Originally published in: Kosem, Iztok/Kosem, Karmen (Eds.): Electronic lexicography in the 21st century.
New applications for new users. Proceedings of eLex 2011, Bled, Slowenien, 10-12 November 2011.
- Ljubljana: Trojina, Institute for Applied Slovene Studies, 2011. Pp. 215-220

Interactive, dynamic electronic dictionaries for text production

D.J. Prinsloo*, Ulrich Heid, Theo Bothma***, Gertrud Faaß****
* Department of African Languages and *** Department of Information Science,
University of Pretoria, Pretoria 0002, South Africa
** Department of Information Science and Natural Language Processing,
Hildesheim University, Marienburger Platz 22, 31141 Hildesheim, Germany and
Department of African Languages, University of Pretoria
E-mail: danie.prinsloo@up.ac.za, heid@uni-hildesheim.de, theo.bothma@up.ac.za, gertrud.faass@uni-hildesheim.de

Abstract

An interactive, dynamic electronic dictionary aimed at text production should guide the user in innovative ways, especially in respect of difficult, complicated or confusing issues. This paper proposes a design for bilingual dictionaries intended to guide users in text production; we focus on complex phenomena of the interaction between lexis and grammar. It will be argued that a dictionary aimed at guiding the user in lexical selection should implement a type of "decision algorithm". In addition, it should flag incorrect solutions and should warn against possible wrong generalisations of (foreign) language learners. Our proposals will be illustrated with examples from several languages, as the design principles are generally applicable. The copulative construction which is regarded as the most complicated grammatical structure in Northern Sotho will be analyzed in more detail and presented as a case in point.

Keywords: bilingual electronic dictionaries; user guidance; text production; dictionary design

1. Introduction

The electronic era was met with great enthusiasm and expectations. Early publications on electronic dictionaries were all about the potential of the new medium and the expected revolution it would bring along, thereby antiquating the paper dictionary in a decade or two. De Schryver (2009), however, rightfully expresses disappointment in respect of the pace of development of electronic dictionaries. More exciting was the introduction of what could be called "true electronic features" such as pop-up boxes, alternative access routes to the data, audible pronunciation and sophisticated search features. Some electronic dictionaries also solve problems in respect of lemmatisation, which cannot be resolved in paper dictionaries. Electronic dictionaries of today, however, could enter a more advanced dimension in fulfilling more sophisticated needs of the users, e.g. if access to data were not only based on a single lemma. Rundell (2009:9) refers to "game changing" developments that have "expanded the scope of what dictionaries can do and (in some respects) changed our view of what dictionaries are for". De Schryver (2009) calls in this context for an adaptive and intelligent dictionary (aiLEX) that will be able to "study and understand its user" and consequently to "present itself to that user". In most cases what is currently offered in dictionaries claiming that they give guidance in text production is in fact still on the level of text reception, and they generally give an overload of information. An interactive and dynamic electronic dictionary aimed at text production should guide the user in innovative ways, especially in respect of difficult, complicated or confusing issues. The underlying lexicographic concepts remain the same. What is at stake here are improvements in the article structure and access possibilities of electronic dictionaries.

2. Phenomena and proposals for their presentation

This paper proposes a design for bilingual dictionaries intended to guide users in text production; we focus on complex phenomena of the interaction between lexis and grammar. Our proposals can be illustrated with examples from several languages, as the design principles are generally applicable. The complex morpho-syntactic phenomena of the South African Bantu languages do particularly require a design of the proposed kind. Adaptivity to individual users (in De Schryver's (2009) sense) is not the main focus of this paper. We assume fixed user profiles for novice and expert users, and task profiles of text production and text reception. Nevertheless, our design allows for more flexibility beyond this simplistic parameterization.

