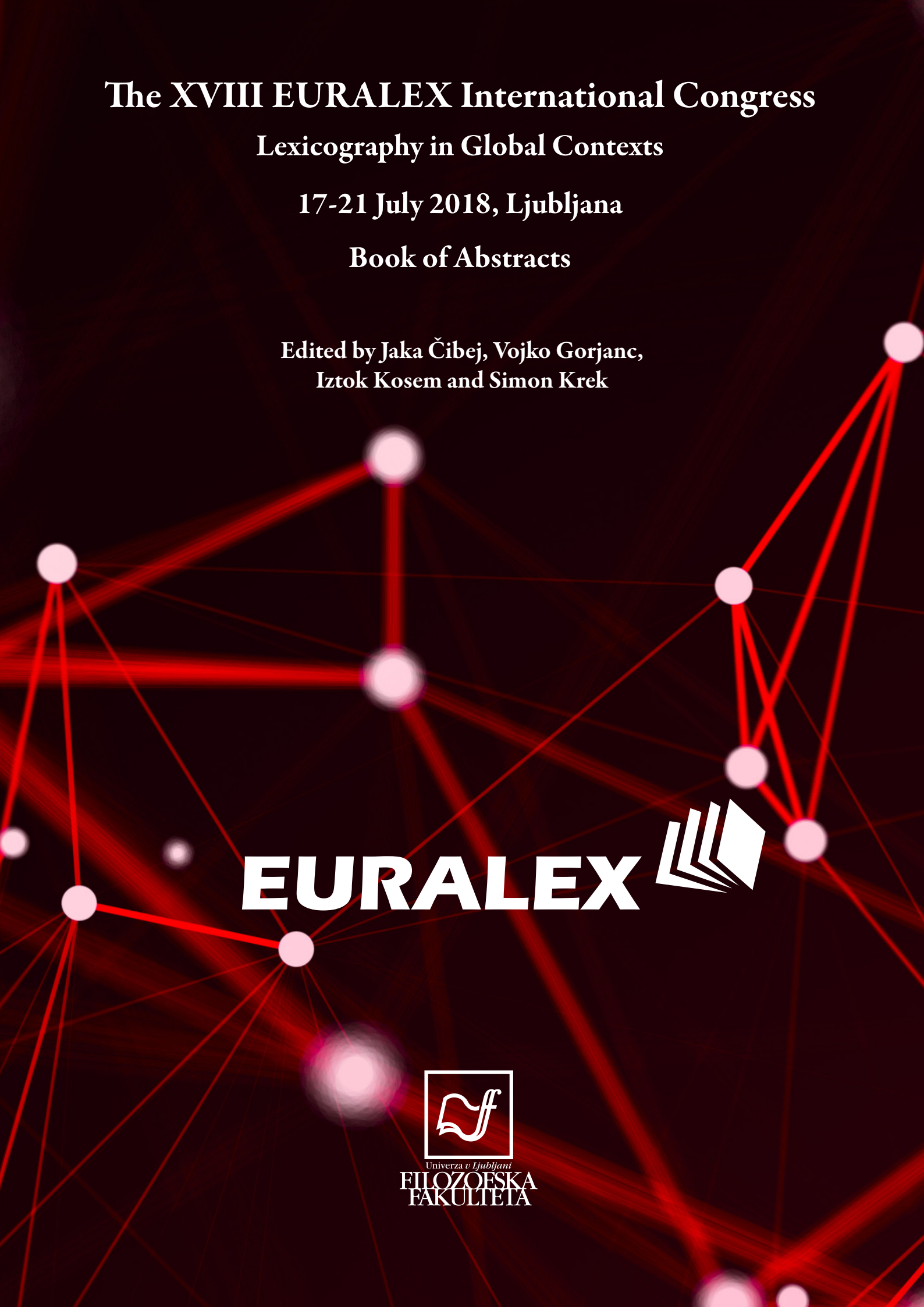
The XVIII EURALEX International Congress

Lexicography in Global Contexts

17-21 July 2018, Ljubljana

Book of Abstracts

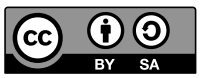
Edited by Jaka Čibej, Vojko Gorjanc,
Iztok Kosem and Simon Krek



Univerza v Ljubljani
FILOZOFSKA FAKULTETA

The XVIII EURALEX International Congress: Lexicography in Global Contexts
Book of Abstracts

Edited by: Jaka Čibej, Vojko Gorjanc, Iztok Kosem and Simon Krek
English language proofreading: Paul Steed
Technical editor: Aleš Cimprič



Published by: Znanstvena založba Filozofske fakultete Univerze v Ljubljani / Ljubljana University Press, Faculty of Arts
Issued by: University of Ljubljana, Centre for language resources and technologies
For the publisher: Roman Kuhar, Dean of the Faculty of Arts, University of Ljubljana

Ljubljana, 2018
First edition, e-edition
Publication is free of charge.

