combination with a weighted two-dimensional noun class system is a most powerful way to visualize the strength of each node and each link in the structure. In a second section we proceed with an indication of how a quantified enumeration of both nominal morphophonology and noun constructions cum linked meanings provides for a representative picture of the various noun-building issues. In a third and final section, we then argue in favour of a three-dimensional semantic-import view of nouns, with as axes noun classes, semantic categories, and corpus frequencies.[1] This is not only a novel but also a most revealing and promising avenue to decode the underlying semantic system of the noun in Lusoga, as well as the noun in any other Bantu language.

Keywords: Lusoga, Bantu, noun class system, corpus linguistics, semantics

---

1. As far as the expression 'semantic import' is concerned, we use import in its historical first use, according to the *Oxford English Dictionary*: "The fact of importing or signifying something; that which a thing (esp. a document, phrase, word, etc.) involves, implies, betokens, or indicates; purport, significance, meaning." (*OED* - import *n*., I. 1), as attested in Shakespeare's "There's letters from my mother: What th' import is, I know not yet." (*All's Well, That Ends Well* - 1601, II. iii. 294).

1. Bantu corpus linguistics

According to Himmelmann (1998, 2006), the main methods of data collection in field-based documentary linguistics are (a) observed communicative events, (b) staged communicative events, and (c) elicitations. As Lüpke (2009:55) points out, "field-based corpora often constitute first documentations", and as such a combination and cross-comparison of the results of methods (a), (b) and (c) is typically required in order to arrive at an adequate description of the language being documented. Lüpke is fully aware of some of the problems with each of these methods in isolation. With regard to the stimuli used in method (b), for example, she points out that "they do not allow a data-driven perspective on the 'genius' of a particular language" (p. 69), and adds that they "yield data that are phonologically, morphologically and syntactically naturalistic, but may present semantic oddities when culturally odd, inappropriate or unusual scenes are depicted" (p. 70). With regard to method (c), she writes: "Elicited data have very low ecological validity – they come into existence under the control of the researcher and are entirely motivated by their research questions" (p. 88). For similar concerns, see Dimmendaal (2001), Mc Laughlin & Sall (2001), or Mithun (2001). The main underlying problem, of course, is that the text corpora which are the result of the transcriptions made of the speech data from method (a) are generally too small. Balancing out methods (a), (b) and (c), as Lüpke (2005a) did in her own PhD on Jalonke (spoken in Guinea), generally results in solid grammatical descriptions. Interestingly, in a subsequent paper Lüpke (2005b) shows how, for statistical analyses, she would still limit herself to a sub-corpus from which the staged communicate events and elicitations have been severed.

To an increasing number of researchers in the language sciences the power of natural language is too compelling indeed, and for major languages this has given rise to the field of corpus linguistics, of which Sinclair (1966) was one of the pioneers. Crucial for corpus linguistics is to have a fair amount of textual data – a large electronic corpus – at one's disposal. For languages of limited diffusion (LLDs, be those minor, minority or endangered languages) this is typically the bottleneck. Transcribing naturally-occurring speech is known to be both time-consuming and costly. However, for more and more LLDs, written material is becoming available (see e.g. Scannell 2007), and for those languages the prospect of applying techniques from the field of corpus linguistics come into view. This prospect has now become a reality for a good number of Bantu languages.

The present article joins a growing body of corpus-based grammatical studies for the Bantu languages. Examples of earlier studies include: a corpus take on the phonetics of Cilubà (L31a), by De Schryver (1999); the first corpus-based diachronic analysis of a linguistics aspect of a Bantu language, *in casu* the locative prefix **ku-** in Zulu (S42), by De Schryver & Gauton (2002); an examination of the intrinsic and contextual semantic import of the Zulu nominal suffix **-kazi**, by Gauton *et al*. (2004); a minute description of the structures of the higher-order locative *n*-grams in Northern Sotho (S32), by De Schryver & Taljard (2006); and a semantic study illustrating the historical relationship between adjectives and enumeratives in Northern Sotho, by Taljard (2006).

What characterizes each of those undertakings is that they uncovered hitherto unknown aspects of the Bantu languages under study. In this sense the present undertaking is of a different magnitude, as the end goal is to write the first learners' grammar for a Bantu language that is entirely sourced from an electronic corpus. The language analysed is Lusoga (JE16), a mostly undocumented language spoken by about two million Basoga in eastern Uganda (UBS 2006:44). This article, then, should be seen as the first in a series that reports on the outcomes as the project proceeds.

To the best of our knowledge, the only published reference grammar that is entirely corpus-based is one for English, namely the *Longman Grammar of Spoken and Written English* (Biber *et al*. 1999). On the one hand one could therefore conclude that the Lusoga grammar project is too daunting; on the other hand the aim is precisely to show that it is not only possible but also desirable to write modern grammars within a corpus-linguistics framework. For one, this allows the compilation of such grammars to be fast-tracked while, even more important, the resulting description is based on actual language usage.

This first report deals with the noun in Lusoga. More in particular, Lusoga nouns are subjected to an in-depth analysis on three levels: (a) morphological (i.e. a study and quantification of the form of the various noun classes, as well as their so-called singular-plural pairings, if any); (b) morphophonological (i.e. a study and quantification of the sound changes when attaching nominal morphemes to roots and stems, as well as a study of the origin of those roots and stems); and (c) semantic (i.e. a study and quantification of the contents of this word category, per noun class, and overall).

2. The Lusoga corpus

The starting point of any study in corpus linguistics is the building of a corpus of texts. Over the course of the past eight years, data was collected with a view to compile the first monolingual dictionary of Lusoga. That dictionary has recently been published (Nabirye 2009a), and given that all the example sentences are based on original fieldwork, *in casu* observed communicative events, we felt that they could form part of a Lusoga corpus. This material was complemented with scanned selections from newspapers, the New Testament and other religious texts, various reports, a series of short stories, as well as transcriptions of conversations, interviews and songs. The distribution of these components is shown in Table 1, together with the number of words – known as tokens – in each section.

Genre	Tokens	%
Dictionary (Eiwanika ly'Olusoga)	305,660	35.00
Newspapers (Kodh'eyo, Ndiwulira)	187,393	21.46
Religious texts (New Testament and others)	199,853	22.88
Reports (from the Busoga clan leaders, private sector, academia, etc.)	24,166	2.77
Short stories (Ababita Ababiri, Ensambo edh'Abasoga, etc.)	150,560	17.24
Transcriptions of conversations, interviews and songs	5,716	0.65
SUM	873,348	100.00

Table 1: Genre distribution in the Lusoga corpus.

As may be seen from Table 1, the Lusoga corpus contains about 870,000 running words (tokens). The transcriptions of conversations, interviews and songs, as well as the dictionary examples – together close to 36% – are reductions of spoken data to text, the other genres were text from the start. Important to observe at this point is that the various orthographies as seen in the original sources were left intact, which implies that the number of orthographically different words – known as types – is slightly inflated compared to a corpus in which the spelling would have been homogenized. As it stands, there are slightly over 150,000 different orthographic words (types) in the Lusoga corpus. Working with a corpus that contains various spellings for some of the same words is not really a hurdle; it only means that one is dealing with some (evenly spread) noise as far as the type counts are concerned; the token counts, however, are always exact. In this article, and for all morphophonological analyses, the spelling introduced in Nabirye (2008) is used. From Table 2 one may further deduce that most sources are recent to very recent, with over 98% produced during the past two decades.

Period	Tokens	%
1960s	16,822	1.93
1970s	–	–
1980s	–	–
1990s	457,978	52.44
2000s	398,548	45.63
SUM	873,348	100.00

Table 2: Period distribution in the Lusoga corpus.

This first version of the Lusoga corpus was not annotated for any linguistic features, as one of the goals of the current study is exactly to uncover those linguistic features. As such, the corpus was not tagged for parts of speech, nor lemmatized.

3. Distributional corpus analysis vs. cognitive semantics

In corpus linguistics one is typically interested in what is common and has predictive power, rather than in what is rare and are outliers. We therefore lifted out all the types in the corpus with a minimum frequency of ten, of which there are roughly 7,000. About one third of those – 2,263 types to be exact – turned out to be nouns. It is these 2,263 noun types, together with their contexts, which constitute the raw material for the study being reported on below. Although it is obviously impossible to make abstraction of received knowledge as far as Bantu grammar is concerned (nor would it be wise to do so), it is true that we took nothing for granted. In practical terms this meant that, for each and every noun candidate, a trained mother-tongue speaker analysed all the (sorted) concordance lines proffered by the corpus query software. It is only following the concurrent consideration, for each noun-type candidate, of (a) the form of the noun prefix, and (b) the form of the concordial agreement morphemes seen in the surrounding context, that nouns were assigned to certain classes. The figure of 2,263 noun types was thus only arrived at once this task was completed. One could therefore say that distributional corpus evidence pinpointed and/or confirmed noun class membership. Moreover, each noun class as a whole was studied and looked at in isolation, disregarding possible (and so-called) singular-plural pairings in a first phase (Section 4). In a second phase relations were uncovered – again following searches through the corpus – leading to noun genders (Section 5). This in turn led to a third phase, namely the pinpointing of the various ways in which nouns are built in Lusoga, together with a study of the applicable sound changes when attaching affixes and roots or stems to one another (Section 6). In addition to these morphological and morphophonological considerations, noun meanings, too, were studied in context (Section 7).

The concurrent analysis of noun class prefixes and concordial agreement morphemes, undertaken in order to assign noun types to classes and genders, does not imply that we subscribe to a mechanistic interpretation of alliterative concord, controlled by syntax. Since the publication of Contini-Morava's *'Things' in a Noun-Class Language* (1996) we know that concords may be "regarded as signals of meanings, not as meaningless or redundant formatives inserted by a 'rule of concord'" (p. 277). The agreement system not being mechanistic, one may actually interpret the system as a cross of lexical collocations and syntactic colligations – with, following Firth (1951 [1957]), collocation the co-occurrence of words, and colligation the co-occurrence of grammatical phenomena. With this one has arrived at "a distributionalist method for lexical semantics: examine the syntagmatic environments in which a word occurs, and you will know more about the kind of word you are dealing with" (Geeraerts 2010:165). Geeraerts (2009:422-3) proposes to view "distributional corpus analysis" of the Sinclair-type as a neostructuralist approach to lexical semantics, with as main characteristic the "radical usage-based rather than system-based approach: it considers the analysis of actual linguistic behaviour to be the ultimate methodological foundation of linguistics" (Geeraerts 2010:168). The present study of the noun in Lusoga, then, is carried out within the theoretical framework of distributional corpus analysis (DCA). As an approach to lexical semantics, one of the goals will therefore also be to say something about

word meaning, or, more specifically for Bantu, the semantic import of each of the various noun classes uncovered.

In a landmark paper Hendrikse & Poulos (1992) argued in favour of an "underlying cognitive organization of the noun universe" (p. 199) and proposed the following "word category continuum" (pp. 207-8) for nouns across the Bantu languages:

Nouns	→	Adjective-like nouns	→	Adverb-like nouns	→	Verb-like nouns
Concrete						**Abstract**
1/2, 3/4, 9/10	5/6, 7/8, 11	12/13, 19, 20, 21, 22		16, 17, 18, 23	14	15

Re-reading Hendrikse & Poulos's paper, one is surprised to see that they succeeded in building a strong argument without presenting a single example from a single Bantu language. It seems as if they took the reader in tow, assuming that that reader would not look too closely.

Others have looked at data, albeit pre-corpus-era dictionary data only. Selvik (2001), for example, in a polysemy analysis of three Tswana (S31) noun classes, used an existing dictionary as a 'fish pond': selecting from it what fits her model (schemas) and throwing back what does not. Apart from the fact that meanings in traditional dictionaries often do not correspond with the meanings that need to be mapped onto the true use as seen in large corpora, the main problem is that Selvik's approach is not random: she uses carefully chosen words as dominoes, creating "networks involving chains of meaning associations" (p. 181). A similar approach, also based on pre-corpus-era dictionary data, may be found in the early work of Contini-Morava (1994, 1997) on Swahili (G42), whereby each noun class prefix is seen as "a distinct linguistic sign, but rather than having a single, invariant meaning, its meaning consists of a network of senses connected to one another both by relations of taxonomic inclusion and by relations of semantic extension such as metaphor and metonymy" (Contini-Morava 2002:7). Even though in her later work Contini-Morava (2002) adds an "indices analysis" to the "polysemy analysis", her approach remains that of a cognitive semanticist, where one "start[s] from an encyclopaedist conception of meaning, in the sense that lexical meaning is not considered to be an autonomous phenomenon, but is rather inextricably bound up with the individual, cultural, social, historical experience of the language user" (Geeraerts 2002:31). This stands in sharp contrast to a neostructuralist approach such as DCA, in which one "trie[s] to demarcate a uniquely linguistic level of meaning" (Geeraerts 2009:424).

In studying the semantic import of the Lusoga noun, we will therefore not entertain any semantic networks consisting of chains of family resemblances, linking members based on common properties, or metaphor and metonymy, nor will we try to recognize prototypes. At the same time, our analysis will be more detailed than the abstract-concrete continuum recognized by Hendrikse & Poulos. Jump-starting some of the results of the Lusoga noun study presented in detail below, and collapsing the data along the lines of the classes/genders suggested by Hendrikse &

Poulos, the graph shown in Figure 1 is obtained. (Observe that the infinitive nouns are not included here, as those are part of a forthcoming study of the Lusoga verb.)

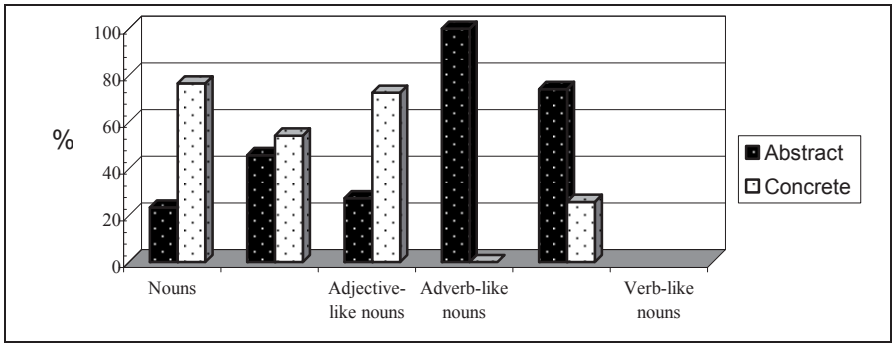

Figure 1: Abstract vs. concrete noun distribution in Lusoga, per group (in terms of types).

At face value, Figure 1 seems to roughly confirm Hendrikse & Poulos's statement, in that the degree of abstractness tends to increase moving through the continuum, with the degree of concreteness decreasing in parallel. Disregarding the fact that the progression is not truly linear, an obfuscating problem is that each group (e.g. Group 2: 5/6, 7/8, 11) is considered in isolation, set out in function of 100%. If one looks at the same data, but for each group now as a part of the total, Figure 2 is obtained.

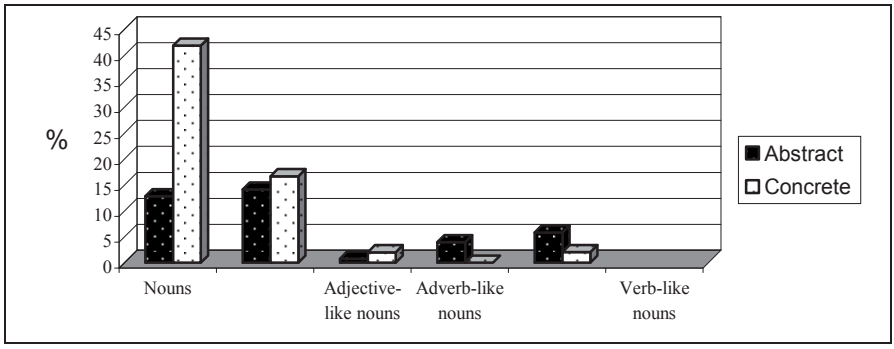

Figure 2: Abstract vs. concrete noun distribution in Lusoga, overall (in terms of types).

