Ruth Vatvedt Fjeld, University of Oslo: How to make a dictionary I?

Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The *macrostructure*[1]

Gilles-Maurice de Schryver*

Research Assistant of the Fund for Scientific Research – Flanders (Belgium)
& Department of African Languages, University of Pretoria, Pretoria, 0002 South Africa
schryver@postino.up.ac.za

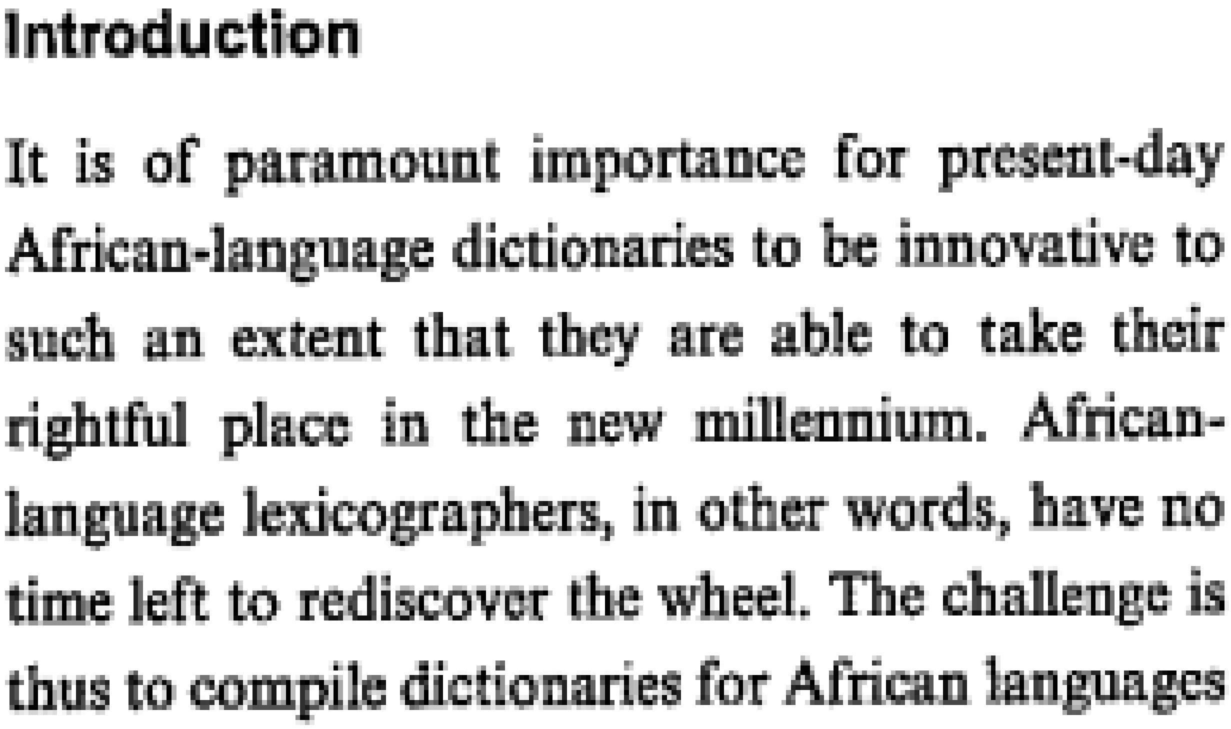
D.J. Prinsloo

Department of African Languages, University of Pretoria, Pretoria, 0002 South *Africa*
prinsloo@postino.up.ac.za

AFRILEX
African Association for Lexicography

Department of African Languages, University of Pretoria
Pretoria 0002, South Africa
Tel.: +27 (0)12 420 2494, Fax: +27 (0)12 420 3163
E-mail: E. Taljard

South African Journal of African Languages



Introduction

It is of paramount importance for present-day African-language dictionaries to be innovative to such an extent that they are able to take their rightful place in the new millennium. African-language lexicographers, in other words, have no time left to rediscover the wheel. The challenge is thus to compile dictionaries for African languages