The editors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0215).

This text was written using the ZRCola input system (ZRCola.zrc-sazu.si),
developed at the Research Centre of the Slovenian Academy of Sciences and Arts in Ljubljana (www.zrc.sazu.si) by Peter Weiss.

DOI: 10.4312/9789610600954
E-book is available at: https://e-knjige.ff.uni-lj.si/

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID=295932416
ISBN 978-961-06-0095-4 (pdf)

Towards a new type of dictionary for Swahili

Gilles-Maurice de Schryver
Centre for Bantu Studies, Ghent University
gillesmaurice.deschryver@UGent.be

1 Problem statement

The Bantu language **Swahili** is ***the* lingua franca of East Africa**, **spoken by up to 100 million** first- and second-language speakers, especially in Tanzania and Kenya, but also in the neighbouring countries to their west and south (Mohamed 2009: iv-v). It is one of the most well-known African languages, and yet, the existing lexicographic output is the result of a century-and-a-half-old *craft* rather than a modern science (Benson 1964). In the present paper a theoretical framework is developed for modern Swahili lexicography.

2 Issues in existing swahili lexicography

What is **common to all the existing dictionaries for Swahili** is that their compilation was a fully **manual process, based on introspection**. The main strategies employed for the construction of the macrostructure were either (i) random, (ii) rule-oriented, or

(iii) enter-them-all approaches. In the random approach "words are simply added whenever they happen to cross the compiler's way", in the rule-oriented approach "a set of rules/guidelines presented in the dictionary's front matter must be followed whenever a word cannot be looked up directly" (so the assumption is that everything is covered 'in theory'), and in the enter-them-all approach "the compilers are obsessed to include all conceivable nominal and verbal derivations [working] through a modular paradigm in order to pursue such a comprehensiveness" (de Schryver & Prinsloo 2001: 219-25). **Modern dictionaries are corpus-based**, not only for the world's major languages (Hanks 2012), but also for the Bantu languages (de Schryver & Prinsloo 2000a, b), with the best aiming to be corpus-driven (de Schryver 2010). Quite surprisingly, corpora of Swahili *have* been used in lexicography, but mainly to evaluate existing dictionaries (Hurskainen 2004, De Pauw *et al.* 2009), rather than for their compilation.

3 Open questions in need of answers

the main idea is to compile a new type of dictionary, one to be **grown as a work in progress**, but to ensure a sound product at all times. This entails that attention needs to go to a number of facets: (i) **Born digital**. In this day and age, the lexicographic tool must first live in a digital environment, with only an optional transfer to paper at a later stage, not the other way round as remains all too common (eLex 2017). Is this theoretically sound for Swahili? (ii) **Define the target user groups**. Too many dictionaries are compiled without a clear picture of who the users will be. Serious thought must be given to this aspect, as it dictates several dictionary compilation decisions (Tono 2009: 39 ff.). Ideally, the project's theoretical base could cater for both native speakers and learners moving between Swahili and English, but also for speakers of Swahili who wish to remain within a Swahili environment. Is this