About 42% of all the nouns in Lusoga are concrete nouns found in Group 1, 17% in Group 2, and 2% in Group 3. In parallel, only 13% of all the nouns are abstract nouns in Group 1, 14% in Group 2, and under 1% in Group 3. For these first three groups, each of the abstract values is thus lower than the concrete ones. The reverse is only seen for Groups 4 and 5.

If anything, Figures 1 and 2 suggest that a more fine-grained approach to the semantic import of the various Bantu noun classes is required. Rather than a blunt distinction between concrete and abstract, we ended up distinguishing between up to ten semantic categories per noun class in our study. In deciding on those ten we were led by the corpus evidence, although, unsurprisingly, our cut-up cuts through

several of the existing semantic mappings found in the Bantu literature (cf. e.g. the summaries in Hendrikse & Poulos (1992:199-201) or Maho (1999:63-99)). No particular claims are made with regard to the definiteness of the ten categories chosen. Rather, the aim is to arrive at a proof of concept for a new way to look at the semantic import of the noun classes in Bantu languages, based on corpus evidence, and to illustrate this for Lusoga. In practical terms, one mother-tongue speaker assigned each of the 2,263 noun types to one or more semantic categories, taking the polysemous and homonymous uses as seen in the corpus into account. Not all uses of each noun type were recorded in the process; the focus was on all the frequent uses.

The overall process followed in our distributional corpus analysis of the Lusoga noun may therefore be summarized as follows:

1. extract all corpus types with a frequency of at least ten;
2. identify noun-type candidates, and for each candidate:
   a. call up corpus lines and concurrently study the form of the noun class prefix and the concordial agreement morphemes;
   b. confirm noun-type status and assign class number;
3. group noun types according to class number, and for each noun type within each class:
   a. search the corpus for possible corresponding (singular/plural) forms; and for each form (original and corresponding, if found):
      i. add one or more glosses (mapping meaning onto use);
      ii. note the morphophonological variation, if any;
   b. assign a one- or two-class gender;
   c. differentiate between inherent and derived noun types, and for the derived ones:
      i. indicate how the noun type is built up (i.e. constructed);
      ii. deduce the generic meaning of the construction (including a consideration of all noun types with identical constructions);
   d. label each with one or more semantic categories;
4. quantify all levels (itimized in step 3) in terms of types and tokens.

4. The Lusoga noun in the corpus

In the section of the corpus looked at – i.e. all nouns with a frequency of at least ten, together with their contexts – a total of 19 different noun classes were found. These are as shown in Table 3, together with their type and token counts.

Class	**1**	**2**	**1a**	**2a**	**3**	**4**	**5**	**6**	**7**
Types (N)	149	155	205	8	171	73	120	130	201
Type %	6.58	6.85	9.06	0.35	7.56	3.23	5.30	5.74	8.88
Tokens (Freq.)	12,633	9,812	12,295	436	7,472	2,406	5,111	6,073	12,072
Token %	11.27	8.75	10.97	0.39	6.67	2.15	4.56	5.42	10.77

Ctd.	8	9	10	11	12	14	15	16	20	23	SUM
	146	385	91	99	61	178	1	8	1	81	2,263
	6.45	17.01	4.02	4.37	2.70	7.87	0.04	0.35	0.04	3.58	100.00
	6,647	15,136	2,705	5,025	1,655	5,909	24	716	11	5,968	112,106
	5.93	13.50	2.41	4.48	1.48	5.27	0.02	0.64	0.01	5.32	100.00

Table 3: Noun distribution in the Lusoga corpus (in terms of types and tokens).

The 2,263 noun types correspond to 112,106 noun tokens. The largest noun class, both in terms of types and tokens, is class 9. (Observe that the type and token distributions correlate rather well; their Pearson correlation coefficient is 0.90.)

Each of these 19 noun classes will now be briefly discussed. The basic facts of the first 15 classes are summarized in three tables each, included as addenda – where N refers to a count of the noun types, Freq. to a count of the noun tokens. In line with a discovery procedure, where no prior assumptions are made, nouns with vs. without their pre-prefixes are counted separately.

4.1. Class 1 (149 types; 12,633 tokens)

Appendix 1.1 shows that 95% of the nouns in class 1 have a corresponding (plural) form in class 2 (e.g. **omulenzi** 'boy', **omuzaile** 'parent'); 5% are only attested in class 1 (e.g. **omumyuka** 'second in command, vice-', **Omulokozi** 'Saviour'). Also, there is only one form of the class 1 noun prefix: (**o**)**mu-**. Appendix 1.2 lists the sound changes that are applicable when this noun prefix is attached to the various roots and stems (the relevant sound changes for the corresponding (plural) form are also listed). All class 1 sound changes are straightforward semivocalizations. Predictably in Bantu, and as seen in Appendix 1.3, the semantic import of class 1 is overwhelmingly pointing to people; with the abstracts even debatable, as philosophical: **omusengwa** 'god'. Halves in the type column (N) are the result of the homonymous and/or polysemous nature of some nouns: **omusumba** 'pastor; god'. Top-frequent members of class 1 include: **omuntu** 'person', **omwana** 'child', **omusaadha** 'man', **omukazi** 'woman', and **omughala** 'girl'.

4.2. Class 2 (155 types; 9,812 tokens)

From Appendix 2.1 one sees that all nouns in class 2 have a corresponding (singular) form in class 1. The class 2 noun prefix is always: (**a**)**ba-**. The class 2 sound changes in Appendix 2.2 are straightforward vowel coalescences, with **a + e > e/_NC** the orthographic rule whereby a long vowel is written as one (but still pronounced long) when followed by a nasal+consonant, as in: **abembi** 'singers'. The semantic import of class 2 is similar to that of class 1, as may be deduced from Appendix 2.3. Top-frequent members of class 2 include: **abantu** 'people', **abaana** 'children', **abasaadha** 'men', **abakazi** 'women', and **abaghala** 'girls'.

4.3. Class 1a (205 types; 12,295 tokens)

About 18% of the nouns in class 1a have a corresponding (plural) form in class 2a; the other 82% are only attested in class 1a (e.g. **duuma** 'maize', **mwogo** 'cassava'). While class 1a nouns are characterized by a zero noun prefix: **ø-**; most class 2a nouns take **ba-** as (plural) prefix (e.g. **maama/bamaama** 'mother/mothers', **bbaabba/babbaabba** 'father/fathers'). For a handful class 2a nouns the (plural) prefix can be either **ø-** or **ba-** (e.g. **malaika** (freq. = 2) or **bamalaika** (freq. = 93) 'angels', **namwandu** (freq. = 3) or **banamwandu** (freq. = 23) 'widows'). Nearly three-quarter (74%) of the types in class 1a still refer to people (e.g. **nabyama** 'chairperson', **kalaani** 'secretary'), although more than half (55%) of those are proper names referring to people (e.g. **Museveni**, **Ndimugezi**), while another 17% are actually personified animals (e.g. **Wankudu** 'Mr/Ms Tortoise', **Wampala** 'Mr/Ms Leopard'). The second largest category is nature (e.g. **zaabbu** 'gold', **musisi** 'earthquake'), followed by both true abstracts (e.g. **isegya** 'spirit', **sitaani** 'devil') and man-made abstracts (e.g. **gulaama** 'grammar', **nantabila** 'verb'). Smaller categories include: flora (e.g. **fene** 'jackfruit', **kaawa** 'coffee') and man-made concretes (e.g. **sigala** 'cigarette', **zaala** 'board game'). Also attested are: liquids (**kyayi** 'tea', **sooda** 'soda') and a human body part (**situka** 'dandruff'). The full distribution, both in terms of types and tokens, is shown in Appendix 3.3.

4.4. Class 2a (8 types; 436 tokens)

Class 2a is very small, as most types from this class are infrequent. The (plural) noun prefix for the few frequent types in class 2a is always: **ba-** (the zero-prefix mentioned under §4.3 is not frequent enough to feature). All nouns in class 2a refer to people (e.g. **badhaadha** 'grandparents', **bamulekwa** 'orphans'), except for two (**bamalaika** 'angels', **bakatonda** 'gods').

4.5. Class 3 (171 types; 7,472 tokens)

All nouns in class 3 take the prefix: (**o**)**mu-**. Three-quarter (75%) of the class 3 noun types also have a corresponding (plural) form in class 4, one quarter (25%) is attested in class 3 only (e.g. **omwenkanonkano** 'gender awareness', **omuwuudu** 'greed'). All class 3 sound changes are straightforward semivocalizations. The semantic import of this class is spread over many categories, including: man-made concretes (e.g. **omugaati** 'bread', **omulyango** 'door'), abstracts (e.g. **omukisa** 'luck; blessing', **omusoso** 'habit'), human body parts (e.g. **omukono** 'hand', **omutwe** 'head'), nature (e.g. **omulilo** 'fire', **omusana** 'sun'), man-made abstracts (e.g. **omusolo** 'tax', **omuluka** 'level of leadership'), liquids (e.g. **omusaayi** 'blood', **omubisi** 'banana brew'), flora (e.g. **omuyembe** 'mango', **omutyele** 'rice'), fauna (e.g. **omusu** 'rat', **omusota** 'snake'), and even people (e.g. **omukwano** 'friend', **omusengo** 'an accused' – homonymous with 'gift').

4.6. Class 4 (73 types; 2,406 tokens)

All nouns in class 4 take the (plural) prefix: **(e)mi-**. Nine out of every ten noun types in class 4 (88%) also have a corresponding (singular) form in class 3, the others (12%) are only attested in class 4 (e.g. **emilaala** 'peace; freedom', **emilonso** 'social norms'). All class 4 sound changes are straightforward semivocalizations. The semantic import of this class is also spread over many categories, and includes: abstracts (e.g. **emidoobaano** 'unsuccessfulness', **emigaso** 'advantages'), human body parts (e.g. **emikono** 'hands', **emitwe** 'heads'), nature (e.g. **emyezi** 'months', **emyaka** 'years'), flora (e.g. **emiti** 'trees', **emizabbibbu** 'date trees'), man-made concretes (e.g. **emitala** 'villages', **emigugu** 'luggage'), and people (e.g. **emikwano** 'friends', **emisengo** 'the accused' – homonymous with 'gifts').

4.7. Class 5 (120 types; 5,111 tokens)

Six out of every ten noun types in class 5 (63%) have a corresponding (plural) form in class 6; the others (37%) are only attested in class 5. There are furthermore two forms of the class 5 noun prefix: **(e)i-** and **(e)li-**. For those with a corresponding (plural) form in class 6, 85% take the prefix **(e)i-** (e.g. **eibandha** 'debt', **eiteeka** 'law'); 15% the prefix **(e)li-** (e.g. **elyato** 'boat', **eliiso** 'eye'). Class 5 nouns without a corresponding (plural) form in class 6 always take the prefix **(e)i-** (e.g. **eibbugumu** 'heat', **eisuubi** 'hope'). The class 5 sound changes are again semivocalizations. Over 60% of the nouns in this class belong to just three semantic categories: man-made concretes (e.g. **eikonelo** 'chair', **eiwanika** 'cemetery; dictionary'), abstracts (e.g. **eisanhu** 'happiness', **eisila** 'emphasis'), and nature (e.g. **eigulu** 'heaven, sky', **eitaka** 'land, soil'). Also found in class 5 are: human body parts (e.g. **eigumba** 'bone', **eiliba** 'skin' – polysemous with 'hide'), flora (e.g. **eitooke** 'banana (cooked)', **eisubi** 'grass'), man-made abstracts (e.g. **eisomo** 'course', **eliina** 'name'), liquids (e.g. **einhila** 'mucus', **eiva** 'sauce'), people (e.g. **eizaile** 'group of children', **eikuukuubila** 'group of people'), and fauna (e.g. **eigi** 'egg', **ikoli** 'eagle').

4.8. Class 6 (130 types; 6,073 tokens)

As many as 63% of the nouns in class 6 have corresponding (singular) forms in class 5 (e.g. **amateeka** 'laws', **amaiso** 'eyes'), just 30% are only attested in class 6 (e.g. **amasaanhalaze** 'electricity', **amatanta** 'saliva'), and a further 5% have corresponding (singular) forms in class 15 (e.g. **amatu** 'ears', **amagulu** 'legs'). There is one case (among the frequent noun types) of a class 6 noun with a corresponding (singular) form in class 9 (**amayumba** 'houses'). The form of the class 6 prefix is always: **(a)ma-**, as may be seen in Appendix 8.1. In gender 5/6, 68% take the noun prefix **(e)i-** in class 5, 32% the noun prefix **(e)li-**. The applicable sound changes are shown in Appendix 8.2. The three main semantic categories, again good for over 60%, are: human body parts (e.g. **amatama** 'cheeks', **amabunda** 'stomach' – polysemous with 'pregnancy'), abstracts (e.g. **amagoba** 'profits',

amazima 'truth'), and man-made concretes (e.g. **amasasi** 'bullets', **amagombe** 'grave'). Smaller categories include: liquids (e.g. **amaziga** 'tears', **amaadhi** 'water'), flora (e.g. **amaido** 'ground nuts', **amenvu** 'bananas (eaten raw)'), fauna (e.g. **amagi** 'eggs', **amooya** 'feathers'), and man-made abstracts (e.g. **masomo** 'courses', **amaina** 'names').

4.9. Class 7 (201 types; 12,072 tokens)

Nine out of every ten noun types from class 7 (89%) also have a corresponding (plural) form in class 8 (e.g. **ekimuli** 'flower', **ekyuma** 'metal'); the others (11%) are only attested in class 7 (e.g. **ekinhagansi** 'respect', **ekitangaala** 'light; transparent; exposure'). The class 7 noun prefix is always: (**e**)**ki-**, and gives way to semivocalizations when attached to vowel-initial roots and stems. When it comes to the semantic import of class 7, one is dealing with a very heterogeneous bag, many of which do not fit any of our ten semantic categories (e.g. **ekigwo** 'a fall or a wrestle to the ground', **ekimega** 'piece cut from a whole (of food); part'). Two categories stand out, however: man-made concretes (e.g. **ekidomola** 'jerrycan', **ekiso** 'big knife') and abstracts (e.g. **ekibi** 'sin', **ekidhuubo** 'thought; idea'). Smaller categories include: flora (e.g. **ekigogo** 'banana plant', **ekibala** 'fruit'), fauna (e.g. **ekisolo** 'animal', **ekinhonhi** 'bird'), nature (e.g. **ekiswa** 'ant hill', **kibali** 'swamp'), human body parts (e.g. **ekigele** 'foot', **ekinkumu** 'thumb' – polysemous with 'signature'), people (e.g. **ekikunsu** and **ekilindi** 'group of people'), and man-made abstracts (e.g. **ekifunze** 'abbreviation', **ekibinuko** 'party; occasion').

4.10. Class 8 (146 types; 6,647 tokens)

In many a way, class 8 is the mirror of class 7. Nine out of every ten noun types from class 8 (86%) have a corresponding (singular) form in class 7 (e.g. **ebimuli** 'flowers', **ebyuma** 'metals'); with the others (14%) only attested in class 8 (e.g. **ebisale** 'rates; fees', **ebyobuwangwa** 'pertaining to social norms and values'). The class 8 noun prefix is always: (**e**)**bi-**, and again gives way to semivocalizations when attached to vowel-initial roots and stems. Here too, the percentage of unclassifiable types (i.e. 'others') is high (e.g. **ebibono** 'doings', **ebikumi** 'tens'), in addition to abstracts (e.g. **ebisilaani** 'bad lucks', **ebyobugaiga** 'riches'), man-made concretes (e.g. **ebizimbe** 'buildings', **ebikopo** 'cups'), man-made abstracts (e.g. **ebyemizaanho** 'pertaining to sports', **ebikoiko** 'question-answer games'), flora (e.g. **ebidhandhaali** 'beans', **ebita** 'gourds'), fauna (e.g. **ebyenhandha** 'fish(es)', **ebiwuuka** 'insects'), human body parts (e.g. **ebikonde** 'fists', **ebyenda** 'intestines; offal'), liquids (**ebizigo** 'body oils'), and people (**ebika** 'clans' – polysemous with 'types').