Early corpus-based dictionaries

- Collins-Birmingham University International Language Database (COBUILD) ->

  The Collins Cobuild English Language Dictionary (Sinclair 1987)

- Oxford-Hachette English-French, French-English Dictionary (OXHA) (Atkins 1994)

Most famous corpusbased dictionaries:

- British National Corpus ->

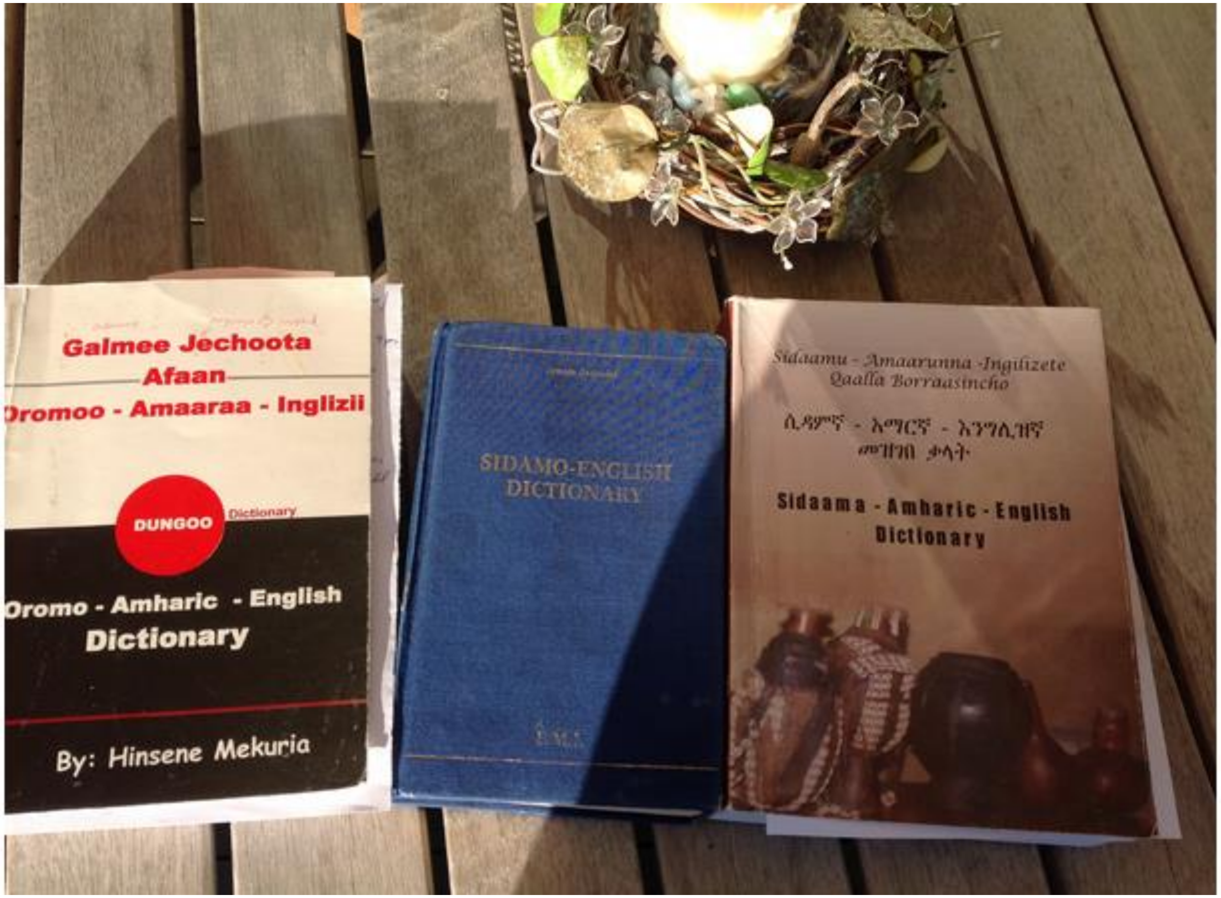
  Oxford Advanced Learners Dictionary of Current English (OALD)

