State-of-the-Art Software to Support Intelligent Lexicography

Gilles-Maurice DE SCHRYVER
Department of African Languages and Cultures, Ghent University,
Ghent, Belgium
Xhosa Department, University of the Western Cape, Bellville,
Cape Town, South Africa
TshwaneDJe Human Language Technology, Pretoria Branch,
Pretoria, South Africa

Abstract This paper presents a proposal for a revolutionary type of electronic dictionary, one in which the potential is explored to link an automatically derived dynamic user profile to the proffered multimedia lexicographic output. Such adaptive and intelligent dictionaries may use the TshwaneLex dictionary production system at their core, to which a string of artificial intelligent components are added. This proposal is illustrated by means of the description of a project to compile an online Swahili to English dictionary. Swahili is both the most widely spoken African language, and the one sub-Saharan language most commonly taught throughout the world. As a theoretical framework for the development of this new type of electronic dictionary, the "fuzzy answer set programming" framework (Van Nieuwenborgh *et al.*, 2007) is advanced.

Keywords Adaptive Lexicography; Artificial Intelligence; Dictionary Production System; Online Dictionaries; Swahili; English

1 The future of lexicography

The future of lexicography is digital, so much is certain. Yet what that digital future will look like, is far less certain. Already the authoritative *Oxford English Dictionary* limits its quarterly updates to the online version only (OED Online), with no plans to ever print that material. What is offered in electronic form, however, by and large mimics what used to be printed. No cognizance is taken of the true power of the digital age. In this regard, it is telling indeed that TIME's *Person of the Year for* 2006 was "you" (Grossman, 2006). TIME magazine saw "collaboration on a scale never seen before," singled out "knowledge Wikipedia and the million-channel people's network YouTube" as prime examples, and pointed out that "the tool that makes this possible is the World Wide Web." In the current proposal, the Web also takes centre stage, and the novel type of lexicography that is proposed revolves entirely around you.

The last lexicographic revolution already dates back to the 1980s, when the highly innovative *Collins COBUILD English Language Dictionary* was published (Sinclair & Hanks, 1987). That dictionary was compiled entirely from scratch, and for the first time meanings were mapped onto use as observed in *real* and *massive amounts* of language data, or thus, in large electronic corpora. Although various corpus query systems and tools have been developed since then, and even though ingenious dictionary writing systems have recently been created, today's dictionaries still very much look like they always have: static. To break out of the straitjacket of the paper world, many a lexicographer has touted the imminent next revolution, viz. the potential of electronic dictionaries to become dynamic. Sadly, nothing is further from the truth. To take stock of the latest state-of-the-art with regard to electronic dictionaries, it is instructive to read what two of the most imminent lexicographers have to say about this in *The Oxford Guide to Practical Lexicography*: "The field is developing so quickly that what we write today is almost guaranteed to be out of date by the time this book appears." (Atkins & Rundell, 2008: 238) This is unfortunately a huge overstatement, as a comparison of their description of the field of electronic dictionaries with even a five-year-old study (De Schryver, 2003a) will easily show. Apart from the addition of some trivial gimmicks to electronic dictionaries, nothing whatsoever has changed.

What is needed is a dictionary, obviously electronic, that is truly adaptive-meaning that it will physically take on different forms in different situations; and one that would do so as intelligently as possible — meaning that it would have the ability to study and understand its user, and based on that to learn how to best present itself to that user. With this, the field has moved to a very different paradigm indeed, to that of *adaptive and intelligent lexicography*, for short aiLEX, the subject of this proposal.

2 From Simultaneous Feedback (SF) to Fuzzy SF

The seeds of what is proposed here can be traced back to my MA and PhD research. In my MA thesis I introduced the concept of *Simultaneous Feedback* (SF), which, in a nutshell and as initially conceived, can be understood as entailing a dictionary-making method in terms of which the release of several small-scale parallel dictionaries triggers off feedback that is instantly channelled back into the compilation process of a main dictionary. Within that framework, the target users guide the compilers near-simultaneously during the entire compilation process. The unabated retrieval of feedback should thus be considered as the main pillar of the methodology. Bilingual (parallel) dictionaries compiled within this framework include three reference works for Cilubà (De Schryver & Kabuta, 1997, 1998, De Schryver, 1999: 55—87) and three for Northern Sotho (Prinsloo & De Schryver, 2000, De Schryver, 2001, De Schryver, 2007).

