Lexicographic treatment of salient features and challenges in the creation of paper and electronic dictionaries

Prinsloo D.

Department of African Languages, University of Pretoria
danie.prinsloo@up.ac.za

Abstract

This paper focuses on the need for lexicographers to study and to treat the salient features of languages satisfactorily and the challenges faced by lexicographers. The focus is on the challenges facing compilers of African language dictionaries and the lack of dictionaries for these languages. It will be argued that lexicographers are expected to fulfil the role of mediators between complicated grammatical structures, on the one hand, and the target users' needs and expectations, on the other. Dictionaries are expected to be inclusive, e.g., providing for and fulfilling user expectations by giving all the required information in the dictionary in order to reduce the need for consultation of external sources. Expectations for future compilation of paper and electronic dictionaries are discussed. It is expected that paper dictionaries will be used in Africa for many years to come but that paper and electronic dictionaries of high lexicographic quality should be compiled simultaneously. The discussion is presented against the background of the transition of African lexicography from Euro-centred dictionary compilation to Afro-centric compilation. African language dictionaries are continuously compiled in Africa, by Africans for Africans.

Keywords: dictionaries; lexicographic treatment; salient features; challenges; African languages

1 Introduction

Lexicographers must make sure that the salient features of the language or languages treated in their dictionaries are well studied and comprehended by themselves before embarking on the arduous task of lemmatising and treating them. What could be an issue in a specific language might be non-problematic and straight forward in another, or in the other member of the language pair in a bilingual dictionary. Lexicographers, although being mother-tongue speakers of the language(s) treated in their dictionaries, should never assume full knowledge of all the salient features of these languages. Many examples of salient features that were missed in dictionaries were detected. In addition to in-depth knowledge of the grammar of the language(s), the lexicographer should also consider all the relevant external issues and challenges impacting on the compilation of the dictionary. A number of specific lexicographic initiatives in Africa involving community engagement will be discussed.

Lexicographers, unfortunately, do not live in an ideal world. They are faced by a multitude of challenges. In this presentation, African languages, specifically the Bantu language family,[1] will be considered as a case in point, i.e., how their salient features should be detected and treated in terms of the intrinsic challenges pertaining to the language(s) as well as how the challenges posed by the environment or setting that the dictionary has to be compiled within should be handled. Challenges pertaining to the language regard, e.g., morphology, syntax, semantics, and pronunciation, as well as compilation traditions and extra-linguistic factors, such as financial and political issues. The aim of this paper is to give an overview of the salient features and main challenges of dictionary compilation for these languages. The aim is neither to provide a mere listing of problematic issues nor to attempt detailed discussion within the limitations of a conference paper. References to resources where the key issues are discussed in more detail will be given for the interested reader. The main focus will be on the impact of these challenges on dictionary compilation for African languages and to suggest best practices for the lexicographer in order to meet them.

2 The Status of Dictionaries and Lexicographic Initiatives

A study on lexicography in Africa, edited by Hartmann (1990), is taken as a point of departure. Thirty years ago, he

[1] The term 'Bantu' got stigmatized during the Apartheid Era in South Africa. Therefore the term 'African' is preferred in South Africa even in reference to what is internationally referred to as 'Bantu languages'. The discussion in this paper is however focused on the Bantu language family and most of the issues described cannot necessarily be generalized to be applicable to other languages on the continent of Africa. To respect the view of those opposed to the term 'Bantu', it will only be used in cases where a distinction between African languages (languages spoken in Africa) versus a member of the Bantu language family is essential.

