https://doi.org/10.14618/ids-pub-11343

Gilles-Maurice de Schryver/Minah Nabirye

TOWARDS A MONITOR CORPUS FOR A BANTU LANGUAGE

A case study of neology detection in Lusoga

Abstract This paper looks at whether, after two decades of corpus building for the Bantu languages, the time is ripe to begin using monitor corpora. As a proof-of-concept, the usefulness of a Lusoga monitor corpus for lexicographic purposes, *in casu* for the detection of neologisms, both in terms of new words and new meanings, is investigated and found useful.

Keywords Monitor corpus; neology detection; new words; new meanings; Bantu; Lusoga

1. Corpus building for the Bantu languages

Corpus building efforts for the Bantu languages remain in their infancy, and not much has changed since the overview published in de Schryver/Prinsloo (2000) – with current corpus sizes typically anywhere between a million and five million tokens. These corpora have mainly been used for dictionary compilation, corpus linguistics, and NLP applications. The last collection of studies in the field of Bantu NLP is already a decade old (De Pauw et al. 2011), and includes studies for all languages from South Africa, as well as Swahili. Recent studies in Bantu corpus linguistics include Dom/de Schryver/Bostoen (2020) for Kisikongo, Kawalya/Bostoen/de Schryver (2021) for Luganda, and Misago/Nshimirimana/Tuyubahe (2021) for Kirundi. Examples of corpus-driven dictionaries compiled for Bantu languages over the past 15 years include de Schryver (2007) for Northern Sotho, de Schryver (2010) for Zulu, and de Schryver/Reynolds (2014) for Xhosa. In all these cases, one or more corpora were built, typically subdivided into a number of sub-corpora reflecting different time periods, genres, and/or topics. The majority of Bantu corpora to date are also 'raw', in that they have not been marked for parts of speech, nor been lemmatised. Also, no project so far has tried to build a 'monitor corpus' for a Bantu language, with which the changing language may be (semi-)automatically tracked (see e.g. Kosem et al. 2021; Kosem 2022). In the current study we attempt exactly that, and apply it to the detection of neologisms in Lusoga, with the aim of improving existing dictionaries for this language.

2. Corpus building for Lusoga

Lusoga is a Great Lakes Bantu language spoken in the Busoga Kingdom, in Eastern Uganda, by about three million people (UBOS 2016, p. 71). Despite a flurry of activity over the past two decades, it may still be classified as a predominantly oral language. During this period, the corpus building effort has been heroically carried forward by a single person (the second author of the present paper), as described in de Schryver/Nabirye (2018). Half a decade ago, the Lusoga corpus stood at a respectable 1.7m tokens (with an oral part of over half a million tokens, 541k more precisely), a corpus mainly used as 'the body of evidence' in writing the first corpus-based grammar of the language (Nabirye 2016). Corpus building continued unabated, and included a special focus on transcriptions of diverse oral data, to reach 3.0m

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

tokens in September 2019 (oral part: 786k; a selection and analysis of which was published in book form: Nabirye 2019).

Within the field of corpus building for the Bantu languages, the Lusoga corpus of 3.0m tokens was considered 'large enough', for it to be able to serve as a base for all future Lusoga studies.[1] Among the tests performed to judge whether or not the Lusoga corpus of 3.0m was also 'stable enough' to act as a reference corpus, stability tests similar to those described by Prinsloo/de Schryver (2001) for the Bantu languages Northern Sotho and Xitsonga were conducted.

Over the past two years, another half a million tokens were collected in addition, bringing the total size of the Lusoga corpus up to 3.5m tokens, nearly a million of them (910k) transcribed material. While it is still a raw corpus, the oral component corresponds to a massive 152 hours of audio recordings; the written component to about 16,000 pages of running text.