In the Cilubà and Northern Sotho projects, the retrieval of feedback had followed the channels of such standard approaches as (natural) participant observation, formal and informal discussions, anonymous mail survey questionnaires, controlled tests, etc. Through a crosscomparison of the results of the various types of feedback, the idea had been to arrive at a representative body of users" desires for each of the respective target user groups. Still, the realization that none of the employed feedback methods is devoid of problems, and that even the balancing out of different types of feedback is only approximate, prompted the search for a straightforward, automatic, neutral and invisible arbiter. Such an unobtrusive arbiter was found in the form of electronic-dictionary log files. In other words, instead of compiling various parallel hard copy dictionaries for the purposes of retrieving feedback, the idea was to make the dictionary available online on the Web while it is still being compiled, in order to log and use feedback truly simultaneously and to be able to interact with the dictionary user in a one-to-one fashion. With this, one arrived at a bold compilation strategy indeed, as users were not only invited to be spectators of "in progress dictionary compilation", but were also, implicitly and informally, led to provide crucial feedback while using that in-progress work. From a dictionary-compilation strategy angle one thus moved from a *discrete* approach to retrieving feedback for a group of users to a *continuous* one for single users, which is why the" electronic adaptation" of SF, which was introduced in my PhD dissertation, was baptized *Fuzzy SF.* Subsequently, the compilation of some Cilubà and Northern Sotho dictionaries also proceeded within the framework of Fuzzy SF (cf. Kabuta *et al.*, 2006, respectively De Schryver & Joffe, 2003).

3 From Fuzzy SF to aiLEX

While the idea of an adaptive and intelligent dictionary is novel, adaptive aspects of related hypermedia systems have already been developed. Particularly relevant in this context are educational hypermedia systems in which the presented information, as well as pointers to additional information, is adapted to the user's knowledge of the topic under consideration (De Bra *et al.*, 2003). Also, some work has been done on adaptive encyclopedia and virtual museum tours (Oberlander *et al.*, 1998). In each case, personalization is achieved by building a model of the user's goals, preferences and expertise (Brusilovsky, 2001), either through explicit interaction with the user, or by observing the user's behaviour. Another line of research has focused on analyzing log files of Web services to improve the static structure of a website, or to dynamically generate websites that are adapted to the expected interests of a certain user. The main techniques being used to this end are co-occurrence-based data-mining techniques and clustering (Mobasher *et al.*, 1999, 2000). Similarly, in the context of Web search, Web usage analysis is employed to personalize search results (Agichtein *et al.*, 2006). In recent years one has seen an increased interest in customized content for mobile devices, in which not only user models are taken into account, but also information about the mobile device being used (e. g. size and resolution of the display), and about the location of the user (Agostini *et al.*, 2005, Bettini & Riboni, 2004). A typical implementation is for example the following: a user approaching a railway station may automatically be provided with information about departing trains to destinations which that user has visited before.

Some of the techniques from these systems may be reused for developing an aiLEX electronic dictionary. Web usage analysis techniques may be applied to the access logs of the dictionary, for example, teaching us hidden relationships between user behaviour and user preferences (e. g. users who perform lookups of type X are usually not interested in linguistic information of type Y). However, the relatively small number of users, and the limited number of lookups per user for even repeat visitors to an online dictionary (compared to Web search engines), as well as the large diversity among users, will limit the usefulness of such techniques in our context. Some hypermedia systems learn user models based on expert knowledge (De Bra *et al.*, 2003), by manually specifying rules to update a user's profile when that user has performed a certain action. One example may be to estimate the user's knowledge of a language based on the kind of words that are being looked up. Rules to implement this may take the form of *If the user mainly searches uncommon words then set the user's level to expert*, together with appropriate, numerical threshold values for "mainly" (e. g. in at least 75% of the cases) and "uncommon" (e. g. occurs at most 20 times in a given corpus).

A hybrid approach seems most promising for our purpose, initially based on expert rules that are allowed to (automatically) evolve based on observed user behaviour. The initial availability of manually defined rules ensures that the system is useful right from the start, while the ability to evolve the rules, based on actual user behaviour, allows both to refine the initial rules and to correct initial assumptions that turn out to be false. This process will be based on user feedback, which could be explicit (whereby users manually update their own profile) and implicit (namely observations of user behaviour that are not consistent with the learned model).