4.11. Class 9 (385 types; 15,136 tokens)

As may be seen from Appendix 11.1, nouns in class 9 have corresponding (plural) forms in either class 10 (49% of the cases) or class 6 (4% of the cases), while the others (47% of the cases) are only attested in class 9. For nouns in gender 9/10, the form of the class 9 noun prefixes are: **(e)N-** (83% of the cases, e.g. **ensonga** 'reason', **ensi** 'world; country') and **(e)ø-** (17% of the cases, e.g. **esaala** 'prayer', **ewiiki** 'week'); for nouns in gender 9, the form of the class 9 noun prefixes are also: **(e)N-** (70% of the cases, e.g. **emmele** 'food', **endhala** 'hunger') and **(e)ø-** (30% of the cases, e.g. **ebbeeyi** 'price; cost', **gomesi** 'female traditional wear'); for nouns in gender 9/6, one instance is found of the noun prefix **eN-** (**enthupa** 'bottle'), the others take **(e)ø-** (e.g. **ebbaluwa** 'letter', **egaali** 'bicycle'). The various (and many) sound changes that apply are listed in Appendix 11.2, the semantic import in Appendix 11.3. Three categories make up more than 70% of all class 9 nouns: man-made concretes (e.g. **engule** 'crown', **empiima** 'short sword'), abstracts (e.g. **ensonhi** 'shyness', **ensaalwa** 'envy'), and fauna (e.g. **entaama** 'sheep', **enkoko** 'chicken'). Smaller categories include: nature (e.g. **emuunienie** 'star', **mpuku** 'cave'), flora (e.g. **emmwanhi** 'coffee bean', **empeke** 'grain' – polysemous with 'solid medicine'), man-made abstracts (**vawulo** 'vowel', **Paasika** 'Easter'), human body parts (e.g. **ennhindo** 'nose', **enkende** 'waist'), people (**poliisi** 'police'), and liquids (**nkolwa** 'sauce of water mixed with salt' – homonymous with 'bird').

4.12. Class 10 (91 types; 2,705 tokens)

As may be seen from Appendix 12.1, nouns in class 10 always have corresponding (singular) forms – most frequently nouns in class 11 (57% of the cases), followed by nouns in class 9 (41% of the cases), and nouns in class 14 (2% of the cases). For the gender 11/10, the form of the class 10 (plural) noun prefix is: **(e)N-** (e.g. **ennimi** 'tongues; languages', **entalo** 'wars'); for the gender 9/10 the forms of the class 10 (plural) noun prefixes are: **(e)N-** (78% of the cases, e.g. **ensonga** 'reasons', **ente** 'cows') and **(e)ø-** (22% of the cases, e.g. **langi** 'colours', **talanta** 'talents'); and for the gender 14/10 the form of the class 10 (plural) noun prefix is: **eN-** (**endwaile** 'diseases'). The various (and many) sound changes that apply are listed in Appendix 12.2, the semantic import in Appendix 12.3. Three categories make up about 70% of all class 10 nouns: abstracts (e.g. **enkabi** 'peace', **entaka** 'stubbornness'), man-made concretes (e.g. **embili** 'palaces', **emmotoka** 'cars'), and human body parts (e.g. **emba** 'jaws', **enkumu** 'nails'). Smaller categories include: man-made abstracts (e.g. **ennhemba** 'songs', **enfumo** 'folk tales'), flora (e.g. **embooli** 'potatoes', **endagala** 'banana leaves'), fauna (e.g. **entaama** 'sheep', **enkoko** 'chickens'), and nature (e.g. **ennaku** 'days' – homonymous with 'sadness').

4.13. Class 11 (99 types; 5,025 tokens)

Three-quarter (76%) of the class 11 nouns have corresponding (plural) forms in class 10 (e.g. **olulimi** 'tongue; language', **olutalo** 'war'); the others (24%) are only

attested in class 11 (e.g. **olwali** 'jocular talk', **Olusooka** 'New Year's day'). The form of the class 11 noun prefix is always: (**o**)**lu-**. Each gender is governed by its own sound changes: For gender 11/10, class 11, changes are only attested when the root-initial letter is the semivowel **y-** (where the sound change itself depends on the environment); and for gender 11 only semivocalizations are attested. Semantically, nearly all nouns belong to just four categories: abstracts (e.g. **olugambo** 'gossip', **olukusa** 'permission'), man-made concretes (e.g. **oluguudo** 'road', **olukoba** 'elastic string; tape measure'), man-made abstracts (e.g. **Olusoga** 'Lusoga', **Olungeleza** 'English'), and nature (e.g. **olusozi** 'hill; mountain', **olunaku** 'day'). Tiny categories include: human body parts (**olwala** 'finger', **oluwusu** 'skin') and flora (**olwendo** 'gourd', **olulagala** 'banana leaf').

4.14. Class 12 (61 types; 1,655 tokens)

Three-quarter (75%) of the nouns in class 12 have a corresponding (plural) form in class 14 (e.g. **akasuwa** 'small pot', **akalulu** 'election; vote'); the others (25%) are only attested in class 12 (e.g. **akanhagansi** 'respect', **akabina** 'bottom, buttocks'). The form of the class 12 noun prefix is always: (**a**)**ka-**. For the gender 12/14 semivocalizations are attested. About one third of the class 12 nouns are man-made concretes (e.g. **akatabo** 'small book', **akamanhiso** 'label'); the other categories include: abstracts (e.g. **akawoowo** 'good scent', **kaladaali** 'pompous behaviour'), human body parts (e.g. **kagulu** 'small leg', **akasolo** 'penis' – homonymous with 'small animal'), fauna (e.g. **akawuuka** 'worm; small insect', **kayima** 'hare'), people (e.g. **akagenge** 'small leper; leprosy', **akasaadha** 'small man'), man-made abstracts (e.g. **akawango** 'affix', **kagambo** 'small word'), nature (e.g. **akabaale** 'small stone', **kasozi** 'small hill; small mountain'), and flora (**akendo** 'small gourd', **kati** 'small stick'). Cutting across the semantic categories, and as may be noted from most glosses in this section, class 12 further contains many diminutives. (More will be said about this aspect in Section 6 below.)

4.15. Class 14 (178 types; 5,909 tokens)

About 87% of the class 14 nouns are only attested in this class (e.g. **obwenzi** 'promiscuity', **obulimi** 'farming'); the other 13% have a corresponding (singular) form in class 12 (e.g. **obusuwa** 'small pots', **obululu** 'votes'). The form of the class 14 noun prefix is always: (**o**)**bu-**. All sound changes in this class are semivocalizations. That class 14 is the abstract class par excellence in Bantu is also confirmed in Lusoga, with seven out of every ten class 14 nouns being true abstracts (e.g. **obusungu** 'anger', **obwilugavu** 'blackness'). The other semantic categories include: nature (e.g. **obulwaile** 'disease(s)', **obwile** 'time; night'), man-made concretes (e.g. **obukwenda** 'money exchanged for love matters', **obulili** 'bed(s)'), flora (e.g. **obutunda** 'passion fruits; passion-fruit juice', **obuwunga** 'seed powder'), fauna (e.g. **obusa** 'cow dung', **obusili** 'small mosquitoes'), man-made abstracts (e.g. **obufumbo** 'marriage institution', **obuwangwa** 'social norms and values'), human body parts (e.g. **obwala** 'fingers; hands', **obwongo** 'brain' –

polysemous with 'intellect'), liquids (**bwino** 'ink', **buugi** 'porridge'), and people (**obwana** 'small children').

4.16. Class 15 (1 type; 24 tokens)

Apart from the infinitive nouns (which are not included in this study), only one other noun type is frequent enough to make it into class 15, namely the human body part: **kutu** 'ear'. Including this noun in class 15 is based on the fact that the form of the noun class prefix is the same as that of the infinitive nouns: (**o**)**ku-**. Doke (1935:64) suggests sub-numbering this class 15a. Its corresponding (plural) form is found in class 6: **matu** 'ears'. (Observe that the frequency of the singular of **magulu** 'legs', mentioned in §4.8, namely **kugulu** 'leg', is only 2, which is why it does not appear here.)

4.17. Class 16 (8 types; 716 tokens)

The form of the class 16 noun prefix is always: (**a**)**wa-**, and invariably refers to locality. Examples include: **wansi** 'down', **waigulu** 'up; above', **wagati** 'in the middle', **awaka** 'at home, in a home', and **wantu** 'a certain place'.

4.18. Class 20 (1 type; 11 tokens)

Only one noun type is frequent enough to make it into class 20: **ogusota** 'big snake'. The form of the class 20 noun prefix is: (**o**)**gu-**. Observe that received Bantu knowledge (see Welmers (1973) for Proto-Bantu, and Kadima (1969) for Lusoga in particular) would place a corresponding (plural) form in class 22, with as plural noun prefix: (**a**)**ga-**, but this plural is unattested in the top-frequent section of the corpus studied. Received knowledge also tells us that class 20 contains augmentatives, which is borne out by this single example.

4.19. Class 23 (81 types; 5,968 tokens)

There are two forms of the class 23 noun prefix: (**e**) **ø-** and (**e**) **bu-**. The pre-prefix **e** 'at; to; from; …; of' is written disjunctively, with the nouns themselves mostly proper names referring to places, whether indigenous or foreign. Frequent examples include: Busoga, Uganda, Jinja, Iganga, Kampala, Africa, Makerere, Bugiri, etc.

5. The Lusoga noun class system

The data presented in Section 4 (§4.1 through §4.19) may now be summarized in various ways. The first is shown in Figure 3, which is a quantified schematic representation of the main relations between the various classes uncovered.

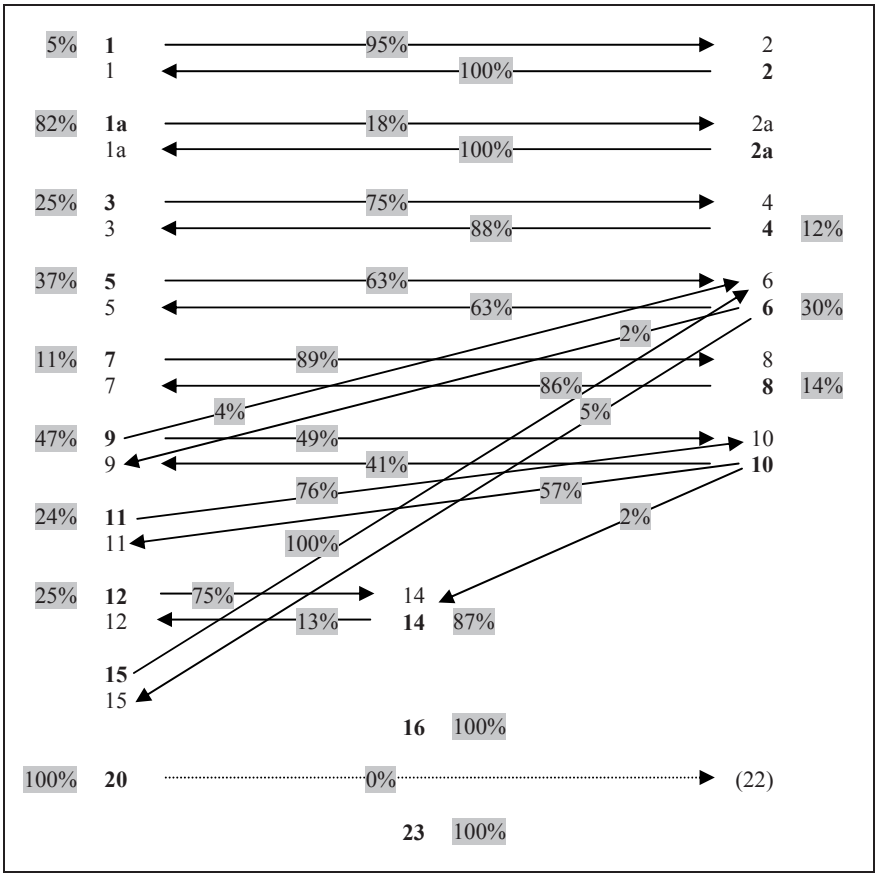

Figure 3: The Lusoga noun class system quantified.

This quantified schematic representation may be read as follows. For example, for gender 3/4: While 75% of the class 3 nouns have a corresponding form in class 4, an even higher number of 88% of the class 4 nouns have a corresponding form in class 3; those without corresponding forms are only attested in class 3 (25%) and class 4 (12%) respectively. Or, for nouns in class 6: When encountering an unknown or new noun in class 6, the chance that it belongs to gender 9/6 is 2%, while it is 5% for gender 15/6, 30% for gender 6, and as much as 63% for gender 5/6. Or even, a (plural) form from class 10 will have a corresponding (singular) form in class 11 in as many as 57% of the cases, in class 9 in 41% of the cases, and in class 14 in only 2% of the cases. Nouns in class 10 thus always have a corresponding (singular) form. Such information is non-trivial, and goes beyond the mere distributional description. In a modern word-based dictionary for Lusoga for example – in other words, in dictionaries that move away from the linguistically elegant but user-unfriendly stem-based approach to lemmatization (cf. De Schryver 2008, Nabirye 2009c) – users can make an 'informed guess' as to where nouns are

most likely to be found when only so-called 'singulars' have been fully treated. Or, in the field of natural language processing, a network such as Figure 3, together with its relative weights, provides crucial information on the likeliness of certain forms/pairs and their meanings. In other words, rather than provide users or machines with all the possible forms, the probable ones can be offered, graded according to their attested occurrence frequencies.

It is convenient to view the left-hand side of Figure 3 (thus classes 1, 1a, 3, 5, 7, 9, 11, 12, 15 and 20) as singular forms, with corresponding plural forms on the right-hand side (thus classes 2, 2a, 4, 6, 8, 10 (and 22)), and vice versa. While this may be useful and correct in a good number of cases, corpus evidence shows that this certainly does not hold for all nouns.

When attempting to uncover the true meaning of each and every Lusoga noun, one should not be tempted to re-project the English glosses back onto the Lusoga forms (compare also Louwrens 1992:110-111). In this regard, one could for example be tempted to assign a singular status to the following class 10 nouns: **enkabi** 'peace' and **entaka** 'stubbornness'. Corpus evidence (in the form of a study of the concordial agreements) in conjunction with the noun meanings in context (assigned to these nouns by a trained mother-tongue speaker) tells us that **enkabi** occurs both as a singular in class 9 (freq. 73) and as a plural in class 10 (freq. 33), even though both may be translated into (idiomatic) English as the single 'peace'. Likewise, **entaka** 'stubbornness' occurs both as a singular in class 9 (freq. 28) and a plural in class 10 (freq. 10). The same is true for singular-plural pairs in other genders, for example: **omudoobaano** 'unsuccessfulness' in class 3 and its corresponding **emidoobaano** 'unsuccessfulness' in class 4. Plural-looking glosses may also confuse. In (the singular) class 12 one for instance finds **akabina** 'buttocks', with a corresponding (plural) form in class 14. In this case it may be handy to use a different gloss: **akabina** 'bottom' and **obubina** 'bottoms'. (To complete the picture: one uses a different noun to refer to one side of the buttocks: **eitako** '(one) buttock'/ **amatako** 'buttocks'.) Yet, there are definitely nouns with singular meanings in **so-**called plural classes: **ebyobuwangwa** 'pertaining to social norms and values' was one of those mentioned above.

