A Session with Glosser-RuG

D.A.Dokter and J.Nerbonne

April 28, 1997

1 Introduction

Glosser-R*u*G is one of the demonstrators developed within the GLOSSER project[1]. This project aimed at applying state-of-the-art techniques in *natural language processing* to the paradigm of *communicative Computer Assisted Language Learning* (CALL) [13] applications. Software was developed to facilitate the task of reading a foreign language by providing information on words. Techniques that were applied in this project include morphological analysis, part-of-speech (POS) disambiguation, aligning bilingual corpora, World-Wide Web technology, and indexing. The project vision foresees two main areas where GLOSSER applications can be used. First, in language learning and second, as a tool for users that have a bit of knowledge of a foreign language, but can not read it easily or reliably. The latter group might not be trying to learn, only to cope with a specific text. A user might, for instance, need to read a software manual, that contains a number of unfamiliar words. GLOSSER provides the user (or learner) with a means of being able to look up information on unfamiliar words in a straightforward and user-friendly manner. Software has been developed for English/Estonian, English/Bulgarian, English/Hungarian and French/Dutch assistance. This paper describes the French-Dutch demonstrator[2].

2 Motivation

If a rudimentary level of instruction in foreign-language grammar is assumed, then a great deal of the learning required in order to be able to read texts in this language is simply vocabulary learning, which is best pursued in lexical context [10, 8]. GLOSSER makes this as easy and accurate as possible: for virtually all words that frequently occur in texts, context is provided, by means of dictionary entries and examples of word use in especially collected (bilingual) corpora. Moreover, this information is accessed in a quick, straightforward manner with an integrated environment. The project has developed demonstrators as a proof of concept, and, in order to promote use, demonstrators run on both UNIX and Windows '95. The prototypes have proven sufficiently robust to support reading of essentially all non-specialized texts. They have further permitted a pilot user study which is being followed up by broader usability studies at two sites. The initial results showed that users enjoyed the 'intelligent dictionary' kind of help the program offers and were a bit faster in reading a text than users of hand-held dictionaries on the same text [1] (this volume).

The demonstrators have been tested by students, but they might also be put to use to support reading directly by people who are not engaged in formal language instruction, or perhaps not even primarily interested in improving their foreign language ability. Given our emphasis on automatic methods applicable to arbitrary texts, a spin-off in support for translations is conceivable.

GLOSSER distinguishes itself from many CALL programs by its emphasis on *language use* as opposed to *drill and test*, by its ability to support nearly any level of text difficulty, and by its emphasis on effectively removing the tedium of dictionary use from intermediate language learning. See [7] for more details on the CALL background against which GLOSSER was developed.

[1] COPERNICUS grant 343

[2] The demonstrator for the other language pairs is described in [5].