conducted a study of lexicography in different regions of Africa, e.g., East, West, South, etc. The opinions echoed by the researchers were that dictionaries for African languages were not of a high lexicographic quality. The main reason given was that existing dictionaries reflected a Euro-centric approach. They were compiled mostly by missionaries from abroad to fulfil their goals, i.e., to assist the missionaries to understand African languages in order to spread the gospel. Such dictionaries were, therefore, not in the first place intended to serve the needs of Africans. In the past decade, the need for dictionaries compiled in Africa by Africans themselves, primarily for speakers of African languages as target users gained momentum, which can thus be called an "Afro-centric approach" to dictionary compilation. Portraying European/western culture instead of African reality is a typical shortcoming in many dictionaries. So, for example, Taljard and Prinsloo (2019: 210) quote an instance where the concept "my house" is represented by an illustration that is unmistakably a typical European dwelling instead of a variant of the houses seen in Africa. In the past decade, the move to Afro-centric dictionary compilation coincided with the decolonisation drive. Several initiatives by entrepreneurs, publishing houses and government-supported agencies were undertaken to give wings to dictionary compilation with a true Afro-centric approach such as IKS (Institute of Kiswahili Studies, https://www.udsm.ac.tz/web/index.php/institutes/iks/the-history-of-the-institute), the Allex Project (http://www.edd.uio.no/allex/aims.html), and the nine national lexicographic units (NLUs) for African languages in South Africa. The requirement for NLUs was the compilation of comprehensive monolingual dictionaries for these languages, which was funded by government. The expectation was that the speech communities of the different languages will eventually take full responsibility for dictionary compilation, including providing the necessary financial resources. There were difficulties faced by these lexicographic initiatives. So, for example, Wolvaardt (2017) reports negatively on the actions of the Pan South African Language Board (PanSALB) responsible for funding and guidance of the South African NLUs.

> […] the national lexicography project, pioneered in the early years of South Africa's democratic transition by some of the country's greatest language activists and academics, […] permitted to degenerate into the scattered efforts of a diminishing band of lexicographers? […] into perpetual begging for adequate funding, the National Lexicography Units (NLUs) hover on the verge of extinction. […] leaves the NLUs where we find them today, desperately trying to justify their existence by producing dictionaries, which, by and large, are based on their feasibility within the constraints of limited funding rather than on any coherent overarching plan. (Wolvaardt 2017: 9)

Financial support from speech communities also did not materialise. The NLUs still rely on PanSALB.
A degree of community engagement, which can be compared to crowdsourcing, materialised where speakers of the language contribute to the extension and updating of the dictionary, e.g., the Xitsonga-English dictionary (https://www.xitsonga.org/dictionary) where the community is involved in correcting dictionary information. Another good example of dedicated community involvement is the Ju|'hoan Children's Picture Dictionary (Jones & Cwi 2014a). In its self-description (Jones & Cwi 2014b), the compilation of this dictionary is described as a collaborative project between the Namibian Ju|'hoan from the Tsumkwe region and academics from various fields. This dictionary clearly indicates an Afro-centric approach to dictionary compilation.
Financial aspects and the fact that African languages are severely under-resourced constitute a major problem in many ways for the lexicographer. In an overview by the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL, 2014), it is stated that "under-resourced languages suffer from a chronic lack of available resources (human-, financial-, time- and data-wise)." This is absolutely applicable to African languages: Compilers of dictionaries for African languages are severely limited in the number of lemmas that can be treated, on the exhaustiveness of treatment of these lemmas, and on the number of pages allowed. This leads to the undesirable situation where the lexicographer must choose between lemmatising, say, 15,000 lemmas with the treatment limited to a few translation equivalents, or 5,000 lemmas with slightly elaborated treatment, still barely fulfilling the need for text reception or dictionary use on demand. See the discussion below regarding dictionaries for text production.

3 Dictionary Compilation for Specific Target Users

Prospective compilers are faced with a situation where dictionaries are required for several thousands of African languages spoken on the content of Africa. Many of these languages do not even have a single dictionary as a reference source. Therefore, the first challenge is to compile, say, a monolingual and a bilingual dictionary for the specific language. Bilingual dictionaries in Africa usually bridge the African language with major languages of the world such as English and French. Lexicographers could depart from the revision of existing dictionaries, where available, or opt for starting afresh with a new compilation.

