Fuzzy SF: Towards the ultimate customised dictionary

Gilles-Maurice de Schryver[*]

DJ Prinsloo[**]

Abstract

Fuzzy SF, a novel concept for an electronic-dictionary package, is presented. In Fuzzy SF, log-file based Artificial Intelligence components enable the implicit retrieval of personalised user feedback with which the package customises each user's own and unique dictionary. To that end, all the data in both the databases and the multimedia (sub)corpora are graded using Fuzzy Sets, so that the package only answers queries on the user's (current) level.

1 Simultaneous Feedback (SF) & Electronic Corpora

Since 1997, the quick compilation of dictionaries within a sound

[*] Department of African Languages and Cultures, Ghent University, Rozier 44, 9000 Ghent, Belgium / gillesmaurice.deschryver@rug.ac.be
G-M de Schryver is currently *Research Assistant of the Fund for Scientific Research – Flanders (Belgium)* and received the 2000 Laurence Urdang Award, administered by Euralex, for implementing Fuzzy SF.

[**] Department of African Languages, University of Pretoria, Pretoria 0002, South Africa / prinsloo@postino.up.ac.za

framework has been our main area of research. This work resulted in the introduction of the theoretical concept of Simultaneous Feedback (SF) (cf. e.g. De Schryver 1999, De Schryver & Prinsloo 2000a, 2000b) and bilingual dictionaries for Cilubà and Sepedi compiled within the framework of this methodology (De Schryver & Kabuta 1997, 1998[2], Prinsloo & De Schryver 2000, De Schryver & Prinsloo *forthcoming*). In a nutshell, Simultaneous Feedback (SF) 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. As such, the target users continuously guide the compilers during the entire compilation process, and the unabated retrieval of feedback can be considered as the main pillar of the methodology. So far, this retrieval of feedback has followed the channels of such standard approaches as (natural) participant observation, formal and informal discussions, anonymous mail survey questionnaires, controlled tests, etc. Through a cross-comparison of the results of the various types of feedback, we attempted to arrive at a representative body of users' desires for each of our particular target user groups. Still, the realisation 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 us to seek a straightforward, automatic, neutral and invisible arbiter. We found that arbiter in the form of electronic dictionaries' (EDs') log files.

According to Moon, the use of electronic corpora *"consolidated into standard dictionary praxis [...] over the period 1986-1996"* (2000: 4). In our own research and our own dictionaries, corpora have always played a crucial role (cf. e.g. De Schryver & Prinsloo 2000c, 2000d, 2000e, Prinsloo & De Schryver 2001a, 2001b).

We will not dwell on how SF through log files and the querying of electronic corpora can contribute to the *compilation* of modern dictionaries. Rather, we would like to show how both these aspects can and should be fully *included in the resulting dictionary* proper. Since we take the gist of SF one step further, through the inclusion of Artificial Intelligence components and Fuzzy Sets, we refer to this new concept as Fuzzy SF. Our aim is that Fuzzy SF should lead to the ultimate customised dictionary. Due to space restrictions, elaborate examples of the different aspects presented in this paper will be given at our web site (<http://www.up.ac.za/academic/libarts/afrilang/elcforall.htm).

2 Fuzzy SF

2.1 A priori, there simply is no dictionary

For presentation purposes, we can begin with the theoretical (and hence idealised) concept of Fuzzy SF. First of all, we should try to make an abstraction of what we perceive when we bring a dictionary to mind (be it a paper version or an electronic one). We must be ready to rethink the very concept itself (yet, paradoxically, at the same time the product must still be a dictionary, and hence, must conform to what a dictionary is according to general lore). Consequently, let us assume we only vaguely know what is wanted as product, namely a tool that helps to retrieve knowledge about language(s). Giving heed to one of today's most often heard requirements for modern reference works, i.e. attention to the user-perspective, that tool must be extremely user-friendly. Not only must the user be able to obtain an answer to a query in as friendly a way as possible, the user must also be enabled to query the tool in the way which suits the user most. As a point of departure of the concept of Fuzzy SF we can say that: *a priori, there simply is no dictionary* with a fixed structure or a fixed access route. Obviously, there is a 'default setting', so we can say that that particular setting is the 'average' dictionary for the expected target user group of the software tool. Yet, even that is not true. In fact, there is a '*set* of default settings'. Indeed, from the first 'handling / query' onwards, the built-in software already decides for which default setting to opt.