To implement such a rule-based strategy, the "fuzzy answer set programming framework" (Van Nieuwenborgh *et al.*, 2007) seems a promising solution. This framework combines the advantages of rule-based languages for knowledge representation with those of fuzzy logics, resulting in a non-monotonic, declarative formalism to model relationships between continuous variables. Thus, rather than using crisp threshold values to define terms like *mainly* and *uncommon*, gradual definitions can be used, according to which a word occurring 20 times, for instance, would be considered uncommon to some degree between 0 (i. e. the term *uncommon* is definitely not appropriate) and 1 (i. e. the term *uncommon* is definitely appropriate). Note that the gradual nature of the conditions of these rules is consistent with the observation that also their conclusion, viz. the familiarity of a user with the language, is a matter of degree. The non-monotonic nature of (fuzzy) answer set programming also makes it an ideal framework for modelling exceptions to rules, which is one of the most straightforward and interesting ways of refining the initial rules. Finally, note that fuzzy answer set programming allows for a seamless integration of qualitative and quantitative user preferences, whereas existing models tend to be suitable only for qualitative (e. g. using traditional answer set programming), or only for quantitative (e. g. using weighted constraint satisfaction) models (Viappiani *et al.* 2002).

4 Objectives of the aiLEX project

4.1 *Fundamental research into the theoretical underpinnings of aiLEX*

Even though one can see a clear line from Simultaneous Feedback to Fuzzy SF to aiLEX — and thus from traditional paper-based feedback, to the use of Web log files as invisible arbiters, to fuzzy answer set programming — the implementation of the proposed theoretical framework opens up a lot of research questions that will need to be answered during the course of this project. The novel aspects of the methodology, however, make this a worthwhile venture. Apart from the truly unconventional notion to bring in concepts from the field of AI (artificial intelligence) into the heart of the lexicographic undertaking, this multidisciplinary aspect may exactly be the way out of the current impasse — dare I say: the current "static" field of lexicography in the digital age.

From a theoretical perspective, the proposed aiLEX research team should be in a position to truly analyze all the theoretical underpinnings, this in contrast to what was the case for the concepts of SF and later Fuzzy SF, which were developed in response to an urgent desideratum, in that dictionaries were needed "now" (as South Africa's *Minister of Lexicography*, M. B. Kumalo, used to put it). SF and Fuzzy SF indeed made the impossible possible, but in the haste to *produce* dictionaries within these frameworks (something all eleven South African *National Lexicography Units*, amongst others, have done by now), not enough attention went to the theoretical aspects. The value of such fundamental underpinnings to the lexicographic community cannot be emphasized enough. Current electronic dictionaries are developed in a total theoretical vacuum, and to date the only in-depth scientific survey remains my own *Lexicographers' Dreams in the Electronic-Dictionary Age* (De Schryver, 2003a).

Looking ahead, the fundamental research that needs to be undertaken within the proposed research project would need to analyze the validity and feasibility of each of the following ten claims (compare De Schryver & Prinsloo, 2001a), all with reference to an aiLEX dictionary or "package" (and, for the time being, expressed in lexicographic terms rather than in applied mathematics ones):

(1) In addition to data being continuously available online, parallel packages (both in print and in electronic format) may be released throughout the endeavour to compile a main package, answering an urgent desideratum to provide users with dictionaries now, and enabling the inclusion of feedback into the very compilation methodology itself.

(2) Since a completed package has been thoroughly "tested" before it is released, it contains user feedback right from the start; and once it is used it (preferably) gathers its feedback indirectly, informally and unknowingly, eliminating any barriers between compilers and users.

(3) The package offers fully fledged default dictionaries, just like any other hard copy or electronic dictionary, and, additionally, each user can retrieve a personally tailored reference work in print or in electronic format.

(4) The package is a family reference work that can be customized for several users, and is continuously re-customized for each single user over time.

(5) The package is primarily descriptive, and includes tools for user-initiated modifications.

(6) The package provides for all linguistically sound lemmatization approaches in parallel, allowing users to decide on the one(s) appropriate for them at the time of consultation.

(7) Both the access to and the visual presentation of the data slots are such that the distinction between onomasiological and semasiological dictionaries disappears.

(8) The package endeavours to be all dictionaries in one, moulding itself according to specific needs and varying with time as a decoding or encoding, monolingual, bilingual or hybrid dictionary, with adjustable /graded difficulty levels.

