User Support in e-Dictionaries for Complex Grammatical Structures in the Bantu Languages

Danie J. Prinsloo*, Theo J.D. Bothma*, Ulrich Heid**
*University of Pretoria, **University of Hildesheim
danie.prinsloo@up.ac.za, theo.bothma@up.ac.za, heid@uni-hildesheim.de

Abstract

This paper discusses direct user guidance as a mechanism in an e-dictionary to provide user support for complex grammatical structures in the Bantu languages. We present a design study to show that user support through direct user guidance can provide solutions in the case of complex concordial relationships between nouns and pronouns. The compilation of the complex relative construction is taken as a case in point. The concept of user support appropriately puts the user in focus. Our approach to user support also caters for the casual, on-the-fly user, who is not interested or in a position to devote time to the learning of a foreign language.

Keywords: e-Dictionaries; Bantu languages; complex grammatical structures; user support; the relative construction; Northern Sotho

1   Introduction

In the Bantu languages, there are many grammatical constructions that are insufficiently treated in current dictionaries because of the complexity of the constructions. Other solutions need to be designed as an integral part of a dictionary and additional levels of user support are required within the dictionary. Such support should be available to the dictionary user "on demand", and different options can be available for a specific information need. The nature of the support could typically also link to a user's level of knowledge of the grammatical system of the language. A user with a very limited knowledge of the language or a casual user, for example, may prefer a machine translation option in the dictionary, with links to the grammar rules which may be consulted on demand. On the other hand, a user who has a fair knowledge of the language may require a different type of support, e.g. through *inter alia* decision trees, structured paths or direct user guidance. Such technologies, integrated in the dictionary, may enable the user to find the correct information at exactly the right level of detail and complexity (s)he requires to solve his/her information need (cf. Bothma 2011).

The innovative use of decision trees and structured paths as tools to support users have been dealt with in some detail in Prinsloo et al. (2011, 2012) and it has been shown that these solutions can provide significant decision support to users for complex text production situations such as copulative constructions and kinship terminology in Northern Sotho. The purpose of such tools is to guide users

to the information they are looking for, i.e. without having to first study complicated grammatical structures in order to find the required information. This guidance process is done through decision trees (i.e. a series of basic choices made by the user) or through structured paths (e.g. visually linking kinship terms in a schematic illustration of a family tree) as discussed in Prinsloo et al. (2011, 2012) or through direct user guidance, as discussed in this paper.

Direct user guidance as an additional technique in the dictionary to provide user support for complex grammatical structures in the Bantu languages is not a solution for all user support. We regard it as a complementary technology that may be used in conjunction with other user support technologies for specific grammatical constructions, available to the user on demand, depending on the user's level of language knowledge, the nature of the information need and choice of support tool. We present a design study to show that user support through direct user guidance can provide solutions in the case of complex concordial relationships between nouns and pronouns. In terms of the Function Theory of Lexicography (Tarp 2008, Bothma and Tarp 2012) the design provides for text production, text reception and cognitive information needs. No case studies or user evaluation of these techniques has been done to date, as we feel that it is important to first define a range of techniques and the range of complex grammatical structures where such techniques could offer relevant user support before any serious implementation in actual real world scenarios would be warranted. This does not mean that small scale prototypes of individual techniques should not be developed to establish the technological feasibility of such techniques. However, to do proper usability studies on such prototypes that are not fully integrated into a full dictionary will have only limited value, as it will not be possible to determine whether (or to what extent) users would use such techniques in real world situations as an integral part of dictionary use. As will be clear from the discussion below and from Prinsloo et al. (2011, 2012), such techniques are made available "on demand", i.e., users are not forced to use them if they feel that their information needs have been solved by the "standard" dictionary article. In every case, the use of such a technique is therefore a conscious choice of the user to find more information or information that is easier to use / digest / apply than the information available in the dictionary, the outer text of the dictionary or other reference tools such as grammar books that the user may have available.

The importance of the user perspective as the main thrust in the compilation of modern dictionaries has been emphasized in numerous publications, e.g., Gouws and Prinsloo (2005), Tarp (2008, 2011, 2012). The concept of user support appropriately puts the user in focus. Compare Tarp's (2012:253) idea of individualization when he refers to "quicker, more accurate and personalized satisfaction of the corresponding user needs". Our approach to user support furthermore does not necessarily put the user into a specific category (e.g., as a learner of the language): it is not profile-based and does not assume that the user will be interested to study a complete grammatical paradigm before being able to produce (or understand) text and speech. We therefore also cater for the casual, on-the-fly user, who is not interested or in a position to devote time to the learning of a foreign language.