3. Towards a monitor corpus for Lusoga

The proof of a pudding is in the eating. It is one thing to judge that a corpus is large and stable enough to be used as a base and reference corpus; it is another entirely to also actually *use* it as such. One valuable use of such a corpus, if it does what it is supposed to do, is to act as a monitor corpus. In their standard textbook, McEnery/Hardie (2012, p. 246) define a 'monitor corpus' as: "A corpus that grows continually, with new texts being added over time so that the dataset continues to represent the most recent state of the language as well as earlier periods." Hanks (2003, p. 53) literally defines a 'dynamic' or so-called 'monitor corpus' in two words: "constantly growing". This may all be good and well and perhaps even feasible for big languages such as English, but for Bantu corpora with their typical modest sizes one can surely not simply keep adding material opportunistically, as one's corpus would lose all its balance and representativeness. As such, given that extreme care is taken to continuously balance out the genres and topics that are being added to the Lusoga corpus, so that it remains representative of both spoken and written Lusoga at all times, some earlier data are sometimes even removed before new material is added (see e.g. de Schryver/Nabirye 2018, § 3.3 vs. § 3.4). In a similar vein, and thus also for reasons of balance and representativeness, 1.6m tokens (of judicial material) have for instance always been kept separate from the main Kirundi corpus (Misago 2018, p. 38), or 2.0m tokens (of religious material) have always been kept separate from the main Luganda corpus (Kawalya 2017, § 5.2 vs. § 5.3 in Chapter 1). In this regard, Kilgarriff's characterisation of how to use monitor corpora for lexicographic purposes is probably more to the point:

> a long-standing vision is the 'monitor corpus', the moving corpus that lets the researcher explore language change objectively (Clear 1988, Janicivic and Walker 1997). The core method is to compare an older 'reference' corpus with an up-to-

[1] In order to put corpus sizes for Great Lakes Bantu languages in context, for the much larger language Kirundi (the national and official language of Burundi, spoken by 8 million people), three scholars contributed to the building of a Kirundi corpus to inform their respective PhDs: Mberamihigo (2014) built and used a Kirundi corpus of 1.9m tokens (oral part: 51k), Nshemezimana (2016) enlarged that to 2.2m tokens (oral part: 196k), and Misago (2018) reached 2.8m tokens (oral part: 418k). For Lusoga's bigger neighbour, Luganda (one of the national languages of Uganda, spoken by 6 million people), Kawalya (2017) built and used a corpus of 4 million tokens for his PhD. Both Kirundi and Luganda have a rich written tradition.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

> the-minute one to find words which are not already in the dictionary, and which are in the recent corpus but not in the older one. (Kilgarriff 2013, p. 81)[2]

The detection of 'new words' is not the only goal though, as dictionary compilers are also, and sometimes even more so, interested in the detection of new usages, and thus 'new meanings' (cf. Hanks 2002), of existing words:

> Monitor corpora are primarily of importance in lexicographic work [...] They enable lexicographers to trawl a stream of new texts looking for the occurrence of new words or for changing meanings of old words. (McEnery/Wilson 2001, p. 30)

Therefore, and in terms of methodology, we will now compare the additional 0.5m Lusoga material to the earlier 3.0m reference corpus. To do so, we make use of the KeyWords tool from WST (Scott 2019), which calculates the 'outstandingness' of each corpus type. The assumption is that we will be able to detect *new words* which entered the language, as well as *new meanings* for existing words. For the first we assume that we can obtain a limited list of new types in the additional 0.5m that were absent from the 3.0m. For the second we assume that a limited list of 'outstanding' types (specifically types used relatively more frequently over the past two years), will hint at extra usages and thus new meanings. While this exercise may seem trivial, it is not, as what one does not want is long lists of so-called 'new words' that are not new at all but were all simply missing from the 3.0m corpus, and/or 'new meanings' that are not new at all but were all simply not used in the 3.0m corpus. That a certain percentage was truly missing or not used is acceptable (no corpus, no matter its size, contains all the words in all its uses for any given language), but to usefully act as a monitor corpus, a useful percentage must be 'new' words and usages, or thus 'neologisms'. If neologisms are indeed detected in this way, lexicographers may also act upon those. That said, if this exercise is successful – in the sense that it results in meaningful data that can be acted upon by dictionary compilers – we can then consider the 3.5m corpus as the new reference and thus new monitor corpus.