(9) The package contains a set of fully integrated built-in multimedia (sub) corpora(i. e. text, computer graphics and audio), from which data are generated automatically when needed (i. e. are queried unperceivingly by the software), and which can also be accessed interactively (i. e. are queried knowingly by the users).

(10) Finally, all multimedia data slots — whether they have been prepared by the lexicographers, have been culled automatically or interactively from the (sub) corpora,or have been supplemented /supplied by the user — are hyperlinked in the package on all levels and in all directions.

If it can be shown that, compared to any principle currently utilized in dictionary-making and compared to any existing multimedia reference work, these ten points are either absent from or constitute important improvements over what is done or available at present, one may well have assembled enough facts and arguments to draw up the needed theoretical underpinnings of aiLEX.

4.2 *New insights into true dictionary use*

Between the theoretical underpinnings on the one hand, and possible tangible products on the other, lies an exciting opportunity to pursue modern dictionary research. This research, in a wåy, is built in, in that there is no escaping it, which makes it such a worthwhile venture.

For example, although the proposal to draw conclusions from some type of online log files in order to improve dictionaries was already expressed in the mid-1980s (Abate, 1985; Crystal, 1986), and although numerous researchers have reiterated this idea in recent years (e. g. Hulstijn & Atkins, 1998; Sobkowiak, 1999; Docherty, 2000; Harley, 2000; Sato, 2000; Pruvost, 2003; Varantola, 2003), very few reports have been published of real-world dictionaries actually making use of this strategy. Notable exceptions are Löfberg (2002), Prószéky & Kis (2002), Jakopin & Lönneker (2004), Bergenholtz & Johnsen (2005), and Müller-Spitzer & Möhrs (2008). Instead, electronic dictionaries cum log files aimed at (manually) adapting online dictionary contents seem to be more popular in *re-*

search environments focusing on vocabulary acquisition (e. g. Hulstijn, 1993; Knight, 1994; Hulstijn & Trompetter, 1998; Laufer, 2000; Laufer & Hill, 2000). When it comes to electronic dictionaries, statements regarding log files are often hypothetical, such as in: "A log file of user access and queries is kept that *should* serve to give insight on how such a service is used" (Popescu-Belis *et al.* , 2002: 1144 [emphasis added]). What is true for log files, is also true for the utilization of direct feedback, whereby users are encouraged to comment online (e. g. Dodd, 1989; Carr, 1997, Considine, 1998; Harley, 2000; Nesi, 2000; Warburton, 2000, Campoy Cubillo, 2004; Ne'eman & Finkel, 2004); that is, reports on what is done with this type of feedback are hard to come by.

The earliest implementation of (the first stages of) adaptive lexicography on the Web was for an *Online Northern Sotho - English Dictionary* (De Schryver & Joffe, 2003). Of the five novelties introduced in that reference work (De Schryver, 2003b: 5—10), one is highly relevant here, namely the so-called "dynamic metalanguage customization". This means that, depending on the interface-language chosen, the output-language of all dictionary metalanguage such as part-of-speech tags, usage labels, cross-reference marker texts, etc. is customized. A world's first for any online dictionary at the time (and to this date), this metalanguage customization is realized in real time and thus dynamically on the Web, and as such this was a first (timid) step towards an aiLEX electronic dictionary.

An analysis of the log files attached to this prototype dictionary, as well as of the online feedback forms, only hints at the vast research possibilities (De Schryver & Joffe, 2004). The outcome of a second prototype dictionary project, for an *Online Swahili - English Dictionary* (Hillewaert & De Schryver, 2004), hints in the same direction (De Schryver *et al.* , 2006). With reference to the current proposal, then, these prototype projects indicate that real electronic dictionaries, used in a natural setting, with no manipulation of research variables whatsoever, can indeed be used to unobtrusively study the dictionary look-up behaviour of particular users. With a well-designed tracking function, any number of individual user's look-up strategies may thus be monitored across time, which is especially relevant for studying language acquisition aspects such as vocabulary retention, and for drawing up the user profiles needed for a fully-fledged aiLEX dictionary.

What therefore needs to be done in the proposed research is to generalize such user profiles, so that one can formalize them. Formalization will lead to two sets of data:

(1) On the one hand this will contribute to the available knowledge on "dictionary use" (a field in which too many armchair-claims continue to be made, as opposed to precise descriptions of what real users actually do, as seen during *actual and unobtrusive* use of real dictionaries, in real settings);

(2) On the other hand language-specific dictionary data will feed into a real dictionary project (for which, cf. the next section).