4 Introspective versus Corpus-based Dictionary Compilation

The advantages and disadvantages of introspective versus corpus-based dictionaries should be carefully considered. It is generally accepted that the utilisation of a corpus can enhance the lexicographic quality of a dictionary on both macrostructural and microstructural levels. In the absence of a corpus, which is the case for most African languages, lexicographers have no option but to compile the dictionary on introspection. The downside of introspective compilation is that words most likely to be looked for by the target users can easily be left out simply because, in the words of Snyman et al. (1990) in *Dikišinare ya Setswana English Afrikaans Dictionary* (DS), they "did not cross the compilers' way". Studies by De Schryver and Prinsloo (2000a) indeed reveal many instances of lemmas most likely to be looked up which

are simply not in the dictionary. However, De Schryver and Prinsloo (2000b) also indicate that consistent application of introspection over time can render good quality lemmalists. For many African languages, it is possible to compile corpora albeit relatively small ones, e.g., comparable in size to initial English corpora such as the Brown Corpus consisting of only one million words. Prinsloo (2015) indicated that even such limited corpora can go a long way in assisting the lexicographer with lemmatisation, sense distinction selection of authentic examples, frequency indication in the dictionary, etc.

5 Words most Likely to be Looked for and User Expectations

The importance of the user perspective has been echoed several times in the literature emphasising the basic fact that dictionaries are judged as good or bad by their users, cf., Gouws and Prinsloo (2005). Many African language dictionaries can be regarded as examples of linguistic achievement but are not user-friendly. Haas' (1962: 48) remark is still relevant after six decades: "a good dictionary is one in which you can find the information you are looking for — preferably in the very first place you look"; likewise, Barnhart (1962: 161) states that "the function of a popular dictionary [is] to answer the questions that the user of the dictionary asks".

Ideally, dictionaries should be compiled for very specific target users, but when the first dictionary for a language is compiled, the only option for the African language lexicographer is to compile a dictionary that can be used by all users, i.e., an unfortunate attempt towards a one-size-fits-all dictionary. In the compilation of such general dictionaries, the lexicographer should maintain a sound balance between descriptiveness and prescriptiveness. On the one hand, prescriptiveness is required, especially in cases where the language is not fully standardised; on the other hand, the lexicographer should guard against excessive purism, e.g., resisting pressure not to enter any loan words in the dictionary.

6 The Lexicographer as Mediator between the User and Complicated Grammatical Systems

African language lexicographers find themselves in the role of mediators between user expectations and complicated grammatical structures. It is not claimed that African languages are the only languages in the world with complicated grammatical systems; the point is that the lexicographer should be fully acquainted with the core grammatical systems in the language(s) treated. For members of the Bantu language family in particular, these core systems are complicated nominal and verbal systems. Nouns are classified into different classes, each generating different sets of concords and pronouns, which are not interchangeable and are elements required to complete sentences and phrases. Verbs occur in eight moods, see Tables 1 and 2 as well as Prinsloo (2020a and 2020b) for a detailed discussion.

Person or noun class	**Example**	**Cp.**	**Sc. 1**	**Sc. 2**	**Oc.**	**Dem.**	**Poss.**	**Ep.**
2nd Person sing.	wena 'you' (singular)		o	wa	go			
2nd Person plural	lena 'you' (plural)		le	la	le			
Class 3	molato 'problem'	mo	o	wa	o	wo	wa	wona
Class 4	melato 'problems'	me	e	ya	e	ye	ya	yona
Class 14	bothata 'difficulty'	bo	bo	bja	bo	bjo	bja	bjona
Class 15	go gopola 'to think'	go	go	gwa	go		ga	gona

Key: Cp. = class prefixes of the noun; Sc. = subject concords; Oc. = object concords; Dem. = demonstratives; Poss. = possessive concords; Ep. = emphatic pronouns

Table 1: Extract from the noun class system in Sepedi.