In Figure 3, class 14 was placed in the middle, as it can appear as a corresponding plural (of nouns in class 12, e.g. **akatale** 'market' / **obutale** 'markets') as well as a corresponding singular (of nouns in class 10, e.g. **obulwaile** 'disease' / **endwaile** 'diseases'). The (locative) classes 16 and 23 were also placed in the middle, as they are not governed by singularity or plurality. Nouns in gender 14 moreover exhibit both singular and plural characteristics, depending on the context. Examples include: **obusobozi** 'ability/abilities', **obuzibu** 'difficulty/difficulties', and **obweyamo** 'reference/references'. The same is noticed for all one-class genders in Figure 3. This is especially so for (in decreasing order) genders 1a, 9, 5 and 6. Examples for gender 1a include: **taaba** 'tobacco/tobaccos', **Saasila** 'Sunday/Sundays', and **nakeewuunia** 'interjection/interjections'; for gender 9: **embuga** 'court/courts', **embalilila** 'budget/budgets', and **mbogo** 'buffalo/buffaloes'; for gender 5: **eisuubi** 'hope/hopes', **igulu** 'heaven; sky/heavens; skies', and **eiva** 'sauce/sauces'; for gender 6: **amaanhi** 'energy/energies', **amakobo** 'conversation/conversations', and **amaka** 'home/homes'. From the moment one takes the context into account, one

thus realizes that *singularia tantum* (the left-hand one-class genders in Figure 3), as well as *pluralia tantum* (the right-hand one-class genders in Figure 3) are often misnomers, as many one-class genders have both singular and plural uses.

Rather than (or in addition to) true plurals, the plural may also refer to (different) types of the item in question. Examples for gender 14 include: **obusungu** 'anger/types of anger', **obunafu** 'laziness/types of laziness', and **obwibuka** 'luck/types of luck'; for gender 1a: **situka** 'dandruff/types of dandruff', **duuma** 'maize/types of maize', and **mwogo** 'cassava/types of cassava'; for gender 9: **emmamba** 'meat/types of meat', **ensaalwa** 'envy/types of envy', and **enkungu** 'dust/types of dust'; for gender 5: **eibbugumu** 'heat/types of heat', **eilalu** 'madness/types of madness', and **iwali** 'jealousy/types of jealousy'; for gender 6: **amasaanhalaze** 'electricity/types of electricity', **amata** 'milk/types of milk', and **amailu** 'greed/types of greed'; etc. Clearly, then, mass nouns often populate the one-class genders.

Further complicating the neat singular-plural pairings is the fact that certain senses will disappear or even appear when one moves between the corresponding classes. For instance, while **akalulu** means 'election; vote', for the corresponding plural **obululu**, only the meaning votes is attested in the corpus – the meaning election was lost. Conversely, while **akatunda** means 'passion fruit', the corresponding plural **obutunda** means 'passion fruits; passion-fruit juice' – the meaning 'passion-fruit juice' was added.

6. Building nouns in Lusoga

In addition to the relations summarized in Figure 3, most if not all classes and genders attract roots and stems, with which new nouns with new non-random meanings are formed. The most obvious is certainly class 12 (and by extension gender 12/14) which not only contains more nouns referring to small items than any other class, but is also used to make new diminutive forms. Transferring the noun root **-yendo** from gender 11/10 to gender 12/14, one consequently obtains: **olwendo** 'gourd'/**ennhendo** 'gourds' > **akendo** 'small gourd'/**obwendo** 'small gourds'. In the process, meanings may also appear or disappear. For example from 7/8 to 12/14: **ekiwuuka** 'insect'/**ebiwuuka** 'insects' > **akawuuka** 'worm; small insect'/**obuwuuka** 'worms; small insects' – where 'worm(s)' has been added to both the singular and the plural; or, also from 11/10 to 12/14: **olwala** 'finger; nail'/**endhala** 'fingers; nails' > **akaala** 'small finger'/**obwala** 'fingers; hands' – where the latter reverted to 'fingers' (rather than 'small fingers', thus losing the small part), while gaining the additional meaning 'hands', and where the meaning 'nail(s)' is also lost in the process.

On a lexical level, noun class 12 (and gender 12/14) as well as its noun prefix (**a**)**ka-** (and noun prefix (**o**)**bu-**) can therefore be seen as a foretoken of diminutives. Class 12 also exhibits a pragmatic aspect, namely that of amelioration, and thus brings together amelioratives. For instance, the difference between **ekinhagansi** 'respect' in gender 7 and **akanhagansi** 'respect' in gender 12 is that the latter has a positive connotation. Depending on the context, referring to small people or things can also mean the opposite pragmatically, and thus refer to pejoratives: **ekintu** 'thing' > **akantu** 'small thing' or 'bad thing'.

Conversely, when roots and stems are moved to class 7 (and gender 7/8), the new forms have an additional augmentative/ameliorative import: **akaso** 'knife'/**obuso** 'knives' > **ekiso** 'big knife; operation'/**ebiso** 'big knives; operations'. Or see the difference between: **olugoye** 'cloth'/**engoye** 'clothes' (gender 11/10, neutral) vs. **ekigoye** 'large cloth'/**ebigoye** 'large clothes' (gender 7/8, augmentative/ameliorative) vs. **akagoye** 'small cloth'/**obugoye** 'small clothes' (gender 12/14, diminutive/ameliorative/pejorative). As seen in §4.18, augmentatives are also found in class 20 (and gender 20/22).

Cross-comparing the various sections of §4 further indicates that personifications and proper names referring to people are only found in gender 1a, that the class 14 noun prefix is the main one used to form abstract concepts, that gender 16 brings together locatives and gender 23 proper names referring to places, and that loanwords are mostly found in gender 9/10.

Of course, a corpus-based approach allows one to go beyond the type of generalizations just discussed, and to fully account for the various noun formation processes, with their linked meanings, together with a quantification of each. This was done for the 2,263 nouns with a frequency of at least ten in the corpus, with the results as shown in Appendix 16.

One may firstly observe that about two thirds of the nouns (1,544 to be exact, or 68%) are simply built by attaching a noun prefix to a noun root (i.e. NP + noun root). As seen above, some of those noun roots may combine with various noun prefixes, and depending on the gender, they acquire varying meanings in the process. In genders 9/6, 15/6 and 20, this is the sole noun formation process. In gender 23 this strategy is used for 98% of the nouns, in gender 6 for 87% of the nouns, etc. as shown in Table 4.

Gender	%	Gender	%
9/6	100.00	12/14, 12	67.86
15/6	100.00	1/2, 1	56.91
20	100.00	7/8, 7, 8	54.74
23	97.53	1a/2a	54.55
6	87.18	1a	52.07
5/6, 5	84.65	16	50.00
9/10, 9	79.80	14	48.39
3/4, 3, 4	77.87	8	25.00
11/10, 11	70.86	14/10	0.00

Table 4: Percentage of nouns formed according to 'NP + noun root'.

Secondly, if two thirds of the nouns are so-called inherent nouns (formed according to NP + noun root), one third must be constructed or derived through other means. A surprisingly high overall number of 93 constructions are seen (in the top-frequent Lusoga section of the corpus looked at), with all those with a frequency of at least two listed and exemplified in Appendix 16. For the genders 1/2 and 1, for example, in addition to 57% 'inherent nouns', 17% follow the pattern 'NP + V + i', 12%

the pattern 'NP + V + a', 9% the pattern 'NP + V + perfective form', etc. Each of those patterns moreover results in a well-defined meaning, here twice 'person who 'verbs'', then 'person who is/has 'verbed'', etc.

As can be deduced from Appendix 16, such derived nouns may be derived from verbs, other nouns, pronouns, numbers, and adjective roots, in combination with various formatives and terminating vowels as affixes and circumfixes.

Quantifying the various patterns, as done in Appendix 16, also goes beyond the mere description within a distributional corpus analytic framework. In addition to applications in lexicography and natural language processing, knowing which patterns are frequent and which ones are not, may for example assist compilers of textbooks in making sure all core issues are covered, while at the same time informing them about the issues that may be carried over to more advanced levels (such as, say, the large number of patterns for class 1a, used to make proper names that refer to people). As a result, language teachers and students alike will be able to focus on what is truly common first.

When building or constructing nouns, sound changes apply, as seen in the various morphophonology tables in the addenda. Here, it may be advantageous to collapse the data as a first approach with teaching purposes in mind (the details per class are covered in the said addenda). Collapsing all the observed sound changes and retabulating them results in the data shown in Table 5.

Rule	Sum N	Rule	Sum N	Rule	Sum N
a+e>e/_NC	3	**N+b>mm/_N**	14	**u+a>wa**	46
a+e>ee	7	**N+b>mb**	74	**u+e>we**	34
a+o>oo	2	**N+g>ŋŋ/_N**	10	**u+i>wi**	36
a+y>e/_NC	2	**N+l>nn/_N**	18	**u+o>wo**	14
a+y>oo	1	**N+l>nd**	47	**u+y>wi/_i**	2
i+a>ii/_D	2	**N+m>mm**	30	**u+y>we/_NC**	8
i+a>ya	61	**N+p>mp**	8	**u+y>wa**	1
i+e>ye	41	**N+w>mp**	60		
i+o>yo	24	**N+y>mp/_i**	15		
i+u>yu	8	**N+y>ndh/_i**	2		
i+y>y	2	**N+y>nnh/_N**	48		
		N+y>mp	4		
		N+y>ndh	33		

Table 5: Collapsed morphophonology data applicable to nouns (in alphabetical order).

When vowels come into contact with other vowels or semivowels, as is the case for the rules in the outer columns of Table 5, processes of vowel coalescence, semivocalization and vowel elision are attested. When a nasal comes into contact

with consonants, glides and semivowels, processes such as syllabification, assimilation and plosivication are attested, as seen in the centre column of Table 5.

The rules listed in Table 5 are mutually exclusive – and as such may easily be memorized by humans, and input into machines – except for one set: **N + y > mp** or **N + y > ndh**. At face value, corpus linguistics has run its course here, as nothing on the surface level helps to disambiguate between these varying sound changes. Indeed, the only way to account for these diverging rules is to postulate an underlying **/p/** from Proto-Bantu ***p**, which weakens to either **[w]** or **[y]** on the surface level, as was done by Hyman & Katamba (1999:369-84, 401-2). As such, PB ***p** weakens and assimilates to **[y]** before front vowels. This results in rules such as:

N + y > mp	**akayindi** / **empindi**	'peas'	**N + [y] (*p) > mp**
N + w > mp	**akawale** / **empale**	'trousers; shorts'	**N + [w] (*p) > mp**

The other consideration is the assimilation of the underlying palatal glide **/j/** (spelled **<y>**) to consonants. Hyman & Katamba (1999:399, 412 note 75) give **/t c k/** realized as **[s]** and **/d l j g/** realized as **[z]** in Luganda (EJ15). The **[z]** is realized as **/dh/** in Lusoga, hence the rule:

N + y > ndh	**akayu** / **endhu**	'house'	**N + /j/ > ndh**
	akayuba / **endhuba**	'sun'	**N + /j/ > ndh**

Corpus linguistics is not entirely powerless on the surface level, however. In the environment of an 'i' the statistics indicate 15 instances of **N + y > mp/_i** versus only 2 of **N + y > ndh/_i**; while in all other environments only 4 cases are attested of **N + y > mp** versus 33 cases of **N + y > ndh**. Both humans and machines are thus very likely to 'get it right' in about 88 to 89% of the cases (i.e. 15 out of 17; 33 out of 37), and this without the need for a recourse to any knowledge of Proto-Bantu.

To complete the picture, one more orthographic convention that applies to the nouns as a whole concerns contractions. These contractions are seen when possessive concords (PCs) 'of' are attached to the nouns that follow, or when nouns are preceded by the conjunction **ni** 'and'. See the left side, respectively right side, of Table 6.

Rule	Sum N	Rule	Sum N
a + a > 'a	30	**i + a > 'a**	15
a + e > 'e	27	**i + e > 'e**	47
a + o > 'o	22	**i + o > 'o**	19

Table 6: Contraction rules applicable to nouns (PCs left, ni right).

When for example applied to the class 23 noun **e** 'at; to; from; …', **ni + e** becomes **n'e**, while Table 7 shows the full paradigm for the PCs (with the underlined forms counted in this study).

Cl.	PC	PC +E	Freq.	PC-pp +E	Freq.	Cl.	PC	PC +E	Freq.	PC-pp +E	Freq.
1	**(o)wa**	**ow'e**	34	**w'e**	2	10	**(e)dha**	**edh'e**	1	**dh'e**	0
2	**(a)ba**	**ab'e**	156	**b'e**	6	11	**(o)lwa**	**olw'e**	0	**lw'e**	1
3	**(o)gwa**	**ogw'e**	5	**gw'e**	0	12	**(a)ka**	**ak'e**	0	**k'e**	0
4	**(e)gya**	**egy'e**	0	**gy'e**	0	14	**(o)bwe**	**obw'e**	0	**bw'e**	3
5	**(e)lya**	**ely'e**	10	**ly'e**	5	15	**(o)kwa**	**okw'e**	4	**kw'e**	0
6	**(a)ga**	**ag'e**	0	**g'e**	0	16	**(o)wa**	**ow'e**	0	**w'e**	0
7	**(e)kya**	**eky'e**	62	**ky'e**	3	20	**(o)gwa**	**ogw'e**	0	**gw'e**	0
8	**(e)bya**	**eby'e**	18	**by'e**	1	22	**(e)ga**	**eg'e**	0	**g'e**	0
9	**(e)ya**	**ey'e**	14	**y'e**	2	23	**(e)ya**	**ey'e**	0	**y'e**	4

Table 7: Contraction rules when attaching a PC 'of' to the class 23 noun E 'at; to; from; …'.

7. The semantic import of the Lusoga noun

We are now in a position to come full circle, and to return to where we started when it comes to the semantic import of nouns in the Bantu languages. To begin with, the semantic categories for each of the 19 Lusoga noun classes uncovered may be set out in function of 100%, as done in Figure 4. Although handy for didactic purposes, a representation such as Figure 4 should be read with care, as all classes are made to look equal when it comes to the number of types in each class. Figure 4 may further be read as the unpacked version of Figure 1, and as such it should be clear that Hendrikse & Poulos's (1992) "continuum interpretation of the Bantu noun class system" is an oversimplification.

An alternative view of this same data may be seen in Figure 5, where the semantic contribution of each class to each semantic category has been plotted, set out in function of 100% for each category.

Figure 4 is a useful view when one needs to summarize the semantic import of each class in isolation; while a study of the so-called singular-plural pairings (e.g. 1a/2a, 3/4, etc.) clearly indicates that the respective classes of the pairs do not exhibit identical distributions. See in this regard for example the different distributions for 1a vs. 2a (in terms of nature vs. abstracts), or 3 vs. 4 (in terms of fauna vs. flora), etc.

Figure 5 is more correct in that one can truly see how each class in isolation contributes meaning. It may come as a surprise, for example, that class 3 contributes nearly as many liquids as class 6.

Both Figure 4 and Figure 5 are partial views, however, and one realizes that each of those two two-dimensional views is actually a projection of a three-dimensional reality. Moreover, setting out each property in function of 100% unnecessarily distorts reality further, as classes with merely a few members – i.e. types – are made to look as important as very large classes. Frequent and infrequent members of each class are also made to look as important as one another in the current views. Therefore, rather than in terms of percentage, and rather than in terms of types, a more realistic view, apart from being three-dimensional, would be one in which true frequencies – i.e. tokens – are plotted. This is exactly what was done in Figure 6.

It is our contention that a three-dimensional representation with (a) noun classes (rather than genders), (b) multiple semantic categories (rather than a concrete-abstract continuum), and (c) tokens (i.e. true corpus frequencies) as axes, is a more realistic view when trying to summarize the semantics of the Bantu noun graphically. Given that the true occurrence of each and every noun is built into this representation (as tokens rather than types are used), this view furthermore reflects true, free-flowing language use.