4. The semi-automatic identification of neologisms in Lusoga

4.1 New words

The default settings of WST's KeyWords were used and, fair enough, a limited number of 55 keywords occurring in at least two of the new texts was found that had not been seen in the 3.0m corpus. An analysis of the categories these 55 'new words' belong to is shown in Figure 1.

One of the 'new words', unsurprisingly, is ***COVID***, a clear neologism. (***Corona*** was also picked up, but because there was already a single mention of it – as the "Corona Hospital" (in California) – it was marked as outstanding; see § 4.2.) As with every non-English monitor corpus, the expectation was that a good number of 'new words' would be proper names

[2] The second reference should be to Janicijevic/Walker (1997); and the title of Clear's (1988) paper is "Trawling the language: Monitor corpora" rather than the misquoted "The Monitor Corpus". Both of these errors are unfortunately found all over the metalexicographic literature. More upbeat: As with so much in our field, the term 'monitor corpus' was coined by John Sinclair, in 1982, or thus four decades ago already (Clear 1988, p. 383).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

and (given that English is the only official language of Uganda) also plain English words. This is borne out; these categories make up 16% and 4% respectively, so 20% in all. Proper names and plain English words are not normally items that warrant inclusion in a Lusoga dictionary.

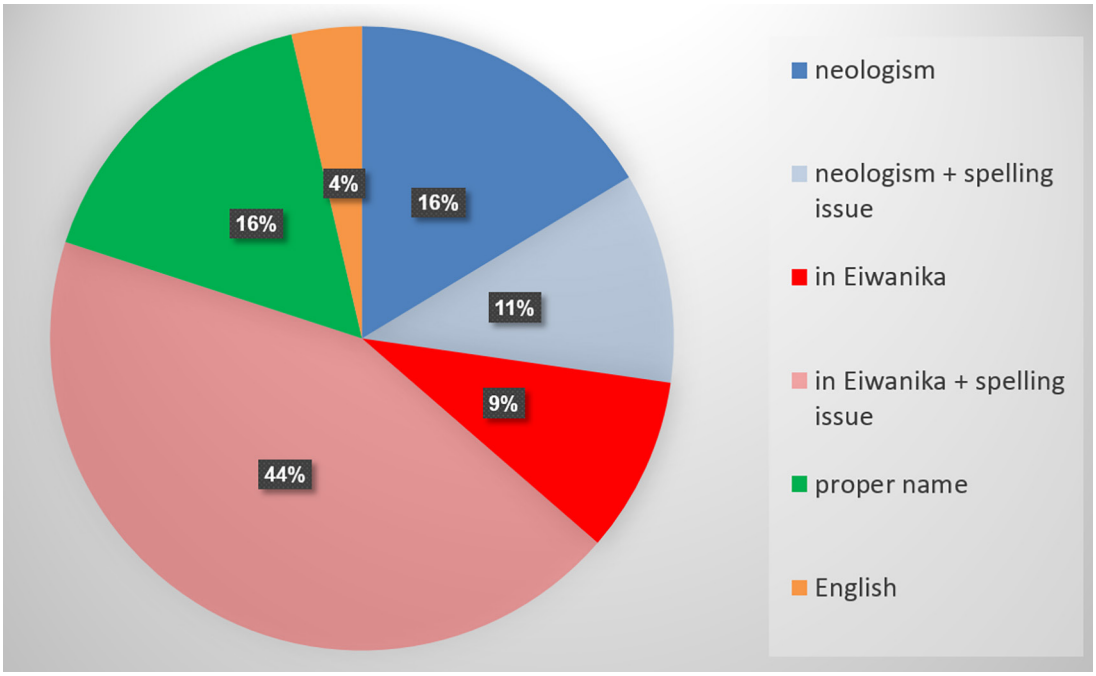

Fig. 1: Analysis of the 'new words' that entered Lusoga between September 2019 and January 2022

In order to analyse the data, it is important to note that no common (standard) orthography has yet been adopted by all those who write in Lusoga, so a tool like WST will also pick up (and give too much weight to) spelling variations – see 'spelling issue' in Figure 1. Had a dictionary been compiled based on the 3.0m corpus, the remaining 80% of the keywords from Figure 1, would all be candidates for inclusion in an update to the dictionary. While such a corpus-based dictionary does not exist, a *non-corpus-based* monolingual dictionary has been compiled, namely the *Eiwanika ly'Olusoga* (Nabirye 2009), online since 2012 at https://menhapublishers.com/dictionary/.