Mood	Positive	Negative
Relative		
Present	**subject concord + verb stem + go** *Kgoši ye e balago melao* 'The king who is reading the laws'	**subject concord + *sa* + verb stem** ending **-e + go** *Kgoši ye e sa balego melao* 'The king who is not reading the laws'
Future	**subject concord + *tlo* + go + verb stem** *Kgoši ye e tlogo bala melao* 'The king who will be reading the laws'	**subject concord + *ka se* + verb stem** ending **-e + go** *Kgoši ye e ka se balego melao* 'The king who will not be reading the laws'
Past	**subject concord + verb stem + go** *Kgoši ye e badilego melao* 'The king who read the laws'	**subject concord + *sa* + verb stem + go** *Kgoši ye e sa balago melao* 'The king who did not read the laws'
Hortative	**subject concord + verb stem** ending **-e** *Kgoši e bale melao* 'The king usually reads the laws'	**subject concord + *se* + verb stem** ending **-e** *Kgoši e se bale melao* 'The king usually does not read the laws'

Table 2: Extract from the verbal mood system in Sepedi.

Lexicographers should serve the users with lexicographic inclusiveness and present the information in such a way that users can find what they are looking for in and what they need to understand from the dictionary. Users should not be obliged to consult external sources, such as grammatical descriptions of the language, which in most cases do not exist anyway. Of specific importance here is the work of Gilles-Maurice de Schryver (2010) in which he attempts to "revolutionize African language lexicography" as well as the publication of the *Oxford Bilingual School Dictionary: Zulu and English* (OZSD) in which the stem tradition for the lemmatisation of nouns was abandoned — nouns were lemmatised as full orthographic words.

7 Paper versus Electronic Dictionaries

Naturally, dictionary compilation of African languages does not stand in isolation — it is influenced by trends and changes in international lexicography. So, for example, the need to compile and consult corpora became an important and desired aspect of dictionary compilation in Africa. Likewise, the dawn of the electronic era brought new opportunities but also new challenges. Most significant is the resolution of stem identification problems in lemmatisation for conjunctively written African languages in electronic dictionaries (see also below). A major challenge, however, is producing good paper and electronic dictionaries simultaneously for African languages. An extreme approach could be to stick to the compilation of only paper dictionaries until paper dictionaries of high lexicographic quality are available for most African languages. The other extreme is to discontinue paper dictionary compilation and disregard lexicographic traditions and approaches in order to focus only on electronic dictionaries. Such an approach would be in line with the decision announced by Michael Rundell in 2012 that Macmillan decided to discontinue printed dictionaries. Rundell (2012: 74) even said "in an ideal world, we would pulp most of this and start from scratch, producing new resources optimally adapted to digital media". Starting afresh was indeed tempting given the African language lexicographic situation. Such a decision would also "free" the lexicographer from the many restraints and misinterpretations about lemmatisation strategies and alphabetical ordering. One of the biggest frustrations to compilers of dictionaries for African languages is the misconception that stem lemmatisation is more scientific than word lemmatisation. Blindly following this belief resulted in situations where the stem tradition was also followed for languages in which words are disjunctively written, ideal for full-word lemmatisation and for instances in which neither the lexicographer nor the user knows what the stem of the noun is in order to look it up. See Van Wyk (1995) for a detailed discussion. The lexicographer ends up in a minefield of lemmatisation approaches, conjunctive versus disjunctive writing systems, and lexicographic traditions.

Rundell (2015: 303) believes that "in many parts of the world, paper dictionaries still have a healthy future ahead of them. He says that "certain types of dictionary — such as those designed for schools, or special-subject dictionaries, or dictionaries of "smaller" languages — may show a preference for print for some time to come". The reality in Africa is indeed that paper dictionaries are expected to be relevant for many years to come. Phillip Louw, Head of Dictionaries and Dictionary Data, Oxford University Press, Cape Town, South Africa (in email correspondence) emphasises that "the dominance of paper dictionaries in the school dictionary market in Africa [is expected] to continue for at least the next ten years".