2.2 De facto, the lingware acts as a dictionary

We will use the term '(Fuzzy SF) package' to refer to the tool to be developed. Even though, a priori, the package does not contain a (customised) dictionary, there is somehow a collection of multimedia data slots that can be combined into a dictionary, and ultimately the aim is that each user will even be able to use and retrieve a personally tailored reference work. It should be evident that we are talking 'electronic'. The package is a disk (CD-ROM(s), DVD(s), etc.) crammed with data (or the equivalent on an Intranet system or even the Internet). This might be the first drawback; by definition, we are dealing with an electronic product from the outset. In extremis, though, there will be the possibility to insert the disk (or type the URL), and to select one option to print one of the (default) dictionaries. At that point the user will have the equivalent of a product that could simply have been taken from a shelf in a bookstore or library. Since that dictionary will have been created by means of SF, this is not unfavourable as a starting point. Yet, for Fuzzy SF 'no interactive feedback means no customised dictionary'.

2.3 A contrario, the package is a customised dictionary

In order to illustrate how we conceptualise the package, we will now reason backwards, beginning with the envisaged product. The only thing user X knows is that the package contains user X's own customised dictionary. The only thing the package 'knows' is how to launch itself and to await the first signal from user X. As noted, from the start it will be possible for any user to press one key (or click one field) to 'receive' a printout of the chosen (default) dictionary. A bit higher on the scale of sophistication is where the user provides an identifier (i.e. whatever (nick) name the user desires), say 'James', and tells the system *how* the chosen dictionary should be printed. Hence James puts in some generalities dealing with layout but also on the slots to be included vs. those to be omitted. Here it is presupposed that James already has a rather clear picture of: a) what a dictionary is, and b) what he wants his own dictionary to look like. If so, James can acquire or access the package and print his personal dictionary within hours, with his identifier included, or simply use the package onscreen with his preferences. This is inessential and straightforward customisation, with no 'intelligence' whatsoever involved, and actually is a 'bad' way of retrieving feedback, since it uses formal / direct questions. Yet, Fuzzy SF must at least be able to 'do' what the current state-of-the-art EDs can, before providing more.

So let us move to more. First of all, the package aims at dealing with *all* the potential users interested in *one* particular target language (e.g. Sepedi) where various metalanguages (e.g. English, besides a plethora of others) enable the functionality for non-mother-tongue speakers of the target language. Secondly, from the first run onwards, James will simply start using those sections that he can 'understand / read' on the screen. He doesn’t need to tell the system which type of dictionary he wants to use (but he *could* if he so wished). Say James is a Londoner who has only just started to learn Sepedi. For him the option *Pukuntšutlhaloši ya Sesotho sa Leboa* ‘Explanatory Sepedi Dictionary’ won’t mean anything, so he won’t even choose that option. He will start his navigation at an option in English, and the software will immediately pick that up. James, unknowingly, has given indirect feedback to the package. Say James has chosen the English phrase ‘I have a word in Sepedi in front of me and I want to find the English equivalent’. Upon this, the software gives a new screen with various possibilities. James has the possibility of answering a couple of questions, or to phrase his query, or to click different options, or even to type in anything he wants on a blank screen. As an illustration, say he types in the phrase *mahlo a magolo*. If the software’s analysis tools are good enough (not too hard in this case), the package will return ‘big eyes’.

It should be noted that, since the software will have analysed this phrase, this would be an appropriate opportunity to ‘show the analysis’ – if James desires, and using the difficulty-level James needs (decided by the software) or indicates (done by James himself). The package could say something like: “The first word, *mahlo* ‘eyes’, is a noun that starts with the letters ma-, so it belongs to class 6. In order to describe the noun *mahlo* ‘eyes’ an adjective construction is used. To that end, the concord of class 6, *a*, is added following the noun *mahlo*. The description is *-golo* ‘big’, which becomes *magolo* when it takes the prefix of class 6. Thus *mahlo a magolo* is a noun + concord + adjective construction meaning ‘big eyes’.” Alternatively, the analysis could be shown graphically (with codes or phrases). It should be clear from this simple example that such an ‘explanation’ is based on a mould in which the data from the search are simply inserted. Hence, *mahlo* replaces an X, 'class 6' a Y, *-golo* a Z, etc. Even within such 'explanations' there should be the possibility to click on e.g. 'class 6' to obtain a screen in which the noun class system of Sepedi is explained, here focusing on class 6.