4.3 *The compilation of a revolutionary Swahili reference work, with a focus on the core software – TshwaneLex*

So far nothing has been said about the language(s) involved. This is so because the underlying theoretical framework to be developed is language-independent. Based on the outcomes of the two prototype online dictionary projects discussed in the previous section, it is also safe to say that large portions of the user profiles are again language-independent — but not all, which means that the modules which feed dictionary data to a particular user at a particular time will have to take both general and language-dependent aspects into account.

What was also clear from the two prototype online dictionary projects was that one will need a very large amount of usage data in order to both develop the theory sensibly, and to compile a real, useful application. Having access to large amounts of data (i. e. millions of dictionary lookups, especially those with a deep time-depth for particular, obviously anonymous, users) is ideal. Given my vast experience with and interest in Bantu lexicography, and at the same time the requirement to make sure the envisaged reference work is used as widely as possible, the compilation of an aiLEX *Swahili to English* electronic dictionary gently imposes itself. Indeed, Swahili, which is used by up to 50 million people in Eastern Africa, is not only the most widely spoken African language, but also the one sub-Saharan language most commonly taught throughout the world. In addition to my contribution in producing the mentioned prototype Swahili dictionary, I have also worked on and co-published about Swahili before, both in lexicography (De Schryver & Prinsloo 2001b, Prinsloo & De Schryver, 2001, De Schryver *et al.*, 2006) and in the field of language technology (De Pauw *et al.*, 2006, 2008; De Pauw & De Schryver, 2008). The resulting Swahili language-technology tools have been made available to the wider scientific community on the African Language

Technology portal aflat. org.

The Swahili to English electronic dictionary will be a hybrid product, in the sense that it will contain both monolingual (Swahili-Swahili) and bilingual (Swahili-English) features. The main dictionary database software has already been developed, namely TshwaneLex, while a new Swahili corpus will be built from scratch. For the extra artificial intelligence modules that will need to be designed, at least one team member will have a background in computer science.

When it comes to the lexicographic work proper, both the theoretical framework and the logs will be kept at hand during the compilation of the data. On a first axis, three levels can be recognized with regard to the contents of the TshwaneLex database proper: (1) the DTD (document type definition), which drives the structure of each and every article, (2) the lexicographic data themselves which populate the database, and (3) the layout /formatting of those data. These three levels are (and have to be) strictly separated. On a second axis, and this with a "smart" dictionary in mind, level (2) will (and has to) be divided into two further layers: (2a) attribute lists (for all metalexicographic data, i. e. all recurrent information which is best selected from closed lists), and (2b) the lexicographic data proper (i. e. the dictionary data themselves, for each and every lemma compiled and derived from corpus data). On a third and final axis, provision will need to be made for the possibility that each of the previous levels is further subdivided into *N* number of layers (with *N* to be derived in bootstrap-fashion as the project proceeds).

Of all these required layers, only the further subdivision of (2a) has so far been empirically tested, and this is what is now known as "dynamic metalanguage customization" (De Schryver & Joffe 2005). Expressed in simple terms, this procedure means that, say, instead of having one definition in the database for a particular sense, one has *N* definitions, graded according to various parameters, chief amongst them (perceived) difficulty. Likewise, and as another simplified example, multiple types of example sentences will be prepared, including, at one extreme, different ranges of pointers directly(in)to the attached corpus lines.

During actual dictionary consultation, any particular user will of course never be presented with all these layers of data, rather, the artificial intelligence modules will present those and only those which the user most likely wants /needs to see at that point in time. Of all the electronic dictionaries of the future proposed to date, an aiLEX dictionary comes closest to the *Virtual Dictionary* put forward by Atkins

(1996). Yet, whereas a Virtual Dictionary is mainly created at the time of dictionary consultation, an aiLEX dictionary aims to build a true user profile, through the continuous retrieval of feedback, with which a tailored reference work is simultaneously assembled. The fact that the first such dictionary could be one for Swahili is an additional reason to be excited.

5 Conclusion

Obviously, the division between theory, dictionary use and dictionary compilation /product will not be as clear-cut as presented in Section 4 above. Rather, there will be a continuous interaction between these three objectives /pillars, and the research will often be of a cyclical nature. Cyclical isn't circular, however.

