TOWARDS A MONITOR CORPUS FOR A BANTU LANGUAGE

A CASE STUDY OF NEOLOGY DETECTION IN LUSOGA

GILLES-MAURICE DE SCHRYVER & MINAH NABIRYE

BANTUGENT – UGENT CENTRE FOR BANTU STUDIES

CORPUS BUILDING FOR THE BANTU LANGUAGES

REALITY CHECK

- Corpus building efforts for the Bantu languages remain in their infancy,
- 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.

REALITY CHECK

- For each language, 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.

TOWARDS THE FIRST MONITOR CORPUS

- 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.

CORPUS BUILDING FOR LUSOGA

- Lusoga is a Great Lakes Bantu language spoken in the Busoga Kingdom, in Eastern Uganda, by about three million people.
- Lusoga may still be classified as a predominantly oral language.

CORPUS BUILDING FOR LUSOGA > GRAMMAR

- The corpus building effort has been heroically carried forward by a single person,
- 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 FOR LUSOGA > TEXTS

- Corpus building continued unabated, and included a special focus on transcriptions of diverse oral data,
- to reach 3.0m tokens in September 2019 (oral part: 786k)

- A selection and analysis of which was published in book form: Nabirye (2019).

MENHA
LINGUISTICS
SERIES

OWAYANGA

Empayo Dhimala Dhaavaamu Olufumo

Minah Nabirye

BOASIAN TRILOGY (DICTIONARY, GRAMMAR, TEXTS)

- Corpus building for Lusoga > Grammar (2016)
- Corpus building for Lusoga > Texts (2019)

><

- Monolingual Lusoga Dictionary (2009): NOT corpus-based

Eiwanika ly'Olusoga (Monolingual Lusoga Dictionary)

FROM 3.0M > 3.5M

- 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.

- 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.

3.5M

- 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.

TOWARDS A MONITOR CORPUS FOR LUSOGA

DEFINITIONS

- In their standard textbook, McEnery & Hardie (2012: 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: 53) literally defines a 'dynamic' or so-called 'monitor corpus' in two words:

  "constantly growing"

DEFINITIONS

- Kilgarriff's (2013: 81) 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-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."

DEFINITIONS

- 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: 30)

METHODOLOGY

- 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.

METHODOLOGY

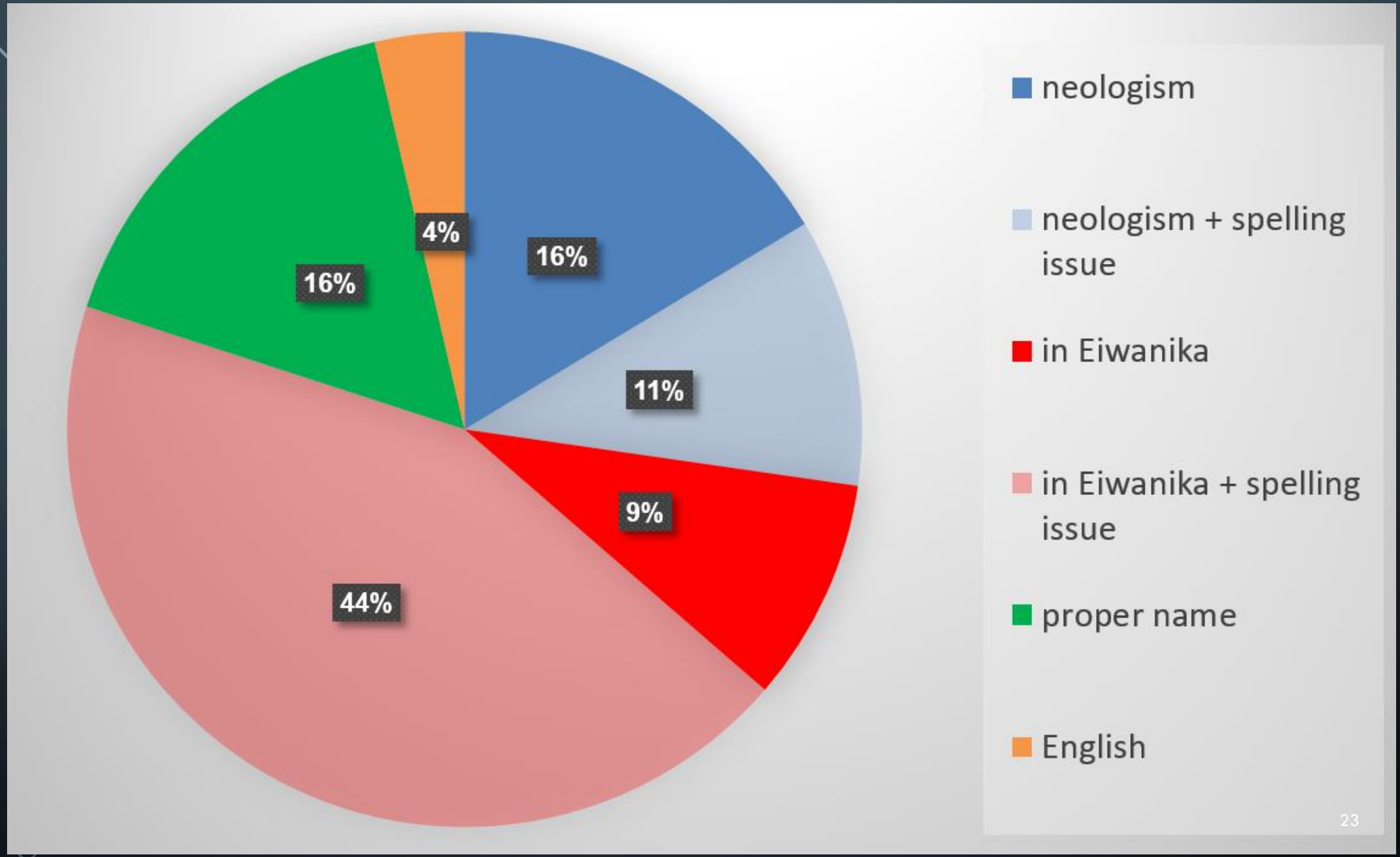
- The assumption is that we will be able to detect new words which entered the language, as well as new meanings for existing words.
- For the first we assume that we can obtain a limited list of new types in the additional 0.5m that were absent from the 3.0m.
- For the second we assume that a limited list of 'outstanding' types (specifically types used relatively more frequently over the past two years), will hint at extra usages and thus new meanings.
- 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.