Thus, when it comes to paper dictionaries versus electronic dictionaries, the recommended approach would rather be for African language lexicographers to persevere with the improvement and compilation of paper dictionaries but also to embark on the compilation of electronic dictionaries for African languages. They should, however, be careful to avoid the typical pitfalls international lexicography fell prey to in the transition from paper dictionaries to electronic dictionaries. These pitfalls mainly revolve around the presentation of paper dictionaries "on computer" with perhaps only a few added electronic features, such as search functions. Sue Atkins (1996: 515-516) is quite adamant on this issue and bluntly states:

> […] dictionaries of the present […] may even come to you on a CD-ROM rather than in book form, but underneath these superficial modernizations lurks the same old dictionary. […] It is up to us to take up the real challenge of the computer age, by asking not how the computer can help us to produce old-style dictionaries better,

but how it can help us to create something new.

The important aspect to realise is that it is a new process in which the features enabled by the computer should be maximally utilised. Such features can be called "true electronic features". Gouws and Tarp (2017: 391) list the following important features applicable to all e-dictionaries:

- Improved search methods and access routes;
- User-based data filtering;
- Less compact article formats with items representing different data categories placed in separate lines;
- Abolition of abbreviations;
- Use of metatexts to introduce sections with specific data categories;
- Use of hidden data, that is data that are not always on display but can be called up when needed;
- Use of pop-up windows and hypermedia to present additional data;
- Inclusion of video and audio options;
- New forms of internal and external linking;
- Interaction between lexicographer and user;
- Continuous updating.

In the same way as OZSD, electronic dictionary designs for Sepedi should include all the required information in the dictionary without the user having to consult external sources. This is, among other things, obtained through a network of pop-up information activated by hovering over or clicking on items for which the user needs more information. Consider a summary of hovering and clicking options available to the user when consulting the Sepedi multi-word lemma *ka se* in Figure 1.