Summarizing the intended research one could say that this project proposal revolves around a revolutionary type of electronic dictionary in which the potential is explored, and exemplified for a Swahili to English dictionary, to link an automatically derived dynamic user profile to the proffered multimedia lexicographic output. Such adaptive and intelligent dictionaries are — by design — excellent tools to study genuine dictionary use, which in turn leads to exciting answers to age-old as well as new lexicographic questions.

Upon completion of the project, there are at least three areas of further impact, with linked additional research. Firstly, even though the focus in this project is on only one type of electronic dictionary, namely "online on the Internet", a different and obvious area to move into with reference works at large is the mobile (phone and other) one. Secondly, the focus is on Swahili, and no doubt colleagues from across the board will want to try out aiLEX for their respective language(s). Thirdly, although limited to lexicography in this project, the implications for say CALL lingware are important. aiLEX, then, has the potential to trigger a positive theoretical and technological tsunami for many years to come.

Acknowledgements

The input of Dr Steven Schockaert, from the Department of Applied Mathematics and Computer Science at Ghent University, is gratefully acknowledged for Section

3. He is a valued member of the "Computational Web Intelligence Team", which specializes in the development and use of computational intelligent methods for the next generation Web applications, using techniques from fuzzy and rough set theory for the enhancement of recommender, question answering and trust systems, as well as for the enrichment of ontologies.

References

Abate, F. R. 1985. Dictionaries Past & Future: Issues and Prospects. *Dictionaries* 7: 270—283.

Agichtein, E., E. Brill, S. Dumais & R. Ragno. 2006. Learning user interaction models for predicting Web search result preferences. In *SIGIR* 2006 (*Proceedings of the* 29*th annual international ACM SIGIR conference on research and development in information retrieval*): 3—10. New York.

Agostini, A., C. Bettini, N. Cesa-Bianchi, D. Maggiorini, D. Riboni, M. Ruberl, C. Sala & D. Vitali. 2005. Towards highly adaptive services for mobile computing. In *Proceedings of IFIP TC8 Working Conference on Mobile Information Systems*: 121—134. Oslo.

Atkins, B. T. S. 1996. Bilingual Dictionaries: Past, Present and Future. In *Euralex* 1996 *Proceedings*: 515—546. Gothenburg.

Atkins, B. T. S. & M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. New York.

Bergenholtz, H. & M. Johnsen. 2005. Log Files as a Tool for Improving Internet Dictionaries. *Hermes*, *Journal of Linguistics* 34: 117—141.

Bettini, C. & D. Riboni. 2004. Profile aggregation and policy evaluation for adaptive Internet services. In *Proceedings of The First Annual International Conference on Mobile and Ubiquitous Systems*: 290—298. Boston.

Brusilovsky, P. 2001. Adaptive Hypermedia. *User Modeling and User-Adapted Interaction* 11 (1—2): 87—110.

Campoy Cubillo, M. C. 2004. Computer-mediated lexicography: An insight into online dictionaries. In Campoy Cubillo, M. C. & P. Safont Jordà. (eds.), *Computer-Mediated Lexicography in the Foreign Language Learning Context*: 47—72. Castelló de la Plana.

Carr, M. 1997. Internet Dictionaries and Lexicography. *International Journal of Lexicography* 10 (3):209—230.

Considine, J. P. 1998. Why do large historical dictionaries give so much pleasure to their owners and users? In *Euralex* 1998 *Proceedings*: 579—587. Liège.

Crystal, D. 1986. The ideal dictionary, lexicographer and user. In Ilson, R. F. (ed.), *Lexicography: An emerging international profession*: 72—81. Manchester.

De Bra, P., A. Aerts, B. Berden, B. de Lange, B. Rousseau, T. Santic, D. Smits & N. Stash. 2003. Aha! The adaptive hypermedia architecture. In *Proceedings of the* 14*th ACM conference on hypertext and hypermedia*: 81—84. New York.

De Pauw, G. & G. -M. de Schryver. 2008. Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes. *Lexikos* 18: 303—318.

De Pauw, G., G. -M. de Schryver & P. W. Wagacha. 2006. Data-Driven Part-of-Speech Tagging of Kiswahili. *Lecture Notes in Artificial Intelligence* 4188: 197—204.

De Pauw, G., P. W. Wagacha & G. -M. de Schryver. 2008. Bootstrapping Machine Translation for the Language Pair English - Kiswahili. In *Proceedings of SREC* 2008: 30—37. Kampala.

De Schryver, G. -M. 1999. *Cilubà Phonetics, Proposals for a corpus-based phonetics from below'—approach* (Recall Linguistics Series 14). Ghent.