Lexical selection in text production can be seen as a decision process. Grammar rules, semantics and communicative intentions, as well as (idiosyncratic, lexicalised) exceptions are among the parameters that influence the choice. Very often these rules are so complex and/or comprehensive that the average user of a dictionary or a grammar text does not (immediately) understand the rules that are being explained, or is simply overwhelmed with the amount of information presented. It is proposed that a dictionary tool is needed to simplify the decision process for the user and/or reduce the amount of information presented to the user to exactly what is needed to address the user's information need. A dictionary aimed at guiding the user in lexical selection should therefore implement a type of "decision algorithm". In addition, it should flag incorrect solutions and should warn against possible wrong generalisations of (foreign) language learners. As it stands in current sources on, for example, Northern Sotho copulatives in dictionaries and grammar books, the guidance given

could be regarded as cognitive aids. Our aim is to address the complexity by moving from the cognitive to text production by means of a selection process. This then also constitutes the rational for linking the dictionary with corpus data.

As a first prerequisite, this type of interactive, dynamic electronic dictionary should guide the user to the production of correct text. Prinsloo (2002) states the role of the lexicographer in this regard as a *mediator* between a complicated linguistic issue on the one hand and the dictionary user on the other, cf. also Tarp's (2011) idea of dictionaries as tools.

Text production support can be at different levels of complexity, for example:

- A simple decision algorithm (decision tree) based on one or two variables, illustrated by means of example sentences with limited additional explanation (available on demand).
- A situation where the grammatical rules are highly complex and follow a complex decision algorithm based on multiple variables, for example, "if *a* then *b* or *c*; if *b* then *d*, but if *c* then *e*, etc.".

Examples of the two levels of complexity will be described below. The first two examples reflect a very simple situation and the third a highly complex one. There are obviously multiple levels of complexity, and the above two reflect the extremes – all such support situations can be plotted on a continuum of complexity, each with its unique type of solution. Each decision tree (with its accompanying explanatory text and number of examples) depends on the nature of the data and the nature of the complexity of the problem.

An example from text understanding is homographic forms with different grammatical functions or meanings.

A case in point is Afrikaans *sy* which can be a personal pronoun (cf. (1)) or a possessive (cf. (2)). The decision algorithm is based on the context: the user verifies the presence of verbal governors (then *sy* is a feminine personal pronoun) or adjacent nominals (then *sy* is always and only a masculine possessive determiner).

(1) Sy het die boeke gekoop
She has the books bought
(She bought the books)

(2) Sy boek
His book

In the above case a simple decision algorithm and a few example sentences followed by a brief explanation should be sufficient to help the user to select the correct interpretation in a text understanding situation, or the correct equivalent in translation from Afrikaans.

Possessive determiners are also a major problem in beginners' text production, e.g. for English speakers learning a Romance language (our examples are in French): while English has different forms depending on the natural gender of the possessor (cf. *his* (masc.) vs. *her* (fem.)), French possessives agree with the grammatical gender of the possessed object, but don't mark the natural gender of the possessor, cf. (3).

(3) son livre (masc.) ("his/her book")
sa famille (fem.) ("his/her family")
ses livres/familles (plural)
("his/her books/families")

The decision algorithm for the selection of possessives thus has to ask for other parameters (number, gender) in French than in English or Afrikaans. Text production support for French possessives therefore requires a different decision algorithm than the above Afrikaans example, but should also be accompanied by a brief grammatical explanation and examples.

As a third example consider the user who wishes to express the basic copulative concepts *is*, *am* and *are* in Northern Sotho (Sepedi), a Bantu language spoken in South Africa. This is a very complex grammatical problem and therefore requires a more complex decision algorithm with multiple variables for text production support. In this case the decision algorithm for the selection of copulatives entails distinguishing between an *identifying* vs. a *descriptive* vs. an *associative* relation existing between the subject and its complement as in (4):

(4)

is
[identifying. copulative], ke lengwalo **(**it is a letter)

[descriptive. copulative], mosadi o bohlale
(the woman is clever)

[associative copulative], Satsope o na le Sara
(Satsope is with Sara)