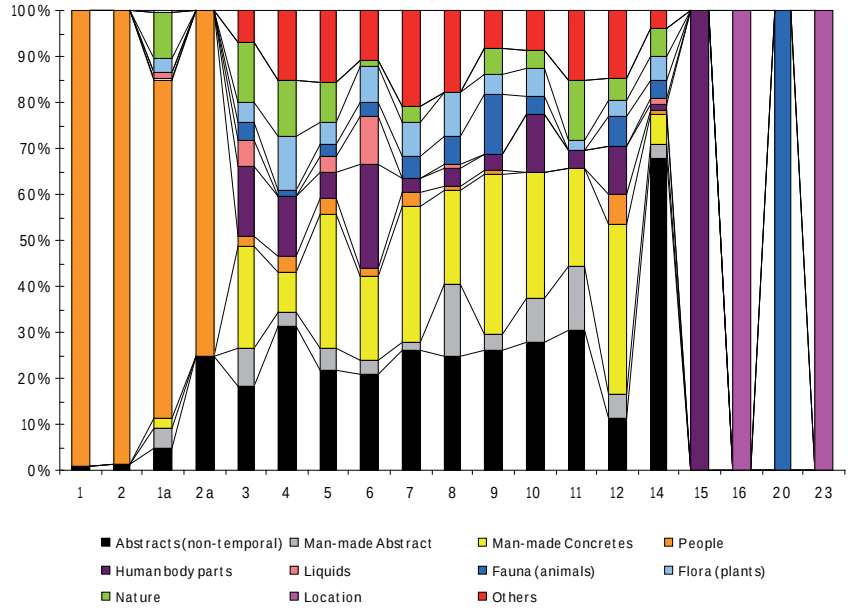

Figure 4: Semantic import of the various noun classes (in terms of types).

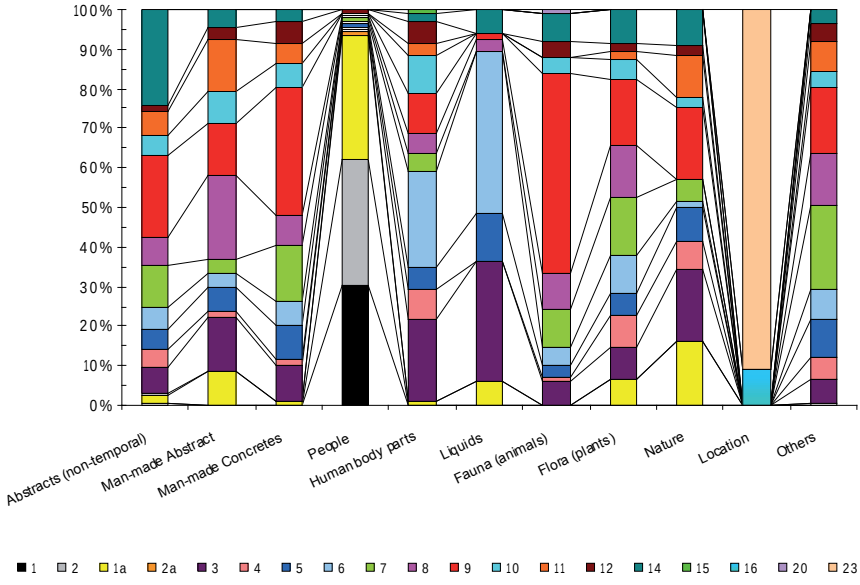

Figure 5: Contribution of the classes to each semantic category (in terms of types).

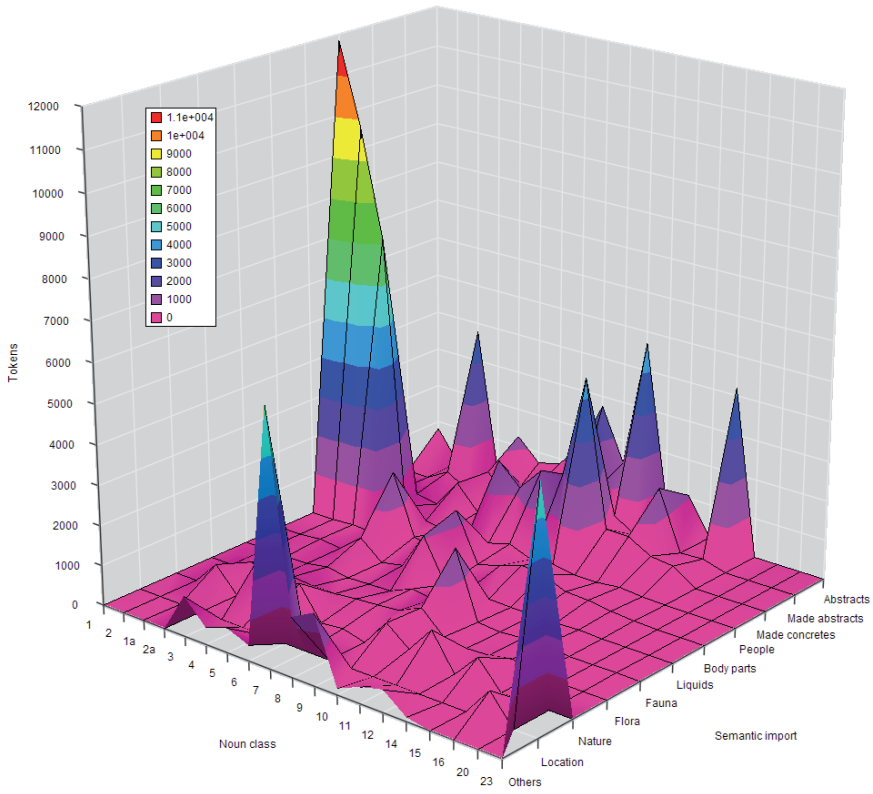

Figure 6: A three-dimensional view of the semantic import of the Lusoga noun.

8. Discussion

The main goal of the above presentation was to illustrate how a distributional corpus analyst could start the grammatical analysis of an undocumented language. As such we hope to have demonstrated its intrinsic value as well as its feasibility. The approach was illustrated for Lusoga, and we are confident that the results also contribute to a better understanding of this particular Bantu language. It stands to reason that studies like the one presented never stand alone. For one, a very large amount of research has already been undertaken for the Bantu languages as a whole, and even though we tried not to be influenced by that earlier work during the building and analysis of the Lusoga corpus itself (Sections 2 and 4–7), one has to concede that it helps to know where one is potentially heading.

For Lusoga in particular we are actually dealing with a mostly undocumented language, as some studies in which Lusoga is featured have indeed been undertaken in the past. These studies include surveys of the interlacustrine Bantu languages, where Lusoga is typically mentioned in comparison only to other languages (e.g. Tucker & Bryan 1957, Matovu 1992, Schoenbrun 1997, Matovu & Walusimbi

2000). Booklets on Lusoga orthography (Kajolya 1990, LULANDA & CRC 2004) and Lusoga grammar (Babyale 1999, Korse 1999a) have also been written. Nabirye (2009b), however, concludes with reference to the former that they are "inconsistent in their description of the Lusoga orthography and their coverage [i]s very shallow" (pp. 178-9), while she characterizes the latter as "a pedestrian consideration of grammar with English translations for tourists" (p. 179). Until the publication of Nabirye's monolingual dictionary (2009a), only wordlists were available, one with English glosses (Korse 1999b), and one with Japanese glosses (Yukawa 2000). As far as we are aware, then, just two scientific publications are entirely dedicated to Lusoga, Steeman (2001) in which a Lusoga play is interlinearized, and Van der Wal (2004) on Lusoga phonology.

8.1. Class system

We are now in a position to summarize the main findings from our distributional corpus analysis (DCA) of the Lusoga noun, and to compare those – where relevant – with outcomes from the earlier studies. To begin with, and with reference to the basic framework of the Lusoga noun class system (Figure 3), one would expect all such frameworks to be rather similar, or even identical. Tucker & Bryan (1957), however, list genders 13, 14/6, and the locatives 17 and 18 for Lusoga, all of which are unattested in our analysis, while they do not mention our attested genders 1a/2a, 9/6, and 14/10. Also, while both studies mention the augmentative, Tucker & Bryan do not mention the diminutive. The main difference, however, lies in our pinpointing of single-class genders in addition: 1, 1a, 3, 5, 7, 9, 11 and 12; and 4, 6, 8 and 14. A comparison with a much later source, Steeman (2001), reveals more or less the same differences: Steeman does not list genders 9/6, 14/10, 15/6, and 23, while listing 17 and 18. He does point out the augmentative and diminutive genders, but none of the single-class genders.

It is not known if techniques other than elicitation were used by Tucker & Bryan, but it is known that Steeman's analysis is based on a single text. We feel that the use of a wide array of texts and text genres, as in our implementation of DCA, allows for a more realistic account. Observe, however, that we deliberately did not consider all noun types from our corpus, as all those with a frequency of less than ten were excluded. While a researcher in a fieldwork setting may be satisfied with a limited number or even a single example of a phenomenon, a distributional corpus analyst will first want to see enough (in our case at least ten instances of) naturally occurring evidence. Larger corpora contain more evidence, by definition, and given that we are currently expanding our Lusoga corpus (adding material from the 1970s and 1980s, as well as transcribing up to a hundred hours of oral material), it will be interesting to see how several of the now excluded nouns will fit into the established noun class system.

In her paper 'Noun Class as Number in Swahili' Contini-Morava (2000) points out "how unilluminating it is to analyze the Swahili data in terms of a binary singular-plural distinction or in terms of class pairing" (p. 11). Instead, she proposes to reanalyse number in Swahili as a combined system of degree of individuation and a continuum of individuation, as shown diagrammatically below:

Continuum → / Degree ↓	concrete individual	abstraction	liquid or continuous mass	mass of homogenous particles	collectivity	replicated individuals
most individuated	1; 3; 5; 7					2; 4; 8
less individuated	11 (includes 14)					
least individuated		6				

Disregarding a few problems with this diagram (such as the lumping of class 14 with class 11, and the absence of gender 9/10 (which she claims is neutral to the scale of individuation and can fall anywhere)), it is true that using a table of two graded scales allows for a more detailed characterization of number in Bantu. Another example of a cognitive semanticist's use of two graded scales in this regard is Hendrikse's (1990:398). Maintaining that, for Southern Bantu, "class 10 is actually nothing else but class 8 stacked onto class 9" (p. 398), he proposes the following diagram to depict the spatial-number properties of the class prefixes in Southern Bantu:

↓ / →	discrete	continuous
multiplex, unbounded	2; 8	4
multiplex, bounded		6
uniplex	1; 3; 5; 7; 9; 11; 14	

We believe that such diagrammatic representations are as generic as our weighted two-dimensional noun class system offered for Lusoga, however. All these approaches, then, are only approximate. They are also the logical outcome of the theoretical frameworks used, cognitive semantics for Contini-Morava and Hendrikse, DCA for us.

Summarizing Sections 4 and 5 we can therefore say that we feel that a notion of the relative distribution of the type and token counts for each noun class (cf. Table 3), in combination with a weighted two-dimensional noun class system (cf. Figure 3) – whereby classes are viewed in isolation in the former, genders in the latter – is a most powerful way to visualize the strength of each node and each link in the structure.

8.2. Construction system

A comparison of the morphophonological rules presented in our work (cf. e.g. Table 5) with the more traditional approach as for example seen in Van der Wal (2004), is decidedly different. Within DCA, one attempts to limit all observations and the analyses thereof to what is observable on the surface level. It was indicated how, in one case, recourse had nonetheless to be taken to Proto-Bantu – up to a point. There may, however, be more theorizing involved. When studying the formation of the noun types, two thirds were found to be inherent, one third derived. A valid question could be: How can one clearly differentiate between the two types? The main strategy used here was to classify nouns as inherent whenever the noun root could

not be right-extended to produce meaningful sequences. Conversely, nouns derived from verbs are typically extendible: add a verbal extension to the verb root, and both the extended verb and the noun derived from this verb stem are meaningful. Also, the final vowel is obligatory on a noun root for it to have any meaning, while it is a grammatical component on a verb root or verb stem. Furthermore, all derived nouns are governed by predictable meanings, as is clear from the derivational formulas cum meanings listed in Appendix 16. Still, a further question could be: How does one know which one is derived from which? Or, could one not postulate that (some of the) verbs are actually derived from nouns? Although we pose the question here, we admit that this issue never surfaced during the analysis. It was, in other words, unproblematic, and may actually be connected to Hopper & Thompson's implicational generalization: "languages often possess rather elaborate morphology whose sole function is to convert verbal roots into Ns, but no morphology whose sole function is to convert nominal roots into Vs" (1984:745).

Summarizing Section 6 we can therefore say that we feel that a quantified enumeration of both nominal morphophonology (cf. e.g. Table 5) and noun constructions cum linked meanings (cf. Appendix 16) provides for a representative picture of the various noun-building issues.

8.3. Semantic system

The three-dimensional semantic-import view for the Lusoga noun offered in Figure 6 is a direct outcome of the DCA framework used. DCA quite literally allows for the addition of a third dimension to the traditional dimensions of classes and genders on the one hand, and semantic categorizations on the other. From the moment Bantuists link the latter two, they seem to undertake this with the aim to do any of three things: (a) disprove that there is a link, (b) prove that there is a link, but only in its original (Proto-Bantu) form, (c) prove that there is a link, which is best analysed within a cognitive framework. Given that the goal in such cases is thus to uncover the existence or non-existence of an (original) underlying system, the data is often manipulated: loanwords (especially recent ones and/or those of non-Bantu origin) may be excluded from the analysis; problematic classes or genders may not be studied; only inherent nouns may be considered (taking out the derived ones); only one form (normally the singular) may be counted for two-class genders; and only noun types may be looked at. For all these aspects our approach has been radically different, again a direct result of DCA: every single frequent noun, no matter its loanword status, was included; all noun classes and genders were studied; both inherent and derived nouns were considered; both forms of all two-class genders were counted; and both noun types and noun tokens were looked at. As a result, Figures 4 and 5 – which give two perspectives on the link between noun classes and semantic categories – should have been more random than any existing description, yet those figures clearly indicate that there is a system, and that that system is not random. The insistence on using occurrence frequencies in naturally occurring language (tokens) rather than single instances of each noun (types), should have thrown another spanner in the works, yet the inselberge seen in Figure 6 forcefully indicate that the system cannot be anything but motivated. This outcome is highly

significant: if with everything against the uncovering of an underlying system, and this moreover for the synchronic study of a single Bantu language rather than Proto-Bantu, one does conclude there is an underlying system, then it becomes worthwhile to start the fine-tuning of the various parameters (+/- loanwords, +/- certain classes or genders, +/- derived nouns, +/- corresponding forms of two-class genders, +/- token counts), in order to make the uncovering a reality. Apart from the extremely high occurrence frequency of classes 1, 2 and 1a nouns (which may indicate that natural language is even more human and anthropomorphic than some assume it already is), the fact that often more than one inselberg may be found along one of the values of either the noun-class axis or the semantic-import axis, may further imply that the semantic import is in those cases actually a composite rather than a single block.

Pursuing this goes beyond the scope of this article, but we hope to report on some of the outcomes in a forthcoming study. One of the reasons for not pursuing this here has to do with the size of the corpus, which needs to be larger for some of the variations to be relevant. For example, and as another type of parameter tuning, one could be interested in knowing the distribution of the semantic categories for the one-class genders 4, 6, 8 and 14, without any interference from (or conflation with) the other genders which include classes 4, 6, 8 and 14 as a corresponding form. The results of this query are shown in Appendix 17.1 through 17.4. For gender 4, for example, and in terms of types, this means that the percentage of true abstracts goes from 32 to 67%. For gender 6, liquids go from 10 to 21%, while human body parts go from 23 to 8%. True abstracts also increase, from 21 to 38%. Gender 8 almost exclusively consists of man-made abstracts now compared to class 8, from 16 to 95%. Gender 14, finally, sees the true abstracts climb from 68 to 76%. While all these 'changes' are in line with expectation, one must keep in mind that most of these counts concern very few noun types only.