2 Phenomena and Data: Grammatical Distinctions as a Problem for Bantu Lexicographers

2.1 The Notion of Grammatical Distinctions

Due to the richness of grammatical distinctions, a given grammatical property may be expressed in many different forms. For example, there are different equivalents for a pronoun such as he in Bantu, determined by the grammatical class of the noun. Nouns in Bantu languages are subdivided into different noun classes and these classes have their own sets of, e.g., subject concords and object concords, as well as different sets of pronouns such as demonstrative, possessive, emphatic and quantitative. This means that in Northern Sotho a basic English pronoun such as he can be expressed by up to ten different subject concords, a form like him by ten object concords and more than 20 pronominal forms. Consider table 1 which distinguishes 15 different noun classes each having their own subject concords (Sc.); object concords (Oc.); demonstratives (Dem.); possessive concords (Poss.); emphatic pronouns (Ep.) and quantitative pronouns (Qp.).

Person or noun class	Example	Sc.	Oc.	Dem.	Poss.	Ep.	Qp.
1st Person singular	nna 'I'	ke	n-				
1st Person plural	rena 'we'	re	re				
2nd Person sing.	wena 'you' (singular)	o	go				
2nd Person plural	lena 'you' (plural)	le	le				
Class 1	monna 'man'	o/a	mo	yo	wa	yena	yohle
Class 2	banna 'men'	ba	ba	ba	ba	bona	bohle
Class 3	molato 'trouble, problem'	o	o	wo	wa	wona	wohle
Class 4	melato 'problems'	e	e	ye	ya	yona	yohle
Class 5	lesogana 'young man'	le	le	le	la	lona	lohle
Class 6	masogana 'young men'	a	a	a	a	ona	ohle
Class 7	selo 'object, thing'	se	se	se	sa	sona	sohle
Class 8	dilo 'objects, things'	di	di	tše	tša	tšona	tšohle
Class 9	ntlo 'hut'	e	e	ye	ya	yona	yohle
Class 10	dintlo 'huts'	di	di	tše	tša	tšona	tšohle
Class 14	bogobe 'porridge'	bo	bo	bjo	bja	bjona	bjohle
Class 15	go reka 'to buy'	go	go		ga		
Class 16	fase 'below'			fa			
Class 17	godimo 'above'	go	go		ga	gona	gohle
Class 18	morago 'behind'			mo			

Table 1: The noun class system of Northern Sotho with a few sets of concords and pronouns.

In table 1 the demonstrative 'this' varies depending on the class of the noun, e.g., class 1: *monna* ***yo*** 'this man' but class 14: *bogobe* ***bjo*** 'this porridge'. Likewise, the possessive 'of' differs for each class, e.g., class 1: *mosadi* ***wa*** *monna* 'wife of the man' but class 2: *basadi* ***ba*** *monna* 'wives of the man'. Concords and pronouns representing subjects and objects also vary according to the nominal class, e.g.:

(1) O e bone 'He saw it'

o (e.g. *monna* class 1)	*e* (e.g. *tau* class 9)	*bone*
he (the man)	*it* (the lion)	saw

2.2 Grammatical Distinctions in the Sentence Context

In table 1 the grammatical distinctions paradigm is mono-dimensional in the sense that it is always given for a single source language item which diverges into a single set of equivalents. More than one instance of grammatical distinction can, however, co-occur in a single construction or phrase:

A single occurrence

Example 1: *he*, as the subject of a sentence (subject concords):

(2) *O/a/le/se/e*   thušitše mosadi.

   He   helped the woman

Example 2: how to express *all* (quantitative pronouns):

(3) Go bolaya   *bohle/yohle/ohle/tšohle*

   To kill   *all*

Two occurrences: he as a subject and them as an object:

What is at stake here is direct guidance in terms of the simultaneous handling of subject and object concords:

(4) *O/a/le/se/e*	tlo	*ba/e/a/di*	thuša
He	will	them	help

He will help *them.*

Three occurrences (the verbal relative): he as a subject, and as a demonstrative and them as an object:

(5) *Yo/wo/le/se/ye*	*a/wo/le/se/e*	*ba/e/a/di*	thušitšego
He	he	them	helped

He who helped *them.*

3   Direct Guidance for Concords and Pronouns

Guidance is given by means of three possible access points depending on the user's need in terms of text production (access points 1 and 2) or text reception (access point 3), and his/her knowledge of the language:

- **Access point 1:** Step by step guidance: build your own Northern Sotho sentence/construction.
- **Access point 2:** The user enters an English phrase and the system then assists him/her in a step-by-step build-up process of the Northern Sotho construction.
- **Access point 3:** The user enters a Northern Sotho phrase and the software analyses it.

Utilising *Access point 1* simply requires the user to enter the Northern Sotho noun and the software will suggest the correct pronoun and subject/object concord from table 1. Where more than one option is applicable, the user has to select the correct one or utilise clickable help functions to guide him/her to the correct option or (s)he can revert to the *Access point 2* option.