Learners of Northern Sotho who want to use copulatives in speech or text production have at best to do intensive study of the copulatives from dictionaries and grammar books. Dictionaries typically provide basic and sometimes inadequate information. Grammar books such as Poulos and Louwrens (1994), on the other hand, provide an overdose (37 pages) of grammatical information, in a desperate effort to cover all the relevant and possible copulatives. Such details may be useful in a cognitive situation where the user would like to learn everything about the copulative, but they are hardly useful in a text production situation where the user simply wants guidance on which form to use. Such information overload could easily lead to "information death" (cf. Bergenholtz & Bothma, 2011). Compare the following extract from their summary of the identifying

copulative:
The identifying copulative
The indicative series *The present tense Principal Identifying* pos. lst and 2nd persons: ***SC - CB*** Classes: ***CP - CB*** neg. 1st and 2nd persons: ***ga - SC - CB*** Classes: ***ga - se - CB*** *Participial* pos. 1st and 2nd person: ***SC - le - CB*** Classes: ***CP - le - CB*** neg. lst and 2nd person: ***SC - se - CB*** Classes: ***CP - se - CB***
The Lemmatization of Copulatives in Northern Sotho 27
The future tense Principal pos. 1st and 2nd person: ***SC - tlô/tla - ba + CB*** Classes: ***CP - tlô/tla - ba + CB*** neg. 1st and 2nd person: ***SC - ka - se - bê + CB SC*** Classes: ***CP - ka - se - bê + CB*** *Participial* pos. 1st and 2nd person: ***SC - tlô/tla - ba + CB*** Classes: ***CP - tlô/tla - ba + CB*** neg. 1st and 2nd person: ***SC - ka - se - bê + CB*** Classes: ***CP - ka - se - bê + CB*** *The past tense Principal* pos. 1st and 2nd person: ***SC - bilê + CB*** Classes: ***CP - bilê + CB*** neg. 1st and 2nd person: ***ga - se - SC - be + CB ga - se - SC2 - a - ba + CB ga - SC2 - a - ba + CB*** Classes: ***ga - se - CP - bê + CB ga - se - SC2 - a - ba + CB1 ga - SC2 - a - ba - CB*** *Participial* pos. lst and 2nd person: ***SC - bilê + CB*** Classes: ***CP - bilê + CB*** neg. lst and 2nd person: ***SC - sa - ba + CB*** Classes: ***CP - sa - ba + CB***
The potential *Principal and participial* lst and 2nd person: pos. ***SC - ka - ba + C*** neg. ***SC - ka - se - bê + CB*** Classes: pos. ***CP - ka - ba + CB*** neg. ***CP - ka - sê - bê + CB***
The subjunctive 1st and 2nd person: pos. ***SC - bê + CB*** neg. ***SC - se - bê + CB*** Classes: pos. ***CP - bê + CB*** neg. ***CP - se - bê + CB*** Note also the compound negative ***SC/CP - se - kê + SC2 - a - ba + CB***
The consecutive lst and 2nd person: pos. ***SC2 - a - ba + CB*** neg. ***SC2 - a - se - bê + CB*** Classes: pos. ***SC2 - a - ba + CB*** neg. ***SC2 - a - se - bê + CB*** Note also the compound negative ***SC2 - a - se - ke + SC2 - a - ba + CB***
The habitual 1st and 2nd person: pos. ***SC - be + CB*** neg. ***SC - se - be + CB - be + CB*** Classes pos. ***CP - be + CB*** neg. ***CP - se - be + CB***
The infinitive pos. ***go - ba + CB*** neg. ***go - se - bê + CB***
The imperative pos. ***e - ba - ng + CB*** or ***ba - a - ng + CB*** neg. ***se - bê - ng + CB***
(Poulos and Louwrens1994:320)

Dictionaries, and especially electronic dictionaries, fail to give even basic receptive guidance or to treat the three main copulative relations in (4). Consider the article for the lemma **is** in the *Sesotho sa Leboa (Northern Sotho) - English Dictionary* (2003) in Figure 1.

In this example two of the three copulative categories, i.e., the identifying and associative copulatives, have not been treated, not to mention giving proper receptive or productive guidance. Paper dictionaries for Northern Sotho reflect the same deficiencies.