Actually, this simple example is much richer in feedback than it appears at first glance. For the Bantu languages, fierce debates have kept scholars busy pondering the best lemmatisation approach for over a century, with some favouring the so-called 'word tradition' for all POSs, others the 'stem tradition', and still others a hybrid approach guided by the type of POS. Now, from James' input the software will 'assume' that James prefers (even though he is not aware of this) the word tradition, and will make a note of that. Say that, during a subsequent query, James types in just *legolo*. In that case the package will return 'big' (and in a grammatical pop-up window that it has taken on the prefix of class 5, etc.) and will give 'extra weight' to the supposition that James wants to work with a word-based dictionary. Yet, if James had typed in *-golo*, the software would immediately have had to reshuffle its weights, giving some credence to the stem tradition now. Just as the 'intelligence' in the package will mark preferences as far as adjectives are concerned (*magolo* and *legolo* vs. *-golo* 'big' above), it will do exactly the same for nouns (*mahlo* (vs. *-ihlo*) 'eyes' above), verbs, MWUs, etc. etc.

What is now the main idea behind this? Say that James, after having utilised the package for a few weeks in the way briefly described, decides to print 'his dictionary'. At that point the software will simply go through all James' preferences (most of which James is unaware of) and print 'James' own customised dictionary'. Or, if James prefers to keep working with an ED, he will 'see' an ED in the format that suits him most. For the example above, this means that both his paper dictionary and screen will either show the adjective 'big' as *-golo* or as *magolo*, *legolo*, etc. The power, especially in the case of Bantu-language dictionaries (for which there has been so much futile debate on the best lemmatisation approaches), is magnificent, for there simply is no best approach anymore. The best approach is precisely the one the user wants. The lexicographers have provided all the possible alternatives, for all possible word / phrase categories, and every single user ends up employing a particular brand of options. Hence, *every single dictionary is tailored to one specific user and hence unique.*

An elementary learner does not remain an elementary learner, but becomes an intermediate one. So James, after having used the package for a few extra months, has possibly triggered the software in such a way that he ends up with a different configuration of preferences. A dictionary produced after a few months of use might very well, and is even bound to, be different from a dictionary produced after just a few weeks' use. With this approach, the very notion of what a dictionary *is* is completely exploded. Even basic notions of what constitutes a 'word', or controversies dealing with words vs. MWUs, or debates on what should and what shouldn't receive lemma-sign status, etc. etc. have become irrelevant. James can search for any string(s) in any field(s) of the database, and conjure up the data of any (combination of) field(s) connected with that. Therefore, *Fuzzy SF effectively explodes the macrostructure.*

2.4 A fortiori, the package is a stratified dictionary

From the moment one explodes the macrostructure, it is hard to refer to the microstructure. Even so, both terms remain on one's lips, and viewed from a certain angle, one notices two objectives when it comes to the 'so-called' microstructure. On the one hand, the package aims to couch the answers to a user's queries in the user's language, i.e. the package utilises explanations which are as close as possible to the user's target-language level (and/or metalanguage level). On the other hand, the articles themselves are formatted entirely to conform to the user's desires.

The first objective is definitely the hardest to implement as the package has the daunting task to 'decipher the quality' of a user's queries, and to decide how to respond, based on this deciphering (the latter, of course, not being the analysis of one single query but of a long battery of inputs). Yet, this point of departure does not seem as hopeless as it might appear. To illustrate this, we can briefly look into three potential candidates for user-level assessment. Firstly, if the average length of a query (expressed in number of spaces and punctuation marks in the query) is long, lexicographers are (i.e. 'the black box is') told that the user is (still) extremely uncertain about the language. At one extreme, if entire sentences (including sub-sentences and numerous punctuation marks) are continuously typed in (as compared to space-less inputs or short phrases), and this over a long period of time, it obviously means the user hasn't made much progress in that particular language. Secondly, if a user keeps searching transparent, yet rare compounds and MWUs, even after years of utilisation, then the software may safely assume that the user (still) hasn't developed the necessary skills to be able to analyse the target language. Thirdly, keeping a log file of all the searches the user has ever made provides invaluable data. If the same successful searches are repeated time and again, and over a long period of time, it obviously enables the software to conclude that poor James isn't making any progress. Conversely, if James repeatedly asks the package the same unsuccessful questions, it indicates that James isn't very inventive when it comes to language. Since we are dealing with a battery of query-assessments, the fact that a user sporadically types in longer queries, and/or occasionally searches transparent yet rare compounds and MWUs, and/or infrequently sins against the log file, does not imply that that user has floating language skills. A dictionary is used and thus queried in many ways; what counts is what deviates from the expected average. In short, a combination of well-designed weighted parameters must enable the software to obtain a likely user's level. And say that James judges that the package's replies are incompatible with his skills, i.e. too simplistic (or too hard), then more advanced (or easier) answers will be just one click away.