feasible? (iii) **Aim for a semi-bilingual reference work**. At face value, opting for a semi-bilingual approach seems like a good idea (Lew 2004). Such a work has characteristics of both a bilingual Swahili-English-Swahili dictionary, and a monolingual Swahili dictionary. But is this hunch also corroborated with actual dictionary use? (iv) **Know the users' lemmatisation needs**. No Bantu dictionary is purely word-based nor purely stem-based; all put the lexical items from the various word classes on a sliding cline between these extremes (de Schryver 2008: 86-87). With the exception of Johnson (1939), there seems to be a broad consensus on how to lemmatise the lexicon in Swahili. Tradition doesn't necessarily correspond to today's (digital) user needs, so one should dare question current practice. (v) **Let a corpus drive the compilation**. With the various preceding 'moving targets' (cf. ii to iv) as compilation proceeds (de Schryver 2005), can a corpus truly guide the macro-, medio- and microstructural compilation? (vi) **Use an existing dictionary writing system (DWS)**. Good off-the-shelf lexicographic software exists (Abel 2012), but can such packages handle all of the above? (vii) **Aim for structured dictionary compilation**. A DWS imposes a rather fixed structure; but is it flexible enough to deal with on-the-fly adaptations of the type envisaged for this Swahili dictionary project? (viii) **Study dictionary use**. In a digital environment, it is possible to unobtrusively study dictionary-use behaviour, while optionally allowing direct feedback in addition (de Schryver & Joffe 2004). In doing so, will one be in a position to check whether the target user groups are as expected (cf. ii)?, will one be able to fine-tune the exact type of dictionary type to work with (cf. iii)?, will one have the means to adapt the lemmatisation strategies (cf. iv)?, will one be able to judge whether the use of a corpus is the right approach (cf. v)?, will one be able to translate the feedback into feasible changes to the DTD or XML schemas (cf. vi)?, and will one end up with a unified overall structure (cf. vii)? The answers to these questions form the base for the sought theoretical framework.

Keywords: Swahili, digital dictionary, semi-bilingual, corpus-driven, user-friendly

References

Abel, Andrea. 2012. Dictionary Writing Systems and Beyond. In: Granger, Sylviane & Magali Paquot (eds). *Electronic Lexicography*: 83–106. Oxford: Oxford University Press.

Benson, T. G. 1964. A Century of Bantu Lexicography. *African Language Studies* 5: 64–91.

De Pauw, Guy, Gilles-Maurice de Schryver & Peter W. Wagacha. 2009. A corpus-based survey of four electronic Swahili–English bilingual dictionaries. *Lexikos* 19: 340–52.

de Schryver, Gilles-Maurice. 2005. Concurrent over- and under-treatment in dictionaries – The *Woordeboek van die Afrikaanse Taal*

as a case in point. *International Journal of Lexicography* 18(1): 47–75.

de Schryver, Gilles-Maurice. 2008. A new way to lemmatize adjectives in a user-friendly Zulu–English dictionary. *Lexikos* 18: 63– 91.

de Schryver, Gilles-Maurice. 2010. Revolutionizing Bantu lexicography – A Zulu case study. *Lexikos* 20: 161–201.

de Schryver, Gilles-Maurice & David Joffe. 2004. On How Electronic Dictionaries are Really Used. In: Williams, Geoffrey & Sandra Vessier (eds). *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004*: 187–96. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

de Schryver, Gilles-Maurice & D.J. Prinsloo. 2000a. Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The macrostructure. *South African Journal of African Languages* 20(4): 291–309.

de Schryver, Gilles-Maurice & D.J. Prinsloo. 2000b. Electronic corpora as a basis for the compilation of African-language dictionaries, Part 2: The microstructure. *South African Journal of African Languages* 20(4): 310–30.

de Schryver, Gilles-Maurice & D.J. Prinsloo. 2001. Towards a sound lemmatisation strategy for the Bantu verb through the use of frequency-based tail slots – with special reference to Cilubà, Sepedi and Kiswahili. In: Mdee, James S. & Hermas J.M. Mwansoko (eds). *Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings*: 216–42, 372. Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam.

eLex. 2017. *Electronic Lexicography in the 21st Century*. Available online at: https://elex.link/elex2017/.

Hanks, Patrick. 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398–436.

Hurskainen, Arvi. 2004. Computational testing of five Swahili dictionaries. In: Karlsson, Fred (ed.). *Proceedings of the 20th Scandinavian Conference of Linguistics, Helsinki, January 7–9, 2004* (Department of General Linguistics Publications 36). Helsinki: University of Helsinki.

Johnson, Frederick. 1939. *A Standard Swahili-English Dictionary (Founded on Madan's Swahili-English dictionary)*. Nairobi: Oxford University Press.

Lew, Robert. 2004. *Which Dictionary for Whom? Receptive use of bilingual, monolingual and semi-bilingual dictionaries by Polish learners of English*. Poznań: Motivex.

Mohamed, Amir A. 2009. *Kiswahili for Foreigners* [3rd revised edition]. Zanzibar: Goodluck Publishers.

Tono, Yukio. 2009. Pocket Electronic Dictionaries in Japan: User Perspectives. In: Bergenholtz, Henning, Sandro Nielsen & Sven Tarp (eds). *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow* (Linguistic Insights): 33–67. Bern: Peter Lang.