In the e-environment it is, however, possible to provide the user with the required guidance on which form is the correct one for a given situation, and to provide exactly the amount of information that is needed for each of the possible choices. In such a case a decision tree will reduce the amount of information considerably and the user can, at any stage, decide that his/her information need has been met and return to his/her primary task, namely to write a text.


Number of results found for 'is': 3

bago
which is/exists
ka bago
which can be

ele
1. (*interjection of astonishment/disapproval*)
2. (= **e le**) is/are

le
1. and, also, too : ***diyunibêsithi le dikôlô*** *universities and schools*
2. is, am, are : ***gê e le morutîši*** *if he/she is a teacher*
3. (*subject/object concord second person plural*) you : ***re a le tseba*** *we know you*
4. he/she, him/her, it
5. (*subject/object concord class 5*) this

le bjalê
even now
le gannyane
not at all, never
le gôna
also

Figure 1: The lemma **is** in the *Sesotho sa Leboa (Northern Sotho) - English Dictionary* (2003)

For example, when the user wants to write "the woman is clever" in Northern Sotho he/she should be guided to *mosadi o bohlale* and guarded from the typical error **mosadi ke bohlale*. The user can then be guided to subsequent levels of decisions, e.g. concerning person and noun class of the subject, tenses and moods, as well as a number of lexicalised exceptions, cf. Appendix 1.

The phenomena sketched above may usefully be presented to the user in terms of subsequent choices, e.g. by means of check boxes, radio buttons, etc. The visual appearance of the interface should make clear that the selections are the result of a decision process involving several steps. Instead of complex tables giving all options, a path through sub-tables should be shown, but together with links to synoptic tables which indeed allow the user to see the full picture if he/she wishes to. For a set of function words of the same category, the basic decision tree is constant. Users will only follow different paths through this tree, depending on their actual needs.

The internal representation of the data should be adapted to the particularized decision-tree-like access to the data. For this, not only synoptic tables of function words, but also a representation of the selection rules is needed, e.g. by means of linked templates.

A number of interface solutions should be considered:

- Just solve the problem, suggest the correct solution and give a visual presentation and link to 'read more' sections such as FAQs or outer texts.

- Supply a link to *read more* information where distinctions on a cognitive level are made.
- Supply a link to guidance on the basis of e.g. *frequently made errors*.
- Give good, typical examples of use throughout.

All envisaged actions should be based upon a grammatical description of the construction to be tackled e.g. pronouns in Afrikaans, English, French or the copulative construction in Northern Sotho. One could argue that these issues have been sufficiently described in standard grammars of these languages. However, one should not assume that the format of these descriptions is such that they are ready to use for our purposes. A reorganization of the data will be necessary.

The process to produce such a dictionary article requires at least three sequential steps, building on one another:

- Step 1 would be to acquire comprehensive and accurate data for the set of rules etc. to be described. This includes the grammatical rules as well as pertinent examples, common errors, etc.
- In Step 2 the lexicographer in collaboration with a database expert needs to reorganise the data so that it will be possible for a programmer to implement a decision tree. This requires at least two processes:
    - The logic of the decision process needs to be worked out very carefully, i.e., what is the logical sequence of the decisions, how much information is required to make and/or support the decisions, when are what type of examples needed, when are links to outer texts required, etc.
    - The data need to be marked up in such a way that each of the data elements defined in the analysis of a specific complex problem can be identified at the required level of granularity. This implies that the database should make provision for such extensions, either by using an extensible XML schema or additional tables and fields in a relational database (depending on the original design of the system), (cf. Bothma (2011)).
- In Step 3 the programmer takes the flow diagram of the decision tree together with all the explanations, examples and linked data, and implements this. The programmer should also design a "user-friendly" interface that is intuitive for the average user and supports him/her to follow the correct trail through the decision tree for the given information need.

3. Exemplification: complex cases of copulative selection

In a text production situation a user can consult the dictionary as an external source to obtain the required information. However, it is also possible that the support the user requires be integrated into a word processor the user is using to construct his/her text. In such a case the user may require feedback on his/her own text production efforts based on his/her grammatical knowledge without specifically consulting the dictionary. In such a case the e-dictionary could be integrated into the word processor as a grammar checker, similar to the features currently available in popular word processing software.