The approach just described implies that, for every lemma sign, a set of graded definitions and translation equivalents is available in the black box, together with a set of graded example sentences, a set of graded grammatical pop-up windows, etc. etc. Actually, several dictionaries (databases) are built into one (and linked), and an initially junior user like James should be able to consult the same package throughout his entire life. With his advancing knowledge, he will simply move (not necessarily knowingly) into ever higher-graded slots. In order to enable 'life-time use', the package will of course need to be upgraded regularly, through the acquisition of an extra disk, a simple download, or automatically online. Such upgrades will typically include new and to-be-removed data slots; to-be-swapped sense and example orders; amendments to, additions to, or to-be-dropped markers and labels, etc. etc.

We will now focus on the second objective, the formatting of the articles proper. Available EDs on the market show the way ahead: users can choose which 'fields' to see and which ones to hide, or customise anything that deals with 'layout', and in very rare instances even consult a 'text corpus'. Of these three aspects, the utilisation of corpora in the final product is by far the most underdeveloped one (and, paradoxically, also the most promising). We would want to see a very close interaction between the prepared / built-in slots of the articles, and those the user can conjure up through querying the attached (sub)corpora. It is well known that lexicographers cannot come up with all possible answers to all potential questions, however if: a) a series of corpora is attached to the ED, b) a handy corpus query tool is provided with the package, and c) 'answers' can be 'shown' in an easy and structured way, then users can get all the (statistical) information they want from the ED package, even if it was not prepared / built-in by the compilers in the first place. Since the (sub)corpora must be integrated in such a way that users of all levels can easily query them, it seems logical to have query tools with different interface levels too. Also, ED corpora needn't be restricted to 'text corpora'. Indeed, in EDs one has the potential to deal with truly multimedia (sub)corpora, fully integrating and fully cross-referencing ED text, computer graphics and audio. Consequently, multimedia (sub)corpora are part and parcel of the ED package. All in all, *querying multimedia (sub)corpora interactively is a fully fledged component of each and every ED article*.

2.5 A pari, the package is a polyaccessible dictionary

We restricted the presentation to one user, James, who performed one specific handling of the package. It goes without saying that each package is a multimedia 'family' reference work. Hence, identifying oneself at the start of each session (not indispensable, but highly recommendable as we've seen) will enable one and the same CD-ROM package to be used by and customised for say a dozen users. (In addition, the default dictionaries will always be usable by anyone.) On Intranet / Internet versions, the number of users is only limited by the size of memory available on the servers.

From the exposition so far, one might get the impression that we are solely dealing with a semasiological presentation of language data. It is however true that, since the macrostructure is exploded in Fuzzy SF, the very notion 'semasiological presentation' is weakened. It might for instance turn out that the software quickly picks up that a particular user tends to favour onomasiological searches. Say Sarah (James' sister) often types in requests such as 'List me the vehicles on two wheels', 'What are the traditional musical instruments played in the area around Pietersburg?', 'Show me how family members call one another', 'Which fruits are only given to animals?', etc. At such a point the software 'realises' that Sarah would in fact like to use the package as if it were organised thematically. The ED will from then onwards present the data in this way. As always, the user will be enabled to *ask* the package for such a presentation too, but the idea here is to provide this possibility for users who are not familiar with this presentation option. In short, Fuzzy SF *harmonises both semasiological and onomasiological approaches to a language's lexicon.*

Finally, it should be clear that the internal structure of a Fuzzy SF ED (i.e. the set of databases, their indexation systems and their hyperlinks) is a complex, multidimensional and fully linked network. Anything that is possibly cross-referenceable is also actually cross-referenced. On top of this, the cross-references also extend to front and back matter; to outside matter: extra reading material, Internet URLs, e-links to the compilers (for suggestions of to-be-added items for instance), etc.; to help files for the dictionary aspects themselves, help files revolving around the package's technical aspects, help files providing a kind of 'user's guide', etc.; and finally to the various multimedia (sub)corpora. As far as the latter are concerned, the primary corpora to be designed are obviously those for the target language. Yet, if time and money allows, energy should also be devoted to the development and full inclusion of corpora for the metalanguages.