THE SEMI-AUTOMATIC IDENTIFICATION OF NEOLOGISMS IN LUSOGA

1. NEW WORDS

RESULTS

- 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 the next FIGURE.



Analysis of the ‘new words’ that entered Lusoga between September 2019 and January 2022

ANALYSIS

- 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 categorised as outstanding by WST.

ANALYSIS

- The quest for neologisms may be rephrased as a quest for candidates to update that dictionary.
- Astonishingly, as many as 53% of the 'new words' were already included in the *Eiwanika*, so they are not new words at all; just 27% are.

NEW WORDS: LOANWORDS

- New loanwords for:
  - ***omusaseredooti*** 'priest' (< Latin *sacerdos* 'priest')
  - ***mwepisikoopi*** 'bishop' (< Latin *episcopus* 'bishop')
  - ***ukarisitia*** 'Eucharist' (< Greek *Eucharist* 'gratitude')
- Surely terms for those concepts were already in the language, not?
- Well, competing religious groups devised their own terms in Lusoga, and 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 Lusoga.

NEW WORDS: 'DERIVABLE' CONCEPTS

- 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)
  - ***mutoto*** 'youngish' (***-to*** 'young' is in)

NEW WORDS: TRUE NEOLOGISMS

- Conversely, others are clearly true neologisms:
- ***akanhomero*** ‘a small pejorative place’ (< ***okunhooma*** ‘despise’)
- ***ekizezengere*** ‘outline; image’ (the personification of ***ekinzenze*** ‘a shadow’)