Let us depart from a most common error scenario in Northern Sotho, for example, the user typing **lesogana ke bohlale*. Learners usually know that *ke* means 'it is' and no distinction is made between *he is*, *she is*, *they are* and *it is* in Northern Sotho: all convert to *it is*, e.g. *(monna) ke morutisi* ' he is (it is) a teacher'. As a second example consider **monna o morutiši* instead of *monna ke morutiši* 'The man is a teacher'. Learners are accustomed to using the subject concord *o* with class 1 nouns in sentence construction and it is the correct form in two out of the 3 copulative relations (descriptive and associative copulatives: so attempting to use it also in the identifying copulative is a common error).

The student types **lesogana ke bohlale* in a word processor linked to the electronic dictionary and all three words are or only the *ke* is flagged as incorrect. A quick solution is offered by means of a suggestion box, in this case offering three possibilities namely *le*, *e le* 'is/am/are' and *e lego* 'who/what is/am/are'. The user who has basic knowledge of the modal system will know which one to select. Most users, however, would need further guidance and this is offered by a decision process guiding him/her through the three possible moods (Indicative *le*, Situative *e le* or Relative *e lego*) of the decision tree for the descriptive copulative with sub-decisions. The process for **monna o morutisi* is similar, i.e. a decision process guiding him/her through the three possible moods (Indicative *ke*, Situative *e le* or Relative *e lego*) of the decision tree for the identifying copulative respectively, with sub-decisions.

3.1 Different levels of user guidance

Figure 2 provides a schematic illustration of a pop-up guidance screen sequence for **lesogana ke bohlale*.

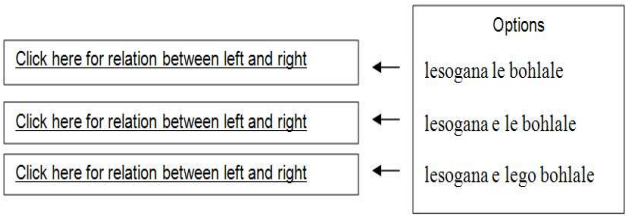

Figure 2: Dictionary feedback for **lesogana ke bohlale*

If more guidance in respect of the descriptive relations in the Indicative, Situative and Relative is required, the user can click the buttons in Figure 2 to display the information given in Figures 3, 4 and 5.