2.6 A posteriori, Fuzzy SF as a dictionary of the next millennia

Compared to any principle currently utilised in dictionary-making and compared to any existing multimedia reference work, the following 10 key novelties of Fuzzy SF are either absent from or would constitute important improvements over what is done or available at present: **1.** Parallel packages are released throughout the endeavour to compile the main package, answering an urgent desideratum to provide users with dictionaries *now*; **2.** Since the package is thoroughly 'tested' before it ever gets launched, it contains user feedback right from the start, and once it is used it (preferably) gathers its feedback indirectly, informally and unknowingly, successfully eliminating any barriers between compilers and users; **3.** The package offers fully fledged default dictionaries, and, additionally, each user can retrieve a personally tailored reference work in print or in ED format; **4.** The package is a family reference work that can be customised for several users, and is continuously re-customised for each single user over time; **5.** The package is primarily descriptive, and includes the possibility for user-initiated modifications; **6.** The notion of 'lemma sign' has become volatile as virtually anything can have lemma-sign status, resulting in a fusion of the macro- and microstructural levels; **7.** Both the access to and the visual presentation of the data slots are such that the distinction between onomasiological and semasiological dictionaries tends to disappear; **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 dictionary, a monolingual, bilingual or hybrid dictionary, and this 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) which automatically generate data 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, culled automatically or interactively from the sub(copora), or supplemented by the user – are hyperlinked in the package on all levels and in all directions.

All in all, we are convinced that Fuzzy SF has the potential to combine some revolutionary concepts with centuries-old traditions. The latter have received no attention, but it is obvious that Fuzzy SF should incorporate all the good aspects to be found in the dictionaries of the past millennia. Integrating Fuzzy SF with those traditions is likely to provide a sound framework for dictionaries of the next millennia.

References

De Schryver, Gilles-Maurice and Ngo S. Kabuta (1997). *Lexicon Cilubà–Nederlands, Een circa 2500-lemma's-tellend strikt alfabetisch geordend vertalend aanleerders-lexicon met decodeer-functie ten behoeve van studenten Afrikaanse Talen & Culturen aan de Universiteit Gent*. Ghent: Recall.

De Schryver, Gilles-Maurice and Ngo S. Kabuta (1998^2). *Beknopt woordenboek Cilubà–Nederlands & Kalombodi-mfùndilu kàà Cilubà (Spellingsgids Cilubà), Een op gebruiksfrequentie gebaseerd vertalend aanleerderslexicon met decodeerfunctie bestaande uit circa 3.000 strikt alfabetisch geordende lemma's & Mfùndilu wa* *myakù ìdì ìtàmba kumwèneka (De orthografie van de meest gangbare woorden)*. Ghent: Recall.

De Schryver, Gilles-Maurice (1999). Bantu Lexicography and the Concept of *Simultaneous Feedback*, Some preliminary observations on the introduction of a new methodology for the compilation of dictionaries with special reference to a bilingual learner's dictionary *Cilubà–Dutch*. (Unpublished MA dissertation, Ghent University.)

De Schryver, Gilles-Maurice and D.J. Prinsloo (2000a). Dictionary-Making Process with 'Simultaneous Feedback' from the Target Users to the Compilers. In Ulrich Heid, Stefan Evert, Egbert Lehmann and Christian Rohrer (*eds*.) (2000). *Proceedings of the Ninth Euralex International Congress, EURALEX 2000, Stuttgart, Germany, August 8th-12th, 2000*: 197–209. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

De Schryver, Gilles-Maurice and D.J. Prinsloo (2000b). The Concept of 'Simultaneous Feedback': Towards a New Methodology for Compiling Dictionaries. *Lexikos 10*: 1–31.

De Schryver, Gilles-Maurice and D.J. Prinsloo (2000c). The compilation of electronic corpora, with special reference to the African languages. *Southern African Linguistics and Applied Language Studies 18/1-4*: 89–106.

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

De Schryver, Gilles-Maurice and D.J. Prinsloo (2000e). Electronic corpora as a basis for the compilation of African-language dictionaries, Part 2: The *microstructure*. *South African Journal of African Languages 20/4*: 310–330.

De Schryver, Gilles-Maurice and D.J. Prinsloo (*eds*.) *forthcoming*. *SeDiPro 2.0, Second Parallel Dictionary Sepêdi–English*. Pretoria: University of Pretoria.

Moon, Rosamund (2000). Congress Report: EURALEX 2000. *ElsNews, The Newsletter of the European Network in Human Language Technologies 9/3*. Available at: <http://www.elsnet.org/publications/elsnews/9.3.pdf>.

Prinsloo, D.J. and Gilles-Maurice de Schryver (*eds*.) (2000). *SeDiPro 1.0, First Parallel Dictionary Sepêdi–English*. Pretoria: University of Pretoria.

Prinsloo, D.J. and Gilles-Maurice de Schryver (2001a). Corpus applications for the African languages, with special reference to research, teaching, learning and software. *Southern African Linguistics and Applied Language Studies 19/1*: 111–131.

Prinsloo, D.J. and Gilles-Maurice de Schryver (2001b). Monitoring the Stability of a Growing Organic Corpus, with special reference to Sepedi and Xitsonga. *Dictionaries: Journal of The Dictionary Society of North America 22*: 85–129.