Taking *Access point 2* as departure, the user can type in "the man is walking" and the software will return the noun (*monna*) + the subject concord (*o*) + the present tense marker (*a*) + the verb *sepela*, guiding the user to build *monna o a sepela*. Clickable help functions and 'more information' boxes are also provided.

Entering a Northern Sotho sentence from Access point 3 will result in the reverse process, e.g. *monna* 'man'(noun) + *o* 'he' (subject concord) + *a* (present tense marker) + *sepela* 'walk' (verb).

The full set of necessary data, for these cases, is thus as follows: In the case of single and two occurrences given in 2.2, subject and object pronouns/concords are independent from each other, and their choice is only conditioned by the noun class of their antecedents.

To be able to provide the above mentioned kinds of guidance for single and two-occurrence of grammatical distinctions a word (token) list tagged for part of speech for nouns, verbs, subject concords, object concords and pronouns, and a basic bilingual dictionary for word forms are required.

To be able to provide guidance for cases like example (5) more than just word lists are needed as an internal knowledge source for the guidance tool: the agreement between subject and relative demonstrative must be encoded as well.

To guide users in the creation of the relative construction, also the morpho-syntactic structure of this construction must be explained.

The verbal relative case (three occurrences) is different in so far as the relative demonstrative ("who") is in grammatical agreement with the subject. In addition to the requirements for single and two-occurrences cases, a basic five-element formation rule for the verbal relative: noun + demonstrative + subject concord + verb + relative suffix (-go) is required.

4   Example of Direct Guidance for the Verbal Relative

4.1   User Support for Text Production

Access point 1: This provides step by step guidance on how to build your own relative construction. The user with a basic knowledge of the grammatical system would like to express "the man who loves her". (S)he knows the different nouns and verbs in Northern Sotho but needs guidance in terms of the concordial system. In this case the user consults the article for "who" in an English to Northern Sotho dictionary, selects the button "Build your own relative construction" and types the Northern Sotho word for "man" = *monna*.

The system subsequently suggests the relevant concords from table 1. In the consultation, the relevant section for the appropriate noun class is highlighted while being placed in context within the grammar table. For cognitive use, clickable options to see larger portions of the table are provided; cf. table 2.

Person or noun class	Example	Dem.	Sc.		Oc.
		This	He/she		Him/her
Class 1	monna *'man'*	yo	o	a	mo
Class 2	banna 'men'	ba	ba	ba	ba
Class 3	molato 'trouble'	wo	o	o	o
Class 4	melato 'problems'	ye	e	e	e
Class 5	lesogana 'young man'	le	le	le	le
Class ...					
		Click to add	Click to add	Click to add	Click to add
		More information	More information	More information	More information
Click for full list	Click for full list	Click for full list	Click for full list		Click for full list

Table 2: The noun class 1 of Northern Sotho highlighted for selection of the correct concords.

Based on his/her knowledge of the relative construction (noun + demonstrative + subject concord + verb + relative suffix (-go)), the user can now build the full relative construction, to arrive at the full phrase *monna yo a mo ratago.* If the user is knowledgeable about the subject concords of class 1 in the relative, (s)he will click the subject concord *a* directly. If the user nevertheless needs more support, e.g., choosing between the subject concords *o* or *a*, the cursor could be momentarily rested on any of the "more information" boxes in the table triggering a pop-up box to support him or her in the selection task.

The user who needs full support can type a complete relative phrase in either English or Northern Sotho as portrayed in access points 2 and 3.

Access point 2: The user enters an English phrase: "The man who loves her", similarly to access point 1 in the dictionary. The system then assists the user in a step-by-step build-up process of the relative construction:

(i)   *the man: the tool provides the correct equivalent from the dictionary, i.e. monna tagged for part of speech as N01 (noun of class 1, cf. table 1);*

(ii) *who: keeping the agreement constraint from the sentence formation rule (noun + demonstrative + subject concord, + verb + relative suffix (-go) in 4.1), the tool extracts the demonstrative for class 1 from the closed-class list of demonstratives, i.e. yo;*

(iii) *(subject concord): The insertion of the SC is coded in the rule for relatives: it requires, in addition to the demonstrative in (ii), the subject concord for the noun in (i). As in (ii), the tool proposes a, i.e., the subject concord for class 1 which is appropriate, among others, for the relative as opposed to o.*

(iv) *her: unspecified, as there is no unique referent e.g., the woman. On the basis of corpus frequency the software suggests the top five ranked possible options for the object concord, i.e. classes 1, 9, 3, 7, 5 (calculated from the Pretoria Sepedi Corpus, PSC). For human nouns, class 1 stands out in terms of frequency, and the selection is therefore for class 1 = mo. If, however, the relevant word is, e.g., from class 7, the user can type the word or click on the full list of object concords and selects the concord se for class 7.*

(v)   *loves: as for (i), the task is only to find the correct Northern Sotho equivalent: rata, plus adding the relative suffix -go which is, as for (iii), built into the relative construction rule.*

In the consultation, the relevant section for the appropriate noun class is highlighted while being placed in context within the grammar table. For cognitive use clickable options to see larger portions of the table are provided, as shown in table 2. The construction rules ensure contextually appropriate highlighting, e.g., only of the subject concord, in step (iii) or of the object concord in step (iv).