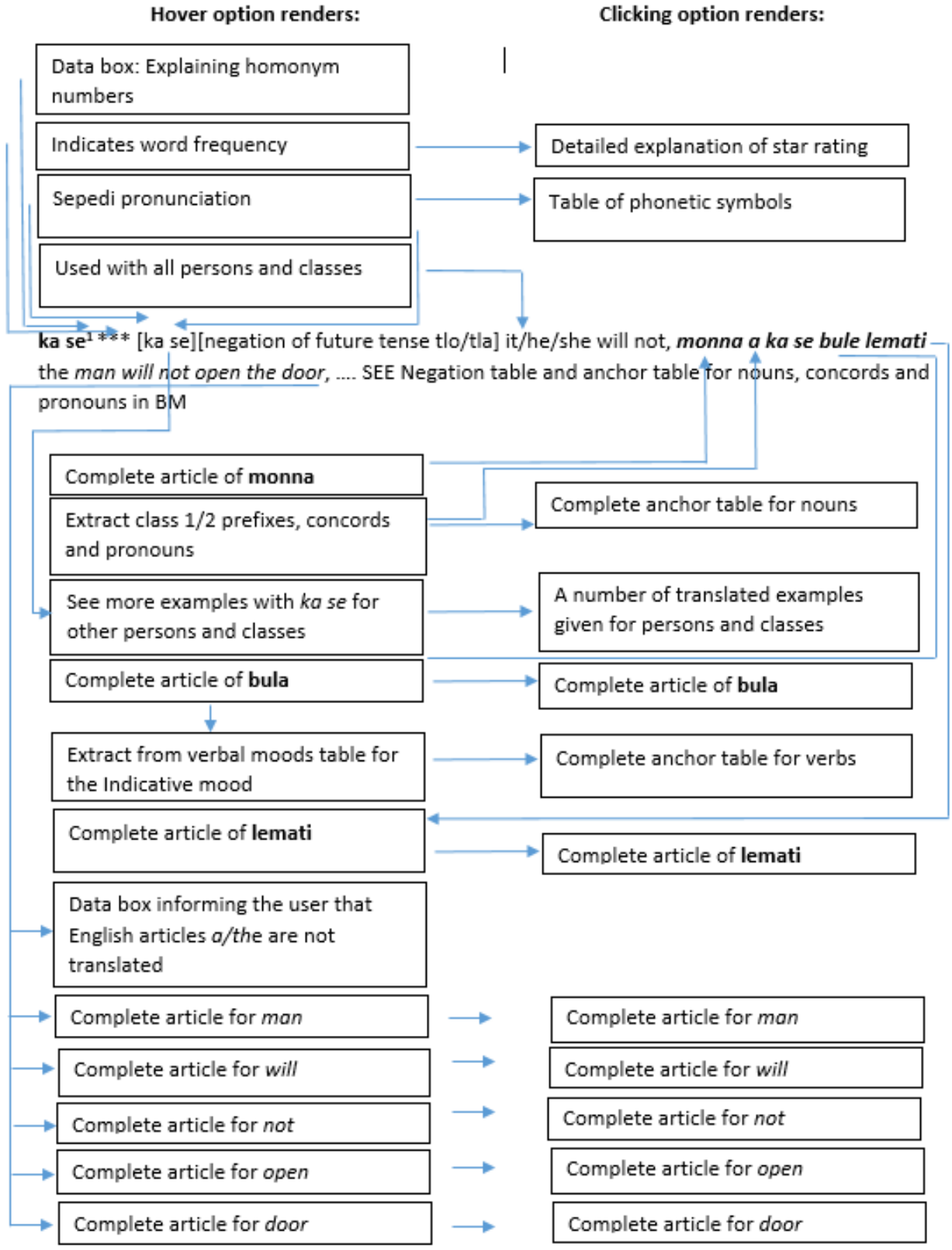

Figure 1: A design for Sepedi indicating pop-up information obtained through hovering and clicking.

8 Dictionaries Suitable for Text Production

One of the major shortcomings in African language lexicography is the lack of dictionaries for guidance in "text production" situations. Most dictionaries barely fulfil the needs of the users for decoding purposes or the use of dictionaries "on demand". So, for example, negation is a complicated issue in Bantu languages — many negation strategies, e.g., *ga*, *sa*, *se*, *ga se* and *ka se*, are distinguished by Prinsloo (2020b) for Sepedi. These strategies are complicated and non-interchangeable. Most dictionaries do not even fulfil the most basic receptive needs of users, not to mention a lack of guidance on when which negation morpheme can be used. Examples of efforts towards giving guidance

in productive dictionary use for Sepedi include a variety of support tools that can be linked to a dictionary such as an assistant for the compilation of isiZulu possessives (Bosch & Faasz 2014) and a sentence constructor, the *Sepedi Helper* for Sepedi.

9 Conclusion

In this paper, it was attempted to give, within the limits of only a few pages, an overview of aspects of the lexicographic treatment of salient features in paper and electronic dictionaries focusing on African language lexicography. This was done in the context of the many challenges faced by the lexicographer in the compilation of dictionaries for African languages.

10 References

Atkins, B.T.S. (1996). Bilingual Dictionaries: Past, Present and Future. In M. Gellerstam, J. Järborg, M. Sven-Göran et al. (eds.) *Euralex '96 Proceedings: Papers Submitted to the Seventh Euralex International Congress on Lexicography*, Göteborg University, pp. 515-546. Göteborg, Sweden.

Barnhart, C.L. (1962). Problems in Editing Commercial Monolingual Dictionaries. In F.W. Householder, S. Saporta (eds.) *Problems in Lexicography,* pp. 161-181. University of Indiana, Bloomington.

Bosch, S.E., Faasz, G. (2014). Towards an Integrated E-Dictionary Application — The Case of an English to Zulu Dictionary of Possessives. In A. Abel, C. Vettori, N. Ralli (eds.) *Proceedings of the 16th Euralex International Congress: The User in Focus 15-19th July 2014*, pp. 739-747. Bolzano, Italy.