Summarizing Section 7 we can therefore say that we feel that a three-dimensional semantic-import view of nouns, with as axes noun classes, semantic categories and corpus frequencies, is not only a novel, but also a most-revealing and promising avenue to decode the underlying semantic system. For the noun in Lusoga, as well as for the noun in any Bantu language.

Acknowledgements

Thanks are due to the two anonymous reviewers who, through their penetrating questions, helped improve this contribution. The usual disclaimers apply.

References

Babyale, S. C. 1999. Gulama w'Olusoga Omukalamu [The Proper Lusoga Grammar] (Unpublished BA dissertation, written in English). Kampala: Makerere University.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Contini-Morava, E. 1994. *Noun Classification in Swahili* (Publications of the

Institute for Advanced Technology in the Humanities, Research Reports, Second Series). Charlottesville: University of Virginia. Available from: http://www2.iath.virginia.edu/swahili/swahili.html.

— 1996. '"Things" in a Noun-Class Language: Semantic Functions of Agreement in Swahili'. In E. Andrews & Y. Tobin (eds), *Toward a Calculus of Meaning: Studies in Markedness, Distinctive Features and Deixis* (Studies in Functional and Structural Linguistics 43), 251-90. Amsterdam: John Benjamins.

— 1997. 'Noun Classification in Swahili: A cognitive semantic analysis using a computer database'. In R. K. Herbert (ed.), *African Linguistics at the Crossroads: Papers from Kwaluseni, 1st World Congress of African Linguistics, Swaziland, 18-22.VII.1994*, 599-628. Cologne: Rüdiger Köppe.

— 2000. 'Noun Class as Number in Swahili'. In E. Contini-Morava & Y. Tobin (eds), *Between Grammar and Lexicon* (Amsterdam Studies in the Theory and History of Linguistic Science, Series IV – Current Issues in Linguistic Theory 183), 3-30. Amsterdam: John Benjamins.

— 2002. '(What) do noun class markers mean?' In W. Reid, R. Otheguy & N. Stern (eds), *Signal, Meaning, and Message: Perspectives on sign-based linguistics* (Studies in Functional and Structural Linguistics 48), 3-64. Amsterdam: John Benjamins.

de Schryver, G.-M. 1999. *Cilubà Phonetics, Proposals for a 'corpus-based phonetics from below'-approach* (Recall Linguistics Series 14). Ghent: Recall.

— 2008. 'A New Way to Lemmatize Adjectives in a User-friendly Zulu – English Dictionary', *Lexikos* 18:63-91.

— & R. Gauton. 2002. 'The Zulu locative prefix ku- revisited: A corpus-based approach', *Southern African Linguistics and Applied Language Studies* 20 (4): 201-20.

— & E. Taljard. 2006. 'Locative trigrams in Northern Sotho, preceded by analyses of formative bigrams', *Linguistics* 44 (1):135-93.

Dimmendaal, G. J. 2001. 'Places and people: field sites and informants'. In P. Newman & M. Ratliff (eds), *Linguistic Fieldwork*, 55-75. Cambridge: Cambridge University Press.

Doke, C. M. 1935. *Bantu Linguistic Terminology*. London: Longmans, Green.

Firth, J. R. 1951 [1957]. 'Modes of Meaning'. In J. R. Firth (ed.), *Papers in Linguistics 1934-1951*, 190-215. London: Oxford University Press.

Gauton, R., G.-M. de Schryver & L. Mohlala. 2004. 'A Corpus-based Investigation of the Zulu Nominal Suffix -kazi: A preliminary study'. In A. Akinlabi & O. Adesola (eds), *Proceedings of the 4th World Congress of African Linguistics, New Brunswick 2003*, 373-80. Cologne: Rüdiger Köppe.

Geeraerts, D. 2002. 'The theoretical and descriptive development of lexical semantics'. In L. Behrens & D. Zaefferer (eds), *The Lexicon in Focus. Competition and Convergence in Current Lexicology*, 23-42. Frankfurt am Main: Peter Lang.

— 2009. 'Currents and undercurrents in lexical semantics, twenty years after'. In E. Beijk *et al*. (eds), *Fons Verborum. Feestbundel voor prof. dr. A.M.F.J. (Fons) Moerdijk, aangeboden door vrienden en collega's bij zijn afscheid van het Instituut voor Nederlandse Lexicologie*, 421-30. Amsterdam: Gopher BV.

— 2010. *Theories of Lexical Semantics*. New York: Oxford University Press.
Hendrikse, A. P. 1990. 'Number as a categorizing parameter in Southern Bantu: An exploration in cognitive grammar', *South African Journal of African Languages* 10 (4):384-400.
— & G. Poulos. 1992. 'A continuum interpretation of the Bantu noun class system'. In D. F. Gowlett (ed.), *African linguistic contributions: presented in honour of Ernst Westphal*, 195-209. Hatfield: Via Afrika.
Himmelmann, N. P. 1998. 'Documentary and descriptive linguistics'. *Linguistics* 36 (1):161-95.
— 2006. 'Language documentation: What is it and what is it good for?' In J. Gippert, N. P. Himmelmann & U. Mosel (eds), *Essentials of Language Documentation* (Trends in Linguistics, Studies and Monographs 178), 1-30. Berlin: Mouton de Gruyter.
Hopper, P. J. & S. A.Thompson. 1984. 'The Discourse Basis for Lexical Categories in Universal Grammar', *Language* 60 (4):703-52.
Hyman, L. M. & F. X. Katamba. 1999. 'The syllable in Luganda phonology and morphology'. In H. van der Hulst & N. A. Ritter (eds), *The Syllable: Views and Facts* (Studies in Generative Grammar 45), 349-416. Berlin: Mouton de Gruyter.
Kadima, M. 1969. *Le système des classes en bantou* (PhD thesis). Leuven: Vander.
Kajolya, J. B. N. 1990. *The Lusoga Orthography*. Jinja: Lusoga Ecumenical Committee.
Korse, P. 1999a. *A Lusoga Grammar*. Jinja: Cultural Research Centre.
— 1999b. *Dictionary Lusoga-English / English-Lusoga*. Jinja: Cultural Research Centre.
Louwrens, L. J. 1992. 'The conceptualisation of spatial relationships as expressed by locative structures', *South African Journal of African Languages* 12 (3): 107-11.
LULANDA & CRC. 2004. *Empandiika y'Olulimi Olusoga Enkalamu / Standard Lusoga Orthography*. Jinja: Lusoga Language Authority.
Lüpke, F. 2005a. *A grammar of Jalonke argument structure* (MPI Series in Psycholinguistics 30; PhD thesis). Nijmegen: Radboud University Nijmegen.
— 2005b. 'Small is beautiful: contributions of field-based corpora to different linguistic disciplines, illustrated by Jalonke'. In P. K. Austin (ed.), *Language Documentation and Description, Volume 3*, 75-105. London: SOAS.
— 2009. 'Data collection methods for field-based language documentation'. In P. K. Austin (ed.), *Language Documentation and Description, Volume 6*, 53-100. London: SOAS.
Maho, J. F. 1999. *A Comparative Study of Bantu Noun Classes* (Orientalia et Africana Gothoburgensia 13; PhD thesis). Gothenburg: Acta Universitatis Gothoburgensis.
Matovu, C. N. 1992. A synchronic description of Lusoga in terms of its relatedness to Luganda (PhD thesis). Kampala: Makerere University.
Matovu, K. B. & L. Walusimbi. 2000. A linguistic survey of the current status of the dialects of some eastern Bantu languages (Unpublished manuscript). Kampala: Makerere University.

Mc Laughlin, F. & T. S. Sall. 2001. 'The give and take of fieldwork: noun classes and other concerns in Fatick, Senegal'. In P. Newman & M. Ratliff (eds), *Linguistic Fieldwork*, 189-210. Cambridge: Cambridge University Press.

Mithun, M. 2001. 'Who shapes the record: the speaker and the linguist'. In P. Newman & M. Ratliff (eds), *Linguistic Fieldwork*, 34-54. Cambridge: Cambridge University Press.

Nabirye, M. 2008. *Compilation of the Monolingual Lusoga Dictionary* (MA dissertation). Kampala: Makerere University.

— 2009a. *Eiwanika ly'Olusoga. Eiwanika ly'aboogezi b'Olusoga n'abo abenda okwega Olusoga* [A Dictionary of Lusoga. For speakers of Lusoga, and for those who would like to learn Lusoga]. Kampala: Menha Publishers.

— 2009b. 'Compiling the First Monolingual Lusoga Dictionary', *Lexikos* 19:177-96.

— 2009c. 'Dictionary Compilation for Mother-tongue Speakers of Bantu Languages'. In R. Zhu (ed.), *Proceedings of the International Seminar on Kangxi Dictionary & Lexicology*, 597-607. Beijing: Beijing Normal University.

Newman, P. & M. Ratliff. (eds) 2001. *Linguistic Fieldwork*. Cambridge: Cambridge University Press.

Scannell, K. P. 2007. Corpus building for minority languages. Available from: http://borel.slu.edu/crubadan/.

Schoenbrun, D. L. 1997. *The Historical Reconstruction of Great Lakes Bantu Cultural Vocabulary: Etymologies and Distributions* (Sprache und Geschichte in Afrika, Supplement 9). Cologne: Rüdiger Köppe.

Selvik, K.-A. 2001. 'When a Dance Resembles a Tree: A Polysemy Analysis of Three Setswana Noun Classes'. In H. Cuyckens & B. Zawada (eds), *Polysemy in Cognitive Linguistics*, 161-84. Amsterdam: John Benjamins.

Sinclair, J. M. 1966. 'Beginning the Study of Lexis'. In C. E. Bazell *et al*. (eds), *In Memory of J.R. Firth*, 410-30. London: Longman.

Steeman, S. 2001. Kintu: an annotated edition of a Lusoga play (MA dissertation). Leiden: Leiden University.

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):165-83.

Tucker, A. N. & M. A. Bryan. 1957. *Linguistic Survey of the Northern Bantu Borderland, Volume 4: Languages of the Eastern Section, Great Lakes to Indian Ocean* (Published for the International African Institute). London: Oxford University Press.

UBS. 2006. *The 2002 Uganda Population and Housing Census, Analytical Report, Population Composition*. Kampala: Uganda Bureau of Statistics.

Van der Wal, J. 2004. Lusoga phonology (MA dissertation). Leiden: Leiden University.

Welmers, W. E. 1973. *African Language Structures*. Berkeley and Los Angeles: University of California Press.

湯川, 恭敏 [Yukawa, Yasutoshi]. 2000. 'ソガ語動詞アクセント試論 [Soga-go dooshi akusento shiron / A tentative tonal analysis of Soga verbs]', *Journal of Asian and African Studies* 60:249-90.

Authors' addresses

Gilles-Maurice de Schryver
Dept of African Languages & Cultures
Ghent University
Rozier 44
B-9000 Ghent
BELGIUM
gillesmaurice.deschryver@UGent.be

Xhosa Dept
University of the Western Cape
Modderdam Road
7535 Bellville South
SOUTH AFRICA

Minah Nabirye
Dept of African Languages & Cultures
Ghent University
Rozier 44
B-9000 Ghent
BELGIUM
mnabirye@gmail.com

Institute of Languages
Makerere University
Kampala
UGANDA

Résumé

Dans cet article, nous montrons comment l'analyse distributionnelle d'un corpus peut être utilisée pour aborder la description d'une langue, pour ainsi dire non documentée. Cette approche est illustrée avec le lusoga (JE16), une langue bantu de la région des Grands Lacs, parlée dans la ville de Jinja (Ouganda) et dans ses environs. L'étude porte sur le nominal en lusoga, en accordant une attention particulière à trois niveaux d'analyse: morphologique, morphophonologique et sémantique.

Dans une première partie, nous montrons que, pour chaque classe nominale, une distribution relative du nombre de types et d'occurrences combinée à un système de classes nominales pondéré à deux dimensions constitue un outil très puisssant pour visualiser la force de chaque noeud et de chaque lien dans la structure. Dans une seconde partie, nous indiquons comment la combinaison d'une énumération quantifiée de la morphophonologie nominale et des constructions nominales avec significations associées fournit une image représentative des divers aspects de construction nominale. Enfin, dans une troisième et dernière partie, nous plaiderons en faveur d'une conception tridimensionnelle d'importation sémantique des noms, avec pour axes les classes nominales, les catégories sémantiques et les fréquences d'apparition dans le corpus. C'est là, non seulement une nouveauté, mais également une voie très révélatrice et prometteuse pour décoder le système sémantique sous-jacent des nominaux en lusoga ou dans toute autre langue bantu.

Appendix 1.1: Class 1 – Morphology.

Gender	N	%	Gender	N	%
1 (omu-) / 2 (aba-)	101	100.00	**1 (omu-)**	6	100.00
1 (mu-) / 2 (ba-)	41		**1 (mu-)**	1	
N	142			7	
%	95.30			4.70	

Appendix 1.2: Class 1 – Morphophonology.

Gender	Rule	N	Gender	Rule	N
1 (omu-) / 2 (aba-)	**u + a > wa**	11	**1 (omu-)**	**u + a > wa**	1
..	**u + e > we**	4			
..	**u + i > wi**	4			
..	**u + o > wo**	1			
1 (omu-) / 2 (aba-)	**a + e > e/_NC**	2			
..	**a + e > ee**	2			
..	**a + o > oo**	1			
1 (mu-) / 2 (ba-)	**u + a > wa**	3			
..	**u + e > we**	2			
..	**u + i > wi**	2			
1 (mu-) / 2 (ba-)	**a + e > ee**	2			

Appendix 1.3: Class 1 – Semantic import.

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	1.5	1.01	804	6.36
Fauna (animals)	0	0.00	0	0.00
Flora (plants)	0	0.00	0	0.00
Human body parts	0	0.00	0	0.00
Liquids	0	0.00	0	0.00
Man-made abstracts	0	0.00	0	0.00
Man-made concretes	0	0.00	0	0.00
Nature	0	0.00	0	0.00
People	147.5	98.99	11,829	93.64
Others	0	0.00	0	0.00
N / Freq.	149	100.00	12,633	100.00

Appendix 2.1: Class 2 – Morphology.

Gender	N	%
1 (omu-) / 2 (aba-)	120	100.00
1 (mu-) / 2 (ba-)	35	
N	155	
%	100.00	

Appendix 2.2: Class 2 – Morphophonology.

Gender	Rule	N
1 (omu-) / 2 (aba-)	**u + i > wi**	10
..	**u + a > wa**	7
..	**u + e > we**	4
..	**u + o > wo**	1
1 (omu-) / 2 (aba-)	**a + e > ee**	3
..	**a + e > e/_NC**	1
..	**a + o > oo**	1
1 (mu-) / 2 (ba-)	**u + a > wa**	1
..	**u + i > wi**	1

Appendix 2.3: Class 2 – Semantic import.

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	2	1.29	30	0.31
Fauna (animals)	0	0.00	0	0.00
Flora (plants)	0	0.00	0	0.00
Human body parts	0	0.00	0	0.00
Liquids	0	0.00	0	0.00
Man-made abstracts	0	0.00	0	0.00
Man-made concretes	0	0.00	0	0.00
Nature	0	0.00	0	0.00
People	153	98.71	9,782	99.69
Others	0	0.00	0	0.00
N / Freq.	155	100.00	9,812	100.00

Appendix 3.1: Class 1a – Morphology.

Gender	N	%	Gender	N	%
1a (ø-)	169	100.00	**1a (ø-) / 2a (ba-)**	32	100.00
			1a (ø-) / 2a (ø-) OR **2a (ba-)**	4	
N	169			36	
%	82.44			17.56	

Appendix 3.2: Class 1a – Morphophonology.