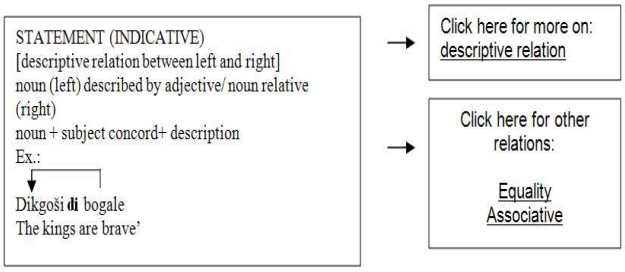

Figure 3: Pop-up 2a: Information boxes for *lesogana le bohlale* in Level 1

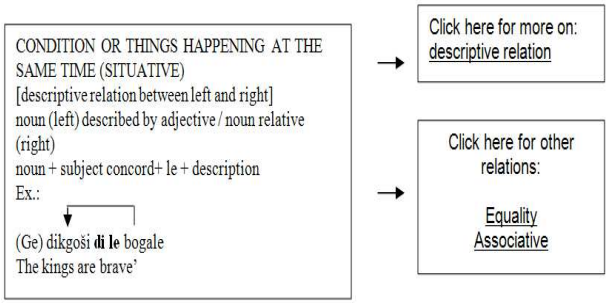

Figure 4: Pop-up 2b: Information boxes for *lesogana e le bohlale* in Level 1

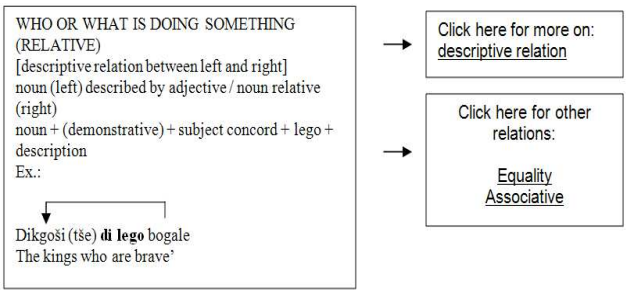

Figure 5: Pop-up 2c: Information boxes for *lesogana e lego bohlale* in Level 1

In each case, the panel given in the left part of the mock-up provides the information needed for text production. Users with more (cognitive) needs can access a fuller picture via the buttons on the right hand side.

3.2 From text production guidance to full grammatical guidance

Pop-up boxes giving more information and typical examples of descriptive relations can be provided on a third level for the Indicative, Situative and Relative. See, for example, additional information for the Indicative in Figure 6.

A second scenario is where comprehensive guidance is required, e.g. when the user wants to know how to say *is* in Northern Sotho. In this case a combination of decision processes is required. These processes are enriched with information from corpora and processed corpus data linked with the dictionary.

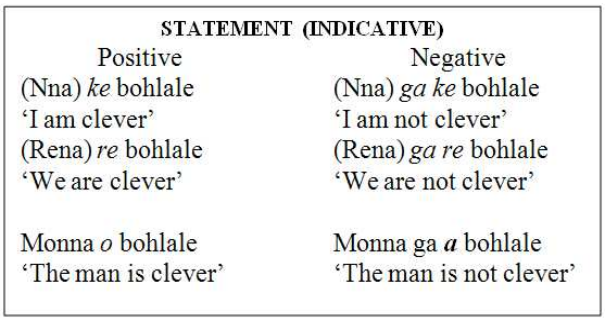

Figure 6: Pop-up 3a: Information boxes for *descriptive relation* in level 2

4. Conclusion

The project described above is driven by two underlying motivations, namely the urge to compile electronic dictionaries that can do better than current ones through maximal utilization of advanced modern technologies and the need for intelligent and dynamic dictionaries guiding the user in new innovative ways. We believe that step-by-step guidance, mainly through sequences of choices, the provision of additional relevant information on request as well as protection against incorrect conclusions, are the cornerstones of the design of such intelligent dictionaries.

5. References

Bergenholtz, H., Bothma, T.J.D. (2011). Needs-adapted data presentation in e-information tools. *Lexikos*, in press.

Bothma, T.J.D. (2011). Filtering and adapting data and information in the online environment in response to user needs. In P.A. Fuertes-Olivera, H. Bergenholtz (eds.) *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. 2011. London & New York: Continuum, pp. 71-102.

De Schryver, G-M. (2009). State-of-the-Art Software to Support Intelligent Lexicography. In R. Zhu (ed.) (2009) *Proceedings of the International Seminar on Kangxi Dictionary & Lexicology*. Beijing: Beijing Normal University, pp. 565–580. Also: http://www.hcxf.cn/read.asp?id=570.

Prinsloo, D.J. (2002). The Lemmatization of Copulatives in Northern Sotho. *Lexikos*, 12, pp. 21-43.

Rundell, M. (2009). The Road to Automated Lexicography: First Banish the Drudgery then the Drudges? In S. Granger, M. Paquot (eds.) *eLexicography in the 21st century: New challenges, new applications, Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*, pp. 9-10.

Sesotho sa Leboa (Northern Sotho) - English Dictionary (2003) http://africanlanguages.com/sdp/.

Tarp, S. (2011). Lexicographical and other e-tools for consultation purposes: Towards the individualization of needs satisfaction. In P.A. Fuertes-Olivera, H. Bergenholtz (eds.) e-*Lexicography: The Internet, Digital Initiatives and Lexicography*. 2011. London & New York: Continuum, pp. 55-70.

Appendix 1: The Copulative in Northern Sotho