De Schryver, G. -M. (ed.). 2001. *Pukuntšutlhaloši ya Sesotho sa Leboa* 1.0 (*PyaSsaL's First Parallel Dictionary*). Pretoria.

De Schryver, G. -M. 2003a. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16 (2): 143—199.

De Schryver, G. -M. 2003b. Online Dictionaries on the Internet: An Overview for the African Languages. *Lexikos* 13: 1—20.

De Schryver, G. -M. 2007. *Oxford Bilingual School Dictionary: Northern Sotho and English / Pukuntšu ya Polelopedi ya Sekolo: Sesotho sa Leboa le Seisimane. E gatišitšweke Oxford.* Cape Town.

De Schryver, G. -M. & D. Joffe. 2003 - . *Online Sesotho sa Leboa (Northern Sotho) - English Dictionary.* Available from http://africanlanguages.com/sdp/.

De Schryver, G. -M. & D. Joffe. 2004. On How Electronic Dictionaries are Really Used. In *Euralex* 2004 *Proceedings*: 187—196. Lorient.

De Schryver, G. -M. & D. Joffe. 2005. Dynamic Metalanguage Customisation with the Dictionary Application TshwaneLex. In *Proceedings of Complex* 2005: 190—199. Budapest.

De Schryver, G. -M., D. Joffe, P. Joffe & S. Hillewaert. 2006. Do Dictionary Users Really Look Up Frequent Words? - On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* 16: 67—83. [Reprinted in *Lexicography*. 2009. Bangalore: Icfai University Press.]

De Schryver, G. -M. & N. S. Kabuta. 1997. *Lexicon Cilubà - Nederlands* (Recall Linguistics Series 1). Ghent.

De Schryver, G. -M. & N. S. Kabuta. 1998. *Beknopt woordenboek Cilubà - Nederlands & Kalombodi-mfùndilu kàà Cilubà (Spellingsgids Cilubà)* (Recall Linguistics Series 12). Ghent.

De Schryver, G. -M. & D. J. Prinsloo. 2001a. Fuzzy SF: Towards the ultimate customised dictionary. *Studies in Lexicography* 11 (1): 97—111. [Also published in *Asialex* 2001 *Proceedings*: 141—146. Seoul.]

De Schryver, G. -M. & D. J. Prinsloo. 2001b. 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 *Proceedings of Kiswahili* 2000: 216—242, 372. Dar es Salaam.

Docherty, V. J. 2000. Dictionaries on the Internet: an Overview. In *Euralex* 2000 *Proceedings*: 67—74. Stuttgart.

Dodd, W. S. 1989. Lexicomputing and the dictionary of the future. In James, G. C. A. (ed.), *Lexicographers and Their Works*: 83—93. Exeter.

Grossman, L. 2006. Time's Person of the Year: You, *TIME*, 13 December 2006. Available from http: //www. time. com/time/magazine/article/0,9171,1569514,00. html.

Harley, A. 2000. Cambridge Dictionaries Online. In *Euralex* 2000 *Proceedings*:85—88. Stuttgart.

Hillewaert, S. & G. -M. de Schryver. 2004. *Online Kiswahili (Swahili) – English Dictionary*. Available from http: //africanlanguages. com/swahili/.

Hulstijn, J. H. 1993. When do foreign-language readers look up the meaning of unfamiliar words? The influence of task and learner variables. *The Modern Language Journal* 77 (2): 139—147.

Hulstijn, J. H. & B. T. S. Atkins. 1998. Empirical research on dictionary use in foreign-language learning: survey and discussion. In Atkins, B. T. S. (ed.), *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*: 7—19. Tübingen.

Hulstijn, J. H. & P. Trompetter. 1998. Incidental learning of second language vocabulary in computer-assisted reading and writing tasks. In Albrechtsen, D. *et al.* (eds.), *Perspectives on Foreign and Second Language Pedagogy: Essays Presented to Kirsten Haastrup on the Occasion of her Sixtieth Birthday*: 191—200. Odense.

Jakopin, P. & B. Lönneker. 2004. Query-driven Dictionary Enhancement. In *Euralex* 2004 *Proceedings*: 273—284. Lorient.

Kabuta, N. S. *et al.* 2006 – . *Nkòngamyakù wa Cilubà – Mfwàlànsa /Dictionnaire Cilubà – Français*. Available from http: //www. ciyem. ugent. be/.