(*no sound changes*)

Appendix 3.3: Class 1a – Semantic import.

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	10	4.9	3,771	30.7
Fauna (animals)	0	0.0	0	0.0
Flora (plants)	7	3.4	256	2.1
Human body parts	1	0.5	43	0.3
Liquids	2	1.0	56	0.5
Man-made abstracts	9	4.4	318	2.6
Man-made concretes	4	2.0	101	0.8
Nature	20	9.8	597	4.9
People	151	73.7	7,133	58.0
Others	1	0.5	20	0.2
N / Freq.	205	100.0	12,295	100.0

Appendix 4.1: Class 2a – Morphology.

Gender	N	%
1a (ø-) / 2a (ba-)	8	100.00
N	8	
%	100.00	

Appendix 4.2: Class 2a – Morphophonology.

(*no sound changes*)

Appendix 4.3: Class 2a – Semantic import.

Semantic import	**N**	**%**	**Freq.**	**%**
Abstracts (non-temporal)	2	25.00	106	24.31
Fauna (animals)	0	0.00	0	0.00
Flora (plants)	0	0.00	0	0.00
Human body parts	0	0.00	0	0.00
Liquids	0	0.00	0	0.00
Man-made abstracts	0	0.00	0	0.00
Man-made concretes	0	0.00	0	0.00
Nature	0	0.00	0	0.00
People	6	75.00	330	75.69
Others	0	0.00	0	0.00
N / Freq.	8	100.00	436	100.00

Appendix 5.1: Class 3 – Morphology.

Gender	**N**	**%**	**Gender**	**N**	**%**
3 (omu-) / 4 (emi-)	80	100.00	**3 (omu-)**	29	100.00
3 (mu-) / 4 (mi-)	49		**3 (mu-)**	13	
N	129			42	
%	75.44			24.56	

Appendix 5.2: Class 3 – Morphophonology.

Gender	**Rule**	**N**	**Gender**	**Rule**	**N**
3 (omu-) / 4 (emi-)	**u + e > we**	3	**3 (omu-)**	**u + a > wa**	1
..	**u + i > wi**	3	**..**	**u + e > we**	1
..	**u + o > wo**	2	**..**	**u + o > wo**	1
..	**u + a > wa**	1			
3 (omu-) / 4 (emi-)	**i + e > ye**	3			
..	**i + o > yo**	2			
..	**i + a > ya**	1			
3 (mu-) / 4 (mi-)	**u + a > wa**	2	**3 (mu-)**	**u + e > we**	1

..	**u + e > we**	2
..	**u + i > wi**	1
..	**u + o > wo**	1
3 (mu-) / 4 (mi-)	**i + a > ya**	2
..	**i + e > ye**	2
..	**i + o > yo**	1

Appendix 5.3: Class 3 – Semantic import.

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	31	18.13	1,264	16.92
Fauna (animals)	6	3.51	103	1.38
Flora (plants)	8	4.68	253	3.39
Human body parts	26	15.20	1,784	23.88
Liquids	10	5.85	457	6.12
Man-made abstracts	14.5	8.48	668	8.94
Man-made concretes	38	22.22	744	9.96
Nature	22	12.87	1,067	14.28
People	3.5	2.05	127	1.70
Others	12	7.02	1,005	13.45
N / Freq.	171	100.00	7,472	100.00

Appendix 6.1: Class 4 – Morphology.

Gender	N	%	Gender	N	%
3 (omu-) / 4 (emi-)	42	100.00	**4 (emi-)**	8	100.00
3 (mu-) / 4 (mi-)	22		**4 (mi-)**	1	
N	64			9	
%	87.67			12.33	

Appendix 6.2: Class 4 – Morphophonology.

Gender	Rule	N
3 (omu-) / 4 (emi-)	**u + a > wa**	2
..	**u + o > wo**	2
..	**u + e > we**	1

..	**u + i > wi**	1
3 (omu-) / 4 (emi-)	**i + a > ya**	2
..	**i + o > yo**	2
..	**i + e > ye**	1
3 (mu-) / 4 (mi-)	**u + a > wa**	1
..	**u + e > we**	1
..	**u + i > wi**	1
..	**u + o > wo**	1
3 (mu-) / 4 (mi-)	**i + a > ya**	1
..	**i + e > ye**	1
..	**i + o > yo**	1

Appendix 6.3: Class 4 – Semantic import.

Semantic import	**N**	**%**	**Freq.**	**%**
Abstracts (non-temporal)	23	31.51	715	29.72
Fauna (animals)	1	1.37	10	0.42
Flora (plants)	8.5	11.64	207	8.60
Human body parts	9.5	13.01	267	11.10
Liquids	0	0.00	0	0.00
Man-made abstracts	2	2.74	27	1.12
Man-made concretes	6.5	8.90	139	5.78
Nature	9	12.33	579	24.06
People	2.5	3.42	37	1.54
Others	11	15.07	425	17.66
N / Freq.	73	100.00	2,406	100.00

Appendix 7.1: Class 5 – Morphology.

Gender	**N**	**%**	**Gender**	**N**	**%**
5 (ei-) / 6 (ama-)	36	85.33	**5 (ei-)**	24	100.00
5 (i-) / 6 (ma-)	28		**5 (i-)**	21	
5 (eli-) / 6 (ama-)	8	14.67			
5 (li-) / 6 (ma-)	3				
N	75			45	
%	62.50			37.50	

Appendix 7.2: Class 5 – Morphophonology.

Gender	Rule	N
5 (eli-) / 6 (ama-)	**i + a > ya**	2
5 (li-) / 6 (ma-)	**i + a > ya**	1

Appendix 7.3: Class 5 – Semantic import.

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	26	21.67	865	16.92
Fauna (animals)	3	2.50	38	0.74
Flora (plants)	6	5.00	209	4.09
Human body parts	7	5.83	174	3.40
Liquids	4	3.33	115	2.25
Man-made abstracts	6	5.00	431	8.43
Man-made concretes	35	29.17	1,695	33.16
Nature	10	8.33	887	17.35
People	4	3.33	82	1.60
Others	19	15.83	615	12.03
N / Freq.	120	100.00	5,111	100.00

Appendix 8.1: Class 6 – Morphology.

Gender	N	%	Gender	N	Gender	N	Gender	N
5 (ei-) / 6 (ama-)	36	68.29	**6 (ama-)**	27	**15 (oku-) /**	3	**9 (en-) /**	1
5 (i-) / 6 (ma-)	20		**6 (ma-)**	12	**6 (ama-)**		**6 (ama-)**	
5 (eli-) / 6 (ama-)	18	31.71			**15 (ku-) /**	4	**9 (n-) /**	1
5 (li-) / 6 (ma-)	8				**6 (ma-)**		**6 (ma-)**	
N	82			39		7		2
%	63.08			30.00		5.38		1.54

Appendix 8.2: Class 6 – Morphophonology.

Gender	Rule	N	Gender	Rule	N	Gender	Rule	N
5 (eli-) / 6 (ama-)	**i+a>ya**	8	**6 (ama-)**	**a+y>oo**	1	**9 (en-) / 6 (ama-)**	**N+y>nnh/_N**	1
..	**i+a>ii/_D**	1				**9 (n-) / 6 (ma-)**	**N+y>nnh/_N**	1
..	**i+y>y**	1						
5 (eli-) / 6 (ama-)	**a+y>e/_NC**	1						
5 (li-) / 6 (ma-)	**i+a>ii/_D**	1						
..	**i+y>y**	1						
5 (li-) / 6 (ma-)	**a+y>e/_NC**	1						

Appendix 8.3: Class 6 – Semantic import.

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	27	20.77	1,518	25.00
Fauna (animals)	4	3.08	58	0.96
Flora (plants)	10	7.69	272	4.48
Human body parts	29.5	22.69	1,332	21.93
Liquids	13.5	10.38	1,064	17.52
Man-made abstracts	4	3.08	115	1.89
Man-made concretes	24	18.46	1,157	19.05
Nature	2	1.54	85	1.40
People	2	1.54	80	1.32
Others	14	10.77	392	6.45
N / Freq.	130	100.00	6,073	100.00

Appendix 9.1: Class 7 – Morphology.

Gender	N	%	Gender	N	%
7 (eki-) / 8 (ebi-)	107	100.00	**7 (eki-)**	12	100.00
7 (ki-) / 8 (bi-)	71		**7 (ki-)**	11	
N	178			23	
%	88.56			11.44	

Appendix 9.2: Class 7 – Morphophonology.

Gender	Rule	N	Gender	Rule	N
7 (eki-) / 8 (ebi-)	**i + a > ya**	5	**7 (eki-)**	**i + e > ye**	1
..	**i + o > yo**	3			
..	**i + e > ye**	1			
..	**i + u > yu**	1			
7 (eki-) / 8 (ebi-)	**i + a > ya**	5			
..	**i + o > yo**	3			
..	**i + e > ye**	1			
..	**i + u > yu**	1			
7 (ki-) / 8 (bi-)	**i + a > ya**	5			
..	**i + e > ye**	2			
..	**i + u > yu**	1			

7 (ki-) / 8 (bi-)	**i + a > ya**	5
..	**i + e > ye**	2
..	**i + u > yu**	1

Appendix 9.3: Class 7 – Semantic import.

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	52.5	26.12	2,738	22.68
Fauna (animals)	9.5	4.73	302	2.50
Flora (plants)	15	7.46	315	2.61
Human body parts	6	2.99	116	0.96
Liquids	0	0.00	0	0.00
Man-made abstracts	3.5	1.74	80	0.66
Man-made concretes	59.5	29.60	1,780	14.74
Nature	7	3.48	146	1.21
People	6	2.99	243	2.01
Others	42	20.90	6,352	52.62
N / Freq.	201	100.00	12,072	100.00

Appendix 10.1: Class 8 – Morphology.

Gender	N	%	Gender	N	%
7 (eki-) / 8 (ebi-)	84	100.00	**8 (ebi-)**	18	100.00
7 (ki-) / 8 (bi-)	42		**8 (bi-)**	2	
N	126			20	
%	86.30			13.70	

Appendix 10.2: Class 8 – Morphophonology.

Gender	Rule	N	Gender	Rule	N
7 (eki-) / 8 (ebi-)	**i + e > ye**	10	**8 (ebi-)**	**i + o > yo**	8
..	**i + a > ya**	6	**..**	**i + e > ye**	6
..	**i + o > yo**	2	**..**	**i + a > ya**	3
..	**i + u > yu**	1			
7 (eki-) / 8 (ebi-)	**i + e > ye**	10			
..	**i + a > ya**	6			
..	**i + o > yo**	2			

..	i + u > yu	1			
7 (ki-) / 8 (bi-)	i + a > ya	4	8 (bi-)	i + a > ya	1
..	i + u > yu	1	..	i + e > ye	1
7 (ki-) / 8 (bi-)	i + a > ya	4			
..	i + u > yu	1			

Appendix 10.3: Class 8 – Semantic import.

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	36	24.66	1,418	21.33
Fauna (animals)	9	6.16	232	3.49
Flora (plants)	14	9.59	373	5.61
Human body parts	6	4.11	263	3.96
Liquids	1	0.68	17	0.26
Man-made abstracts	23	15.75	981	14.76
Man-made concretes	30	20.55	1,861	28.00
Nature	0	0.00	0	0.00
People	1	0.68	23	0.35
Others	26	17.81	1,479	22.25
N / Freq.	146	100.00	6,647	100.00

Appendix 11.1: Class 9 – Morphology.

Gender	N	%	Gender	N	%	Gender	N	%
9 (eN-) / 10 (eN-)	99	82.54	9 (eN-)	71	70.00	9 (eN-) / 6 (ama-)	1	6.25
9 (N-) / 10 (N-)	57		9 (N-)	55				
9 (e-) / 10 (e-)	4	17.46	9 (e-)	4	30.00	9 (e-) / 6 (ama-)	4	93.75
9 (ø-) / 10 (ø-)	29		9 (ø-)	50		9 (ø-) / 6 (ma-)	11	
N	189			180			16	
%	49.09			46.75			4.16	

Appendix 11.2: Class 9 – Morphophonology.

Gender	**9 (eN-) / 10 (eN-)**	**9 (eN-) / 10 (eN-)**	**9 (N-) / 10 (N-)**	**9 (N-) / 10 (N-)**	**9 (eN-)**	**9 (N-)**
Rule	**N**	**N**	**N**	**N**	**N**	**N**
N + b > mm/_N	2	2	0	0	2	2
N + b > mb	14	14	10	10	2	3
N + g > ŋg/_N	1	1	1	1	1	0
N + l > nn/_N	1	1	3	3	2	1
N + l > nd	8	8	1	1	5	3
N + m > mm	5	5	1	1	4	4
N + w > mp	9	9	3	3	8	6
N + y > nnh/_N	5	5	6	6	4	2
N + y > mp/_i	1	1	1	1	1	2
N + y > ndh	5	5	2	2	3	6

Appendix 11.3: Class 9 – Semantic import.

Semantic import	**N**	**%**	**Freq.**	**%**
Abstracts (non-temporal)	100	25.97	4,694	31.01
Fauna (animals)	49	12.73	1,712	11.31
Flora (plants)	17	4.42	636	4.20
Human body parts	13	3.38	471	3.11
Liquids	0.5	0.13	17	0.11
Man-made abstracts	14	3.64	563	3.72
Man-made concretes	134.5	34.94	4,485	29.63
Nature	22	5.71	688	4.55
People	3	0.78	93	0.61
Others	32	8.31	1,777	11.74
N / Freq.	385	100.00	15,136	100.00

Appendix 12.1: Class 10 – Morphology.

Gender	**N**	**Gender**	**N**	**%**	**Gender**	**N**
11 (olu-) / 10 (eN-)	35	**9 (eN-) / 10 (eN-)**	22	78.38	**14 (obu-) / 10 (eN-)**	2
11 (lu-) / 10 (N-)	17	**9 (N-) / 10 (N-)**	7			

		9 (e-) / 10 (e-)	1	21.62
		9 (ø-) / 10 (ø-)	7	
N	52		37	2
%	57.14		40.66	2.20

Appendix 12.2: Class 10 – Morphophonology.

Gender	Rule	N	Gender	Rule	N
11 (olu-) / 10 (eN-)	**u + y > we/_NC**	3	**9 (eN-) / 10 (eN-)**	**N + b > mb**	2
11 (olu-) / 10 (eN-)	**N + b > mb**	3	**..**	**N + l > nd**	1
..	**N + l > nn/_N**	2	**..**	**N + m > mm**	1
..	**N + l > nd**	1	**..**	**N + w > mp**	1
..	**N + p > mp**	5	**..**	**N + y > mp/_i**	1
..	**N + w > mp**	1	**..**	**N + y > nnh/_N**	1
..	**N + y > nnh/_N**	3	**9 (eN-) / 10 (eN-)**	**N + b > mb**	2
11 (lu-) / 10 (N-)	**N + b > mb**	1	**..**	**N + l > nd**	1
..	**N + l > nn/_N**	1	**..**	**N + m > mm**	1
..	**N + l > nd**	2	**..**	**N + w > mp**	1
..	**N + p > mp**	1	**..**	**N + y > mp/_i**	1
..	**N + w > mp**	2	**..**	**N + y > nnh/_N**	1
..	**N + y > mp/_i**	1	**9 (N-) / 10 (N-)**	**N + b > mb**	1
			..	**N + y > mp/_i**	1
Gender	**Rule**	**N**	**9 (N-) / 10 (N-)**	**N + b > mb**	1
14 (obu-) / 10 (eN-)	**N + l > nd**	2	**..**	**N + y > mp/_i**	1

Appendix 12.3: Class 10 – Semantic import.