CCURL. (2014). Proceedings overview: *Workshop on collaboration and computing for under-resourced languages in the Linked Open Data Era*. Accessed at: http://www.ilc.cnr.it/ccurl2014/ [01/06/2016].

De Schryver, G.M. (2010). Revolutionizing African language lexicography — a Zulu case study. In *Lexikos 20*, pp. 161-201.

De Schryver, G.M., Prinsloo, D.J. (2000a). Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The macrostructure. In *South African Journal of African Languages,* 20(4), pp. 290-309.

De Schryver, G.M., Prinsloo, D.J. (2000b). (In)consistencies and the Miraculous Consistency Ratio of '(x 1.25)4 = x 2.44', A perspective on corpus-based versus non-corpus-based lemma-sign lists. In *Fifth International Conference of AFRILEX*, *July 2000*. University of Stellenbosch.

(DS) Snyman, J.W., Shole, J.S. & Le Roux, J.C. (1990). *Dikišinare ya Setswana English Afrikaans Dictionary*. Pretoria: Via Afrika.

Gouws, R.H., Prinsloo, D.J. (2005). *Principles and practice of South African lexicography*. Stellenbosch: African Sun Media.

Gouws, R.H., Tarp, S. (2017). Information Overload and Data Overload in Lexicography. In *International Journal of Lexicography,* 30(4), pp. 389-415.

Haas, M.R. (1962). What belongs in the bilingual dictionary? In F.W. Householder, S. Saporta, (eds.) *Problems in Lexicography*, pp. 45-50. University of Indiana, Bloomington.

Hartmann, R.R.K. (ed.). **(**1990). *Lexicography in Africa. Progress Reports from the Dictionary Research Centre Workshop at Exeter, 24-26 March 1989: Exeter Linguistic Studies 15*. Exeter: University of Exeter Press.

Jones, K., Cwi, T.F. (2014a). *Ju|'hoan children's picture dictionary. Interactive disc-gallery*. Pietermaritzburg: University of KwaZulu-Natal Press.

Jones, K., Cwi, T.F. (2014b). *Ju|'hoan children's picture dictionary. Information leaflet*. Pietermaritzburg: University of KwaZulu-Natal Press.

(OZSD) De Schryver, G.M. (ed.). (2010). *Oxford Bilingual School Dictionary: Zulu and English / Isi-chazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgisi*. OUP Southern Africa, Cape Town.

Prinsloo, D.J. (2015). Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited corpus. In *Lexikos,* 25*,* pp. 285-300.

Prinsloo, D.J. (2020a). Detection and lexicographic treatment of salient features in e-dictionaries for African languages. In *International Journal of Lexicography,* 33(3), pp. 269-287.

Prinsloo, D.J. (2020b). Lexicographic treatment of negation in Sepedi Paper dictionaries. In *Lexikos,* 30*,* pp. 321-345.

Rundell, M. (2012). It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical. In R.V. Fjeld, J.M. Torjusen, (eds.) *Proceedings of the 15th Euralex International Congress, 7-11 August 2012*, Oslo, pp. 47-92.

Rundell, M. (2015). From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. In *Lexikos,* 25, pp. 301-322.

Sepedi Helper, Accessed at: (http://sepedihelper.co.za) [20/08/2020].

Taljard, E., Prinsloo, D.J. (2019). African Language Dictionaries for Children — A Neglected Genre. In *Lexikos,* 29, pp. 199-223.

Van Wyk, E.B. (1995). Linguistic assumptions and lexicographical traditions in the African languages. In *Lexikos,* 5, pp. 82-96.

Wolvaardt, J. (2017). South Africa's National Lexicography Units: time for a reboot? In *2nd International Conference of the African Association for Lexicography (Afrilex)*, pp. 9-10. Rhodes University, Grahamstown.

Acknowledgement

This research is supported in part by the South African Centre for Digital Language Resources (SADiLaR). The Grantholder acknowledges that opinions, findings and conclusions or recommendations expressed are those of the authors.