THE SEMI-AUTOMATIC IDENTIFICATION OF NEOLOGISMS IN LUSOGA

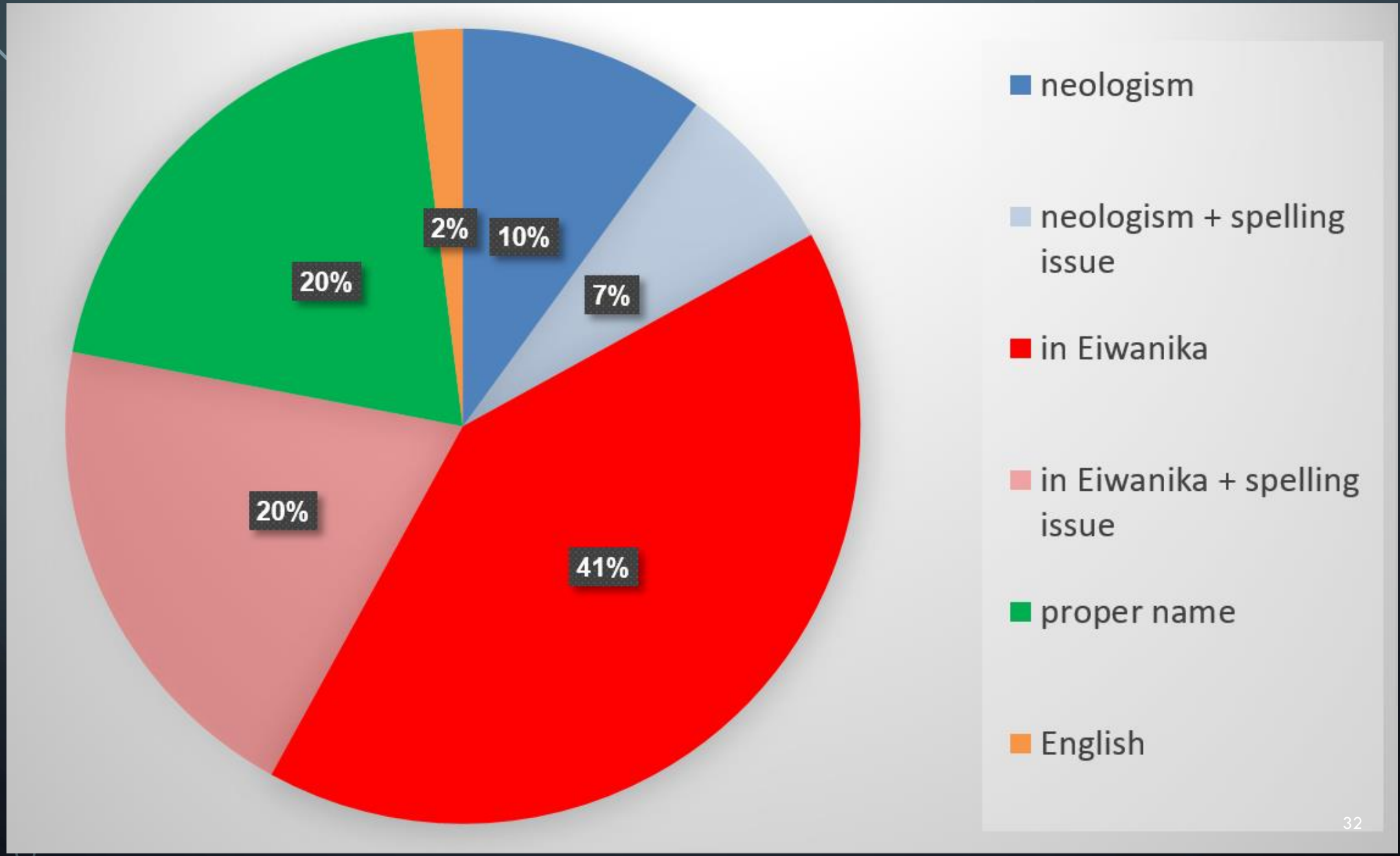
2. NEW MEANINGS

RESULTS

- 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’).

RESULTS

- 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 the next FIGURE.



Analysis of the top 100 ‘outstanding words’ when comparing Lusoga between September ’19 and January ’22

ANALYSIS

- A notable proper name that is now far more outstanding is that of ***Gabula***, the title of the current Busoga King.

ANALYSIS

- 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.

FACING THE FACTS: LEXICON VS. GRAMMAR

- 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)).

FACING THE FACTS: LEXICON VS. GRAMMAR

- 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.

FACING THE FACTS: WORD-STATUS

- 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***).

FACING THE FACTS: SPELLING

- 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.

NEW MEANINGS

- 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')

NEW MEANINGS

- ***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')
- ***kituufu*** 'it is true' (adjective < ***obutuufu*** 'truthfulness').

DISCUSSION AND CONCLUSION

CONCLUSION: NEW WORDS

- As Kilgarriff (2013: 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.

CONCLUSION: NEW MEANINGS

- In terms of 'new meanings', we presented an analysis of the top 100 outstanding words only, where we saw that the wheat was less forthcoming (17%).
- 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*.

DISCUSSION

- However, upon also considering recall and precision when going down the list of potential new meanings, 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.

- All in all, we are confident that the dawn of monitor corpora for the Bantu languages has arrived.