4.2   User Support for Text Reception

Access point 3: A Northern Sotho phrase: *monna yo a mo ratago* (or part thereof, e.g., *ratago, mo ratago, yo a mo,* etc.). The software analyses the phrase in terms of the formation rule for the relative (automatically or user selected), in fact reversing the strategy explained with respect to access point 2 for "the man who loves her".

5   Conclusion and Future Work

User support through direct guidance (and other support mechanisms) for complex grammatical structures allows the user to navigate via the shortest route to the information (s)he is looking for in an dictionary without having to work through long and often complicated grammar-type representations of complex grammatical structures. Such guidance is always available on demand, i.e., the user is not forced to work through any such support mechanisms if (s)he finds that the "standard" data in the dictionary are sufficient to solve his/her information need in a given situation. However, if more information is needed or if the standard presentation of the information (be this in the dictionary, in outer texts or in reference tools) is too difficult or complex to be easily understood, the user would have an alternative mechanism (or alternative mechanisms) to obtain the relevant information. It also successfully combats information overload and fulfils the needs of not only the learner of the language but also of the casual on-the-fly-user of the language; its flexibility is intended to provide a step towards individualization.

Different access points are available to the user depending on his/her pre-existing knowledge. It is not a profile-based dictionary. We envisage that such mechanisms be implemented as "plug-in modules" in entries of specific lemmas of the dictionary, i.e., an additional link is shown to the user on screen which (s)he can follow on demand. Since such modules can exist independently from the dictionary database, it would be feasible to reuse them in other environments as well. It would therefore be feasible to use such tools as writing tools integrated in a word processor, again activated by the user on demand, if (s)he requires to check the correct formulation of a complex grammatical construction, similar to spelling and grammar checkers that currently occur in popular word processing software.

Future work includes the development of a working prototype and possibly the full-scale implementation of user support for complex structures proposed in this paper as a module of electronic dictionaries. Identifying and categorising additional support techniques and developing prototypes and the full-scale implementation of such additional support techniques are also envisaged, as well as identifying further complex grammatical structures for which additional user support techniques may need to be developed. We will also investigate the possibility of the reuse of all such modules in writing tools for user support.

6   References

Bothma, T.J.D. (2011). Filtering and adapting data and information in the online environment in response to user needs. In: Fuertes-Olivera, PA and Bergenholtz, H (Eds.). 2011. *e- Lexicography: The Internet, Digital Initiatives and Lexicography.* London & New York: Continuum. Pp. 71-102.
Bothma, T.J.D. & Tarp, S. (2012). Lexicography and the relevance criterion. *Lexikos* 22: 86–108.

Gouws, R.H. and Prinsloo, D.J. (2005). *Principles and Practice of South African Lexicography.* Stellenbosch: SUN PReSS.

Prinsloo, D.J., Heid, U., Bothma, T.J.D. & Faaß, G. (2011). Interactive, dynamic electronic dictionaries for text production. In Kosem, I. and Kosem, K. 2011. *Electronic Lexicography in the 21st Century. New Applications for New Users. Proceedings of eLex 2011.* Bled, 10-12 November 2011. Bled: Trojína. Pp. 215-220. http://www.trojina.si/elex2011/elex2011_proceedings.pdf.

Prinsloo, D.J., Heid, U., Bothma, T.J.D. & Faaß, G. (2012). Interactive, dynamic electronic dictionaries for text production. *Lexikos* 22: 290-320. http://lexikos.journals.ac.za/pub/article/view/1009/526.

PSC, Pretoria Sepedi Corpus. University of Pretoria.

Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge. General lexicographical theory with particular focus on learner's lexicography.* (Lexicographica Series Maior 134). Tübingen: Niemeyer.

Tarp, S. (2011). Lexicographical and other e-tools for consultation purposes: Towards the individualization of needs satisfaction. In: Fuertes-Olivera, P.A. & Bergenholtz, H. (Eds.) 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography.* London & New York: Continuum. Pp. 54-70.

Tarp, S. (2012). Online dictionaries: today and tomorrow. In: *Lexicographica 28.* 253-267.

Acknowledgement