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	25.5	28.02	745	27.54
Fauna (animals)	3.5	3.85	45	1.66
Flora (plants)	5.5	6.04	105	3.88
Human body parts	11.5	12.64	409	15.12
Liquids	0	0.00	0	0.00
Man-made abstracts	8.5	9.34	220	8.13

Man-made concretes	25	27.47	787	29.09
Nature	3.5	3.85	237	8.76
People	0	0.00	0	0.00
Others	8	8.79	157	5.80
N / Freq.	91	100.00	2,705	100.00

Appendix 13.1: Class 11 – Morphology.

Gender	**N**	**%**	**Gender**	**N**	**%**
11 (olu-) / 10 (eN-)	44	100.00	**11 (olu-)**	10	100.00
11 (lu-) / 10 (N-)	31		**11 (lu-)**	14	
N	75			24	
%	75.76			24.24	

Appendix 13.2: Class 11 – Morphophonology.

Gender	Rule	N	Gender	Rule	N	Gender	Rule	N
11 (olu-) / 10 (eN-)	u + y > we/_NC	3	11 (lu-) / 10 (N-)	u + y > we/_NC	2	11 (olu-)	u + e > we	2
..	u + y > wi/_i	1	..	u + y > wi/_i	1	..	u + o > wo	2
..	u + y > wa	1				..	u + a > wa	1
11 (olu-) / 10 (eN-)	N + b > mm/_N	1	11 (lu-) / 10 (N-)	N + b > mm/_N	1	11 (lu-)	u + a > wa	3
..	N + b > mb	3	..	N + b > mb	3	..	u + e > we	3
..	N + g > ŋŋ/_N	2	..	N + g > ŋŋ/_N	2	..	u + i > wi	1
			..	N + l > nn/_N	1			
..	N + l > nd	3	..	N + l > nd	3			
..	N + p > mp	1	..	N + p > mp	1			
..	N + w > mp	2	..	N + w > mp	1			
..	N + y > nnh/_N	3	..	N + y > nnh/_N	2			

Appendix 13.3: Class 11 – Semantic import.

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	30	30.30	1,077	21.43
Fauna (animals)	0	0.00	0	0.00
Flora (plants)	2	2.02	27	0.54
Human body parts	4	4.04	177	3.52
Liquids	0	0.00	0	0.00
Man-made abstracts	14	14.14	1,643	32.70
Man-made concretes	21	21.21	780	15.52
Nature	13	13.13	911	18.13
People	0	0.00	0	0.00
Others	15	15.15	410	8.16
N / Freq.	99	100.00	5,025	100.00

Appendix 14.1: Class 12 – Morphology.

Gender	N	%	Gender	N	%
12 (aka-) / 14 (obu-)	27	100.00	**12 (aka-)**	7	100.00
12 (ka-) / 14 (bu-)	19		**12 (ka-)**	8	
N	46			15	
%	75.41			24.59	

Appendix 14.2: Class 12 – Morphophonology.

Gender	Rule	N
12 (aka-) / 14 (obu-)	**u + a > wa**	2
..	**u + e > we**	1

Appendix 14.3: Class 12 – Semantic import.

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	7	11.48	143	8.64
Fauna (animals)	4	6.56	49	2.96
Flora (plants)	2	3.28	57	3.44
Human body parts	6.5	10.66	152	9.18
Liquids	0	0.00	0	0.00
Man-made abstracts	3	4.92	138	8.34
Man-made concretes	22.5	36.89	460	27.79

Nature	3	4.92	35	2.11
People	4	6.56	66	3.99
Others	9	14.75	555	33.53
N / Freq.	61	100.00	1,655	100.00

Appendix 15.1: Class 14 – Morphology.

Gender	**N**	**%**	**Gender**	**N**	**%**
14 (obu-)	110	100.00	**12 (aka-) / 14 (obu-)**	22	100.00
14 (bu-)	45		**12 (ka-) / 14 (bu-)**	1	
N	155			23	
%	87.08			12.92	

Appendix 15.2: Class 14 – Morphophonology.

Gender	**Rule**	**N**	**Gender**	**Rule**	**N**
14 (obu-)	**u + i > wi**	10	**12 (aka-) / 14 (obu-)**	**u + a > wa**	3
..	**u + a > wa**	6	**12 (ka-) / 14 (bu-)**	**u + a > wa**	1
..	**u + e > we**	6			
..	**u + o > wo**	2			
14 (bu-)	**u + e > we**	3			
..	**u + i > wi**	2			
..	**u + o > wo**	1			

Appendix 15.3: Class 14 – Semantic import.

Semantic import	**N**	**%**	**Freq.**	**%**
Abstracts (non-temporal)	121	67.98	4,187	70.86
Fauna (animals)	7	3.93	110	1.86
Flora (plants)	9	5.06	176	2.98
Human body parts	3	1.69	70	1.18
Liquids	2	1.12	26	0.44
Man-made abstracts	5	2.81	172	2.91
Man-made concretes	12	6.74	391	6.62
Nature	11	6.18	663	11.05
People	1	0.56	17	0.29
Others	7	3.93	107	1.81
N / Freq.	178	100.00	5,909	100.00

Appendix 16: Noun constructions and linked meanings in the Lusoga corpus.

Gender	Construction	N	%	Meaning	Example
1/2, 1	NP + noun root	173	56.91		
	NP + V + i	53	17.43	person who 'verbs'	**-iba** 'steal' > **omwibi** 'thief'
	NP + V + a	35	11.51	person who 'verbs'	**-somesa** 'teach' > **omusomesa** 'teacher'
	NP + V + perfective form	28	9.21	person who is/has 'verbed'	**-siba** 'tie; lock' > **omusibe** 'prisoner'
	NP + V + u	8	2.63	person who 'verbs'	**-tamiila** 'get drunk' > **omutamiivu** 'drunkard'
	NP + V + o	5	1.64	person in the institution of 'verbing'	**-fumba** 'cook' > **omufumbo** 'married person'
	NP + na + N-pp	2	0.66	person of the 'noun'	**amateeka** 'laws' > **abanamateeka** 'lawyers'
	SUM	304	100.00		
1a/2a	NP + noun root	24	54.55		
	NP + ka + V + a	4	9.09	person who 'verbs'	**-bona** 'see' > **kabona** 'pastor'
	NP + ise + N-pp	2	4.55	male who heads the 'noun'	**entebe** 'chair' > **isentebe** 'chairman'
	NP + ise + Pronoun	2	4.55	male who heads or belongs to 'pronoun'	**-bo** 'them' > **isebo** 'sir'
	NP + na + N-pp	2	4.55	female whose 'noun' is lost	**omwandu** 'richness; dowry' > **banamwandu** 'widows'
	NP + PC7 + NP2-pp + V + a	2	4.55	person who represents the 'verbed'	**-zinga** 'encircle' > **Kyabazinga** 'King of Busoga'
	Other	8	18.18		
	SUM	44	100.00		

1a	NP + noun root	88	52.07		
	NP + wa + N-pp	24	14.20	person who is 'noun'	**endhovu** 'elephant' > **Wandhovu** 'Mr/Ms Elephant'
	NP + na + N-pp	17	10.06	time, state or manner of the 'noun'	**engeli** 'way' > **nangeli** 'adverb'
	NP + ki + V + a	6	3.55	person who 'verbs'	**-lunda** 'herd' > **Kilunda** 'Herdsman' (proper name)
	NP + NP1-pp + V + i	3	1.78	person who 'verbs'	**-sika** 'pull' > **Musisi** 'Earthquake' (proper name)
	NP + inhe + plural N-pp	2	1.18	wife of the person who represents the 'plural noun'	**abantu** 'people' > **Inhebantu** 'Queen of Busoga'
	NP + ise + Pronoun	2	1.18	male who heads or belongs to 'pronoun'	**-ife** 'us' > **iseife** 'our father'
	Other	27	15.98		
	SUM	169	100.00		
3/4, 3, 4	NP + noun root	190	77.87		
	NP + V + o	46	18.85	the result of 'verbing'	**-zaanha** 'play; act' > **omuzaanho** 'play; show; game; sport'
	NP + V + a	4	1.64	that which 'verbs'	**-nhwa** 'drink' > **omunhwa** 'mouth'
	NP + V + i	4	1.64	the base / basis of 'verbing'	**-kula** 'grow' > **omukuzi** 'time'
	SUM	244	100.00		
5/6, 5	NP + noun root	171	84.65		
	NP + V + o	27	13.37	place of 'verbing'	**-somela** 'study' > **eisomelo** 'school'
	NP + Number	2	0.99	'number' used definitely	**kumi** 'ten' (neutral use) > **eikumi** 'ten' (definite use)
	Other	2	0.99		
	SUM	202	100.00		

6	NP + noun root	34	87.18		
	NP + V + a	2	5.13	that which is 'verbed'	**-kina** 'dance' > **amakina** 'dance'
	NP + V + o	2	5.13	the result of 'verbing'	**-koba** 'say' > **amakobo** 'conversation'
	Other	1	2.56		
	SUM	39	100.00		
7/8, 7, 8	NP + noun root	179	54.74		
	NP + V + o	73	22.32	the result / occasion of 'verbing'	**-wandiika** 'write' > **ekiwandiiko** 'document; report'
	NP + V + a	36	11.01	that which 'verbs'	**-bala** 'grow well' > **ebibala** 'fruits'
	NP + V + perfective form	23	7.03	that which is 'verbed'	**-tonda** 'create' > **ekitonde** 'creation'
	NP + N	6	1.83	belonging to the 'noun'	**ennhandha** 'lake' > **ebyennhandha** 'fish'
	NP + V + i	5	1.53	the state of 'verbing'	**-faanana** 'look like' > **ekifaanani** 'picture'
	NP + Number	3	0.92	'number' used definitely	**kumi** 'ten' (neutral use) > **ebikumi** 'hundreds' (definite use)
	NP + NP15 + V + enclitic '-ku'	2	0.61	representation of 'verb'	**-bonela** 'see from' > **ekyokubonelaku** 'example'
	SUM	327	100.00		
8	NP + noun root	5	25.00		
	NP + N	15	75.00	pertaining to the institution of the 'noun'	**obufuzi** 'leadership' > **ebyobufuzi** 'pertaining to politics' [with **obufuzi** < **-fuga** 'lead']
	SUM	20	100.00		

9/10, 9	NP + noun root	324	79.80		
	NP + V + a	41	10.10	the state of 'verbing'	**-tegeka** 'prepare; organize' > **entegeka** 'preparation; organization'
	NP + V + o	38	9.36	the result of 'verbing'	**-tegeela** 'understand' > **entegeelo** 'sense'
	NP + V + i	3	0.74	that which 'verbs'	**-bula** 'get lost' > **embuzi** 'goat'
	SUM	406	100.00		
9/6	NP + noun root	18	100.00		
11/10,	NP + noun root	107	70.86		
11	NP + V + o	29	19.21	the result of 'verbing'	**-kiika** 'meet' > **olukiiko** 'meeting'
	NP + V + a	8	5.30	that which is 'verbed'	**-emba** 'sing' > **olwemba** 'song'
	NP + V + perfective form	2	1.32	that which 'verbs'	**-kwa** 'woo, court' > **olukwe** 'cunning plan'
	NP + Number	2	1.32	'number' used definitely	**kumi** 'ten' (neutral use) > **enkumi** 'thousands' (definite use)
	Other	3	1.99		
	SUM	151	100.00		
14/10	NP + V + perfective form	2	100.00	the result of 'verbing'	**-lwala** 'get sick' > **endwaile** 'diseases'
12/14,	NP + noun root	57	67.86		
12	NP + V + o	19	22.62	that which is 'verbed'	**-wanga** 'join' > **akawango** 'affix'
	NP + V + a	5	5.95	that which 'verbs'	**-nhwa** 'drink' > **kanhwa** 'the inside of the mouth'
	NP + V + perfective form	3	3.57	place where something is 'verbed'	**-tala** 'trade' > **akatale** 'market'
	SUM	84	100.00		

14	NP + noun root	75	48.39		
	NP + V + i	23	14.84	the state of 'verbing'	**-sobola** 'be able' > **obusobozi** 'ability'
	NP + V + perfective form	16	10.32	in a 'verbed' state	**-esiga** 'trust' > **obwesige** 'trustworthiness'
	NP + adjective root	11	7.10	the quality or state of being 'adjective'	**-bi** 'ugly; bad' > **obubi** 'ugliness; badness'
	NP + V + a	11	7.10	the state of 'verbing'	**-yinza** 'be able' > **obuyinza** 'power; authority'
	NP + V + u	10	6.45	the condition of 'verbing'	**-guma** 'be hard; be strong' > **obugumu** 'hardness; strongness'
	NP + V + o	6	3.87	the result of 'verbing'	**-eyama** 'pledge' > **obweyamo** 'reference'
	NP + a + PC7 + NP2-pp + V + a	3	1.94	institution which represents the 'verbed'	**-zinga** 'encircle' > **Obwakyabazinga** 'Busoga kingship'
	SUM	155	100.00		
15/6	NP + noun root	8	100.00		
16	NP + noun root	4	50.00		
	NP + N	4	50.00	locativized 'noun'	**ka** 'home' > **awaka** 'at home, in a home'
	SUM	8	100.00		
20	NP + noun root	1	100.00		
23	NP + noun root	79	97.53		
	NP + V + N	2	2.47	where the 'noun' 'verbs'	**-gwa-** 'fall' + **ndhuba** 'sun' > Bugwandhuba 'West'
	SUM	81	100.00		

With N = noun; NP = noun prefix; -pp = minus pre-prefix; PC = possessive concord; V = verb; and the formatives: INHE = female head (wife of ISE); ISE = male head, leader; KA = elevator; KI = describer; NA = specifier; WA = personifier.

Appendix 17.1: Semantic import of gender 4 (to be compared with Appendix 6.3).

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	6	66.67	159	63.86
Fauna (animals)	0	0.00	0	0.00
Flora (plants)	0	0.00	0	0.00
Human body parts	0	0.00	0	0.00
Liquids	0	0.00	0	0.00
Man-made abstracts	0	0.00	0	0.00
Man-made concretes	0	0.00	0	0.00
Nature	1	11.11	21	8.43
People	0	0.00	0	0.00
Others	2	22.22	69	27.71
N / Freq.	9	100.00	249	100.00

Appendix 17.2: Semantic import of gender 6 (to be compared with Appendix 8.3).

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	15	38.46	688	37.09
Fauna (animals)	2	5.13	32	1.73
Flora (plants)	2	5.13	74	3.99
Human body parts	3	7.69	69	3.72
Liquids	8	20.51	226	12.18
Man-made abstracts	0	0.00	0	0.00
Man-made concretes	9	23.08	766	41.29
Nature	0	0.00	0	0.00
People	0	0.00	0	0.00
Others	0	0.00	0	0.00
N / Freq.	39	100.00	1,855	100.00

Appendix 17.3: Semantic import of gender 8 (to be compared with Appendix 10.3).

Semantic import	N	%	Freq.	%
Abstracts (non-temporal)	1	5.00	12	3.33
Fauna (animals)	0	0.00	0	0.00
Flora (plants)	0	0.00	0	0.00
Human body parts	0	0.00	0	0.00
Liquids	0	0.00	0	0.00

Man-made abstracts	19	95.00	348	96.67
Man-made concretes	0	0.00	0	0.00
Nature	0	0.00	0	0.00
People	0	0.00	0	0.00
Others	0	0.00	0	0.00
N / Freq.	20	100.00	360	100.00

Appendix 17.4: Semantic import of gender 14 (to be compared with Appendix 15.3).

Semantic import	**N**	**%**	**Freq.**	**%**
Abstracts (non-temporal)	118	76.13	4,132	76.97
Fauna (animals)	3	1.94	54	1.01
Flora (plants)	5	3.23	101	1.88
Human body parts	1	0.65	21	0.39
Liquids	2	1.29	26	0.48
Man-made abstracts	4	2.58	158	2.94
Man-made concretes	6	3.87	147	2.74
Nature	11	7.10	653	12.16
People	0	0.00	0	0.00
Others	5	3.23	76	1.42
N / Freq.	155	100.00	5,368	100.00