The quest for neologisms may then be rephrased as a quest for candidates to update that dictionary. Astonishingly, as many as 53% of the 'new words' from Figure 1 were already included in the *Eiwanika*, so they are not new words at all; just 27% are. The latter include new loanwords for ***omusaseredooti*** 'priest' (< Latin *sacerdos* 'priest'), ***mwepisikoopi*** 'bishop' (< Latin *episcopus* 'bishop'), and ***ukarisitia*** 'Eucharist' (< Greek *Eucharist* 'gratitude'), but also concepts that can only be 'derived', using language-internal processes, from other words already in the *Eiwanika*, and which are thus debatable neologisms, such as ***obukurisitu*** 'Christianity' (***Omukristo*** 'Christian' is in the *Eiwanika*), ***omuyumo*** 'entertainer' (***ekinhumo*** 'party' is in), or the reduplicated form ***mutoto*** 'youngish' (***-to*** 'young' is in). Conversely, others are clearly true neologisms: ***akanhomero*** 'a small pejorative place' (< ***okunhooma*** 'despise') or ***ekizezengere*** 'shadow' (the personification of ***ekinzenze*** 'a shadow').

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

Regarding the first three religious terms here (for 'priest', 'bishop', and 'Eucharist') one may wonder why we label them neologisms, as surely terms for those concepts were already in the language. Suffice it to say that competing religious groups devised their own terms in Lusoga, and that with the recent publication and now inclusion in the Lusoga corpus of Roman Catholic material, these 'new' terms (for old concepts) have now also officially entered the Lusoga language.[3]

4.2 New meanings

In addition to the 55 'new words', WST also lists 1,251 'outstanding words': 815 'positive keywords' (= words that are relatively more frequent in the new 0.5m material compared to the monitor corpus of 3.0m), and 436 'negative keywords' (= words that are relatively less frequent in the new material compared to the monitor corpus, and may thus be 'disappearing from the language'). Of the positive keywords, 466 occur in at least two of the new 0.5m corpus files, while 349 occur in just one of the new files. For the purposes of the present paper, we will only look at the top 100 positive keywords that occur in at least two new texts. An analysis of the categories these 'top 100' belong to is shown in Figure 2.

In Figure 2, the proportion of proper names has slightly grown compared to Figure 1 (to 20%), while that of English has gone down (to 2%). A notable proper name that is now far more outstanding is that of ***Gabula***, the title of the current Busoga King. In terms of candidate new meanings, as many as 61% turn out to have been properly covered in the *Eiwanika*, with their various meanings; yet 17% have not. Some of these 17% indicate that a number of function words which are the result of grammatical constructions had better been lemmatised in the *Eiwanika*, such as the connectives (construction = pronominal prefix + ***-a***), and that some combinations also warrant lemma-sign status, such as ***-liwo*** 'be present' (< ***-li*** 'to be' + ***wo*** (locative)), or ***me ni*** 'and then' (< ***me*** 'and then' + ***ni*** (focus)). These, of course, are neither new words nor new uses; yet the software has (correctly!) picked them out as candidate entries. So here the use of a monitor corpus for Lusoga has not detected new meanings, but forces lexicographers to face the facts; and the fact is that more grammar needs to be entered into the central lemma-sign list of a dictionary.