- New editions of COBUILD-dictionaries

Corpus is important at three stages:

- Establishment of the nomenclature
- Entry preparation
- Revision (editing)

Three Ethiopian dictionaries



African dictionaries has a need for

- lemmmatised frequency lists
- data on lemma-sign distributions across sub-corpora

Most frequently used words in NBC :

COBUILD 2	**LDOC3**
5 diamonds – 700	W1 - 1000
4 diamonds – 1900	W2 - 2000
3 diamonds – 3400	W3 - 3000
2 diamonds – 6600	
1 diamond – 14700	

John Sinclair (1995):

The words in the five frequency bands are of immense importance to learners because they make up 95 % of all spoken and written English.

A sad situation
(or a good starting point):

- Lexicographical activities on the various indigenous African languages […have] resulted in a wide range of dictionaries. Unfortunately, the majority of these dictionaries are the products of limited efforts not reflecting a high standard of lexicographical achievement. (Gowus 1990:55)

Lemma selection

- How to find the words to include in the dictionary?
- Selection is guided by usefulness, and usefulness is determined by the degree to which terms most likely to be looked for are included. (Gove 1961)

Take a corpus …

- Extract a frequency list
- Compare it with the dictionary to identify and rectify mismatches
- Identify the one-, two-, and three thousand cut-off points
- Mark the corresponding dictionary entries accordingly

(Kilgarriff 1997:136)

Why is corpus frequency so important?

- To explore a corpus, means to see the language as conditioned by its social functions, and so the choice of words to go into the dictionaries, is determined by sophisticated assessments of frequency and the user’s needs (Osselton 1983:21)
- … we do not include words just because they are odd or interesting (Sinclair 1995:ix)

Compilation of lemma-sign list

You need two types of information:

- A lemmatised frequency list
- The lemma-sign distributions across certain sub-corpora

Be aware of:

The rank or position of items in ordered frequency lists

- Overall counts being the total number of occurencies of items in the entire corpus
- The distribution of those items across the different sub-corpora or sources.

Other considerations

Frequency is but one of many parameters according to which lexicographers decide that a word should be in the dictionary. Other criteria are, for example, contrastive relevance, disponibility, text-type specific relevance, etc. (Doherty & Heid, 1998:339)

From corpus to lemma sign list

- Most dictionaries of African languages are non-corpus dictionaries, documenting vocabulary from textbooks and special topics (according to the author’s interest).
- A small corpus of some hundred thousand tokens (running words) can help a lot in supplementing such a lemma list.

Documenting the central vocabulary

- In a small corpus, more than the half amount of the words will occur only once (hapax legomena).
- Most of the other words will be very frequent words, and belong to a lemma list for a small dictionary.

Zipf’s law (1949)

- The most common item has twice as many occurrences as the second most common, thre times as many as the third, a hundred times as many as the hundredth etc…

  (Kilgarriff 1997:136)

The lexicographic corpus

- A lexicographic corpus should be lemmatised. That means that all instances of a word is included in the same count:
- E.g. the verb *aim* includes *aims, aiming, aimed* but excludes the non-verbal forms *aim* and *aims*.
- Result: a lemmatised frequency list with POS-tags.

Frequency counts presented in the dictionary?

- Frequency is an information to the user of the usability of the lemma.
- Frequency is a documentation that the lemma is actually in use in the corpus.
- Written or spoken?
- Try to find a notation that is easy to understand and will be used!

Conclusion

- Frequency counts derived from corpora ensures that extremely frequent lemma signs ar not accidentally omitted from a dictionary.
- Precious space is not allocated to lemma signs unlikely to be looked up.
- Corpora can be put to good use in revising existing dictionaries.