Knight, S. 1994. Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The Modern Language Journal* 78 (3): 285—299.

Laufer, B. 2000. Electronic dictionaries and incidental vocabulary acquisition: does technology make a difference? In *Euralex* 2000 *Proceedings*: 849—854. Stuttgart.

Laufer, B. & M. Hill. 2000. What Lexical Information do L2 Learners Select in a Call Dictionary and How Does it Affect Word Retention? *Language Learning & Technology* 3 (2): 58—76.

Löfberg, L. 2002. *Miksi sanat eivät löydy sanakirjasta? Tapaustutkimus MOT Englanti* 4.0 (MA Thesis). University of Tampere.

Mobasher, B., R. Cooley & J. Srivastava. 1999. Creating adaptive web sites through usage-based clustering of URLs. In *Proceedings of KDEX* 1999 (*Knowledge and Data Engineering Exchange*): 19—25. Chicago.

Mobasher, B., R. Cooley & J. Srivastava. 2000. Automatic personalization based on web usage mining. *Communications of the ACM* 43 (8): 142—151.

Müller-Spitzer, C. & C. Möhrs. 2008. First ideas of user-adapted views of lexicographic data exemplified on OWID and *elexiko*. In *Proceedings of COGALEX* 2008 (*COLING* 2008 *workshop on Cognitive Aspects of the Lexicon*) : 39—46. Manchester.

Ne'eman, Y. & R. Finkel. 2004. Rav-Milim Online. *Kernerman Dictionary News* 12 : 28—31.

Nesi, H. 2000. Electronic Dictionaries in Second Language Vocabulary Comprehension and Acquisition : the State of the Art. In *Euralex* 2000 *Proceedings* : 839—847. Stuttgart.

Oberlander, J. , M. O'Donnell, C. Mellish & A. Knott. 1998. Conversation in the museum : experiments in dynamic hypermedia with the intelligent labelling explorer. *New Review of Hypermedia and Multimedia* 4 (1) : 11—32.

OED Online. Available from http : //www. oed. com/.

Popescu-Belis, A. , S. Armstrong & G. Robert. 2002. Electronic Dictionaries - from Publisher Data to a Distribution Server : the DicoPro, DicoEast and RERO Projects. In *Proceedings of LREC* 2002 : 1144—1149. Las Palmas.

Prinsloo, D. J. & G. -M. de Schryver (eds.) . 2000. *SeDiPro* 1.0, *First Parallel Dictionary Sepêdi - English*. Pretoria.

Prinsloo, D. J. & G. -M. de Schryver. 2001. Taking Dictionaries for Bantu Languages into the New Millennium - with special reference to Kiswahili, Sepedi and isiZulu. In *Proceedings of Kiswahili* 2000 : 188—215. Dar es Salaam.

Prószéky, G. & B. Kis. 2002. Development of a Context-Sensitive Electronic Dictionary. In *Euralex* 2002 *Proceedings* : 281—290. Copenhagen.

Pruvost, J. 2003. Some Lexicographic Concepts Stemming from a French Training in Lexicology (1). *Kernerman Dictionary News* 11 : 10—15.

Sato, H. 2000. Multi-Functional Software for Electronic Dictionaries. In *Euralex* 2000 *Proceedings* : 863—870. Stuttgart.

Sinclair, J. M. & P. Hanks. 1987. *Collins COBUILD English Language Dictionary*. London.

Sobkowiak, W. 1999. *Pronunciation in EFL Machine-Readable Dictionaries*. Poznań.

TshwaneLex Dictionary Production System. Available from http : //tshwanedje. com/tshwanelex/.

Van Nieuwenborgh, D. , M. De Cock & D. Vermeir. 2007. An introduction to fuzzy answer set programming. *Annals of Mathematics and Artificial Intelligence* 50 (3) : 363—388.

Varantola, K. 2003. Linguistic corpora (databases) and the compilation of dictionaries. In Van Sterkenburg, P. (ed.), *A Practical Guide to Lexicography* : 228—239. Amsterdam.

Viappiani, P. , P. Pu & B. Faltings. 2002. Acquiring User Preferences for Personal Agents. In *Proceedings of the AAAI Fall Symposium* : 53—59. North Falmouth.

Warburton, Y. 2000. The Oxford English Dictionary - From OED to OED Online. *International Journal of Lexicography* 13 (2) - *EURALEX Newsletter* : 7—8.