[3] The work concerned is the Roman missal (Gonza 2018); which despite being dated 2018 only became available in late 2019, whereupon it was scanned, OCRed, and heavily processed (by the first author of this paper, to take out all the English parts) before it was added to the corpus. The Protestant Bible (BSU 2014) was already in the corpus.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

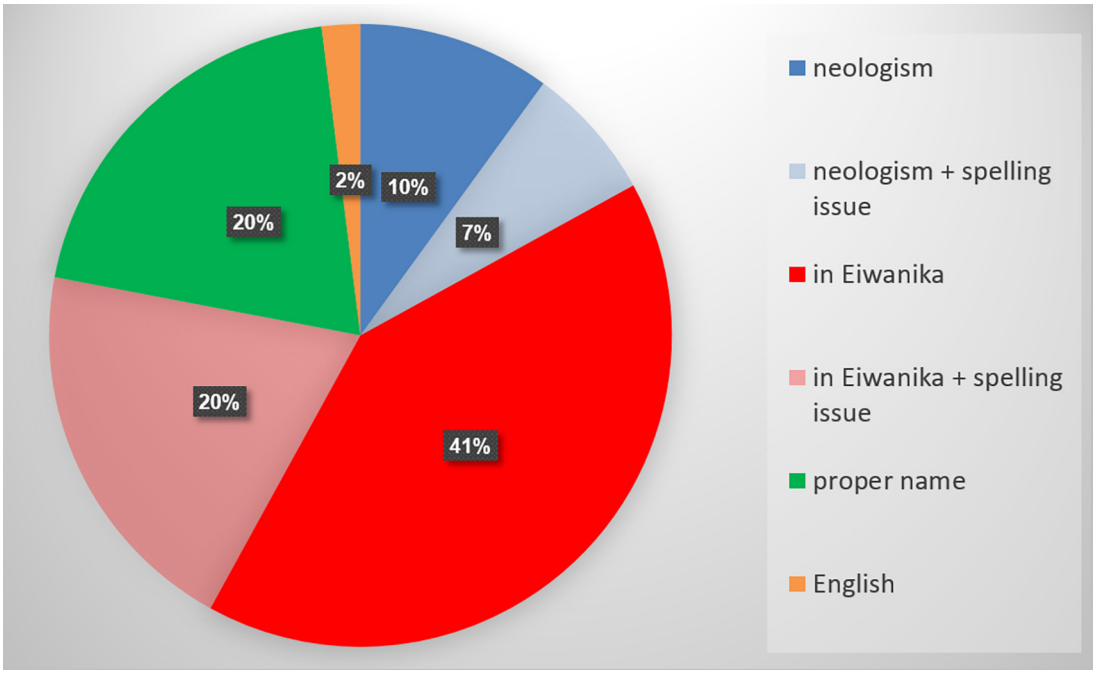

Fig. 2: Analysis of the top 100 'outstanding words' (in at least two corpus files) when comparing Lusoga between September 2019 and January 2022

Full words not lemmatised in the *Eiwanika* include ***lebe*** 'so and so', as well as the interjection ***eee***. The specific but non-descript meaning 'so and so' may be considered a near-neologism; it was hardly there before but now entered the language 'in force'. Similarly for the unspecific interjection *eee*, while not lemmatised in the *Eiwanika*, it *was* used once in a single example (under the lemma ***(a)keewuunia***).

An interesting language change is ***eisakaramentu*** 'sacrament': ***saakalamentu*** was lemmatised in the *Eiwanika*, but the monitor corpus now indicates that the form with a noun class prefix has become far more acceptable than it used to be.

The remainder are all clear cases of neologisms, as these are words that acquired new and very specific meanings. These include: ***ebyeghongo*** 'things used to pray; gifts' (deverbative < ***okuwonga*** 'to give offerings in church'), ***amaingira*** 'the process of entering' (deverbative < ***okwingila*** 'to enter'), ***ekitaloodheka*** 'that which is difficult to relay' (deverbative < cl. 7 noun prefix + negative marker ***-ta*** + ***okuloodha*** 'to relay' + stative extension), ***olugololiro*** 'in a straight manner' (deverbative < ***okugolola*** 'to make straight'), and ***kituufu*** 'it is true' (adjective < ***obutuufu*** 'truthfulness').

5. Discussion and conclusion

As Kilgarriff (2013, p. 82) correctly pointed out: "The nature of the task is that the automatic process creates a list of candidates, and a lexicographer then goes through them to sort the wheat from the chaff. There is always far more chaff than wheat." In terms of 'new words', adding half a million Lusoga tokens to a corpus of 3 million tokens, revealed just 55 items, so having to sort the wheat (which turned out to be 27%) from the chaff manually for such a small amount is more than doable. In terms of 'new meanings', we presented an analysis of the top 100 outstanding words only, where we saw that the wheat was less

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

forthcoming (17%). The full details of going from raw data to analysis may be found in Addenda 1 and 2.[4]

While Kilgarriff does not give us an indication of an acceptable ratio of wheat to chaff, apart from informing us that it is inherently low, we feel that the exercise for Lusoga was worthwhile, as we did pinpoint enough useful material to update the *Eiwanika*. As a result, we are confident that the dawn of monitor corpora for the Bantu languages has arrived.

However, upon also considering recall and precision when going down the list of potential new meanings [to be presented during the actual talk only, as space constraints do not allow for a full description here], we are dealing with a case of diminishing returns: The recall does indeed go up, but at an increasingly punishing precision. Another bottleneck, especially with hopes of automating the process in future, revolves around the various spellings used among the Basoga community; but this is a language-specific problem, not a Bantu-wide one.

References

BSU. 2014. Baibuli. Ekibono kya Katonda. Omuli n'ebitabo ebyetebwa deuterokanoniko/apokurifa [Bible. The word of God, which also has the books known as Deuteronomy/Apocrypha]. Kampala.

Clear, J. (1988): Trawling the language: monitor corpora. In: Snell-Hornby, M. (ed.): ZuriLEX '86 Proceedings: Papers read at the EURALEX International Congress. Tübingen, pp. 383–389.

De Pauw, G./de Schryver, G.-M./Pretorius, L./Levin, L. (2011): Introduction to the special issue on African Language Technology. In: Language Resources and Evaluation 45 (3), pp. 263–269.

de Schryver, G.-M. (2007): Oxford bilingual school dictionary: Northern Sotho and English / Pukuntšu ya Polelopedi ya Sekolo: Sesotho sa Leboa le Seisimane. E gatišitšwe ke Oxford. Cape Town.

de Schryver, G.-M. (2010): Oxford bilingual school dictionary: Zulu and English / Isichazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgisi, Esishicilelwe abakwa-Oxford. Cape Town.

de Schryver, G.-M. (2020): Linguistics terminology and neologisms in Swahili: rules vs. practice. In: Dictionaries: Journal of The Dictionary Society of North America 41 (1), pp. 83–104 + 14 pages of supplementary material online.

de Schryver, G.-M./Nabirye, M. (2018): Corpus-driven Bantu lexicography, Part 1: Organic corpus building for Lusoga. In: Lexikos 28, pp. 32–78.

de Schryver, G.-M./Prinsloo, D. J. (2000): The compilation of electronic corpora, with special reference to the African languages. In: Southern African Linguistics and Applied Language Studies 18 (1–4), pp. 89–106.

de Schryver, G.-M./Reynolds, M. (2014): Oxford bilingual school dictionary: IsiXhosa and English / Oxford isiXhosa-isiNgesi English-isiXhosa Isichazi-magama Sesikolo. Cape Town.

Diki-Kidiri, M. (ed.) (2008): Le vocabulaire scientifique dans les langues africaines. Pour une approche culturelle de la terminologie. (= Collection Dictionnaires et Langues). Paris.

[4] Note that it has not been the purpose of this paper to also analyse the various linguistic strategies used to create neologisms in Lusoga, even though some of the evidence for this may be deduced from the Addenda. For a recent Bantu example in this domain, see de Schryver (2020) who deals with term creation processes in Swahili. In Africa more generally, many scholars have worked on this as well; see for instance the edited collection by Diki-Kidiri (2008) for case studies in West and Central Africa.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

Dom, S./de Schryver, G.-M./Bostoen, K. (2020): Kisikongo (Bantu, H16a) present-future isomorphism: a diachronic conspiracy between semantics and phonology. In: Journal of Historical Linguistics 10 (2), pp. 251–288.

Gonza, R. K. (2018): Misaale mu Lusoga. Jinja.

Hanks, P. (2002): Mapping meaning onto use. In: Corréard, M.-H. (ed.): Lexicography and natural language processing. A festschrift in honour of B.T.S. Atkins. Euralex, pp. 156–198.

Hanks, P. (2003): Lexicography. In: Mitkov, R. (ed.): The Oxford handbook of computational linguistics. Oxford, pp. 48–69.

Janicijevic, T./Walker, D. (1997): NeoloSearch: Automatic detection of neologisms in French Internet documents. In: Proceedings of the 1997 Joint International Conference of the Association for Computers and the Humanities and the Association for Literary & Linguistic Computting. Kingston, Ontario, pp. 93–94.

Kawalya, Deo. 2017. A corpus-driven study of the expression of modality in Luganda (Bantu, JE15). PhD dissertation. Ghent.

Kawalya, D./Bostoen, K./de Schryver, G.-M. (2021): A diachronic corpus-driven study of the expression of possibility in Luganda (Bantu, JE15). In: International Journal of Corpus Linguistics 26 (3), pp. 336–369.

Kilgarriff, A. (2013): Using corpora as data sources for dictionaries. In: Jackson, H. (ed.): The Bloomsbury companion to lexicography. London, pp. 77–96.

Kosem, I. (2022): Trendi – a monitor corpus of Slovene. In: Klosa-Kückelhaus, A./Engelberg, St./Möhrs, Ch./Storjohann, P. (eds): Proceedings of the XX EURALEX International Congress: Dictionaries and Society. Mannheim.

Kosem, I./Krek, S./Gantar, P./Holdt, Š. A. /Čibej, J. (2021): Language monitor: tracking the use of words in contemporary Slovene. In: Kosem, I./Cukr, M./Jakubíček, M. /Kallas, J./Krek, S./Tiberius, C. (eds): Electronic Lexicography in the 21st Century (eLex 2021): Post-Editing Lexicography. Proceedings. Brno, pp. 514–528.

Mberamihigo, F. (2014): L'expression de la modalité en kirundi: exploitation d'un corpus électronique. PhD dissertation. Brussels/Ghent.

McEnery, T./Hardie, A. (2012): Corpus linguistics: method, theory and practice (= Cambridge Textbooks in Linguistics). Cambridge.

McEnery, T./Wilson, A. (2001): Corpus linguistics: an introduction. 2nd Edition. (= Edinburgh Textbooks in Empirical Linguistics). Edinburgh.

Misago, M.-J. (2018): Les verbes de mouvement et l'expression du lieu en kirundi (bantou, JD62): une étude linguistique basée sur un corpus. PhD dissertation. Ghent.

Misago, M.-J./Nshimirimana, E./Tuyubahe, P. (2021): Usages grammaticaux du verbe -guma 'rester' en kirundi (JD62): Une étude linguistique basée sur un corpus. In: Language in Africa 2 (1), pp. 3–40.

Nabirye, M. (2009): 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] (= Linguistics Series 1). Kampala.

Nabirye, M. (2016): A corpus-based grammar of Lusoga. PhD dissertation. Ghent.

Nabirye, M. (2019): Owayanga: Empayo Dhimala Dhaavaamu Olufumo [Speak regularly: conversations have a tendency to become legends]. (= Linguistics Series 5). Kampala.

Nshemezimana, E. (2016): Morphosyntaxe et structure informationnelle en kirundi: Focus et stratégies de focalisation. PhD dissertation. Ghent.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

Prinsloo, D. J./de Schryver, G.-M. (2001): Monitoring the stability of a growing organic corpus, with special reference to Sepedi and Xitsonga. In: Dictionaries: Journal of The Dictionary Society of North America 22, pp. 85–129.

Scott, M. (2019): WordSmith Tools, version 7. http://www.lexically.net/wordsmith/.

UBOS (2016): The National Population and Housing Census 2014 – Main Report. Kampala.

Contact information

Gilles-Maurice de Schryver
BantUGent – UGent Centre for Bantu Studies, Ghent University
& Department of African Languages, University of Pretoria
gillesmaurice.deschryver@UGent.be

Minah Nabirye
BantUGent – UGent Centre for Bantu Studies, Ghent University
minah.nabirye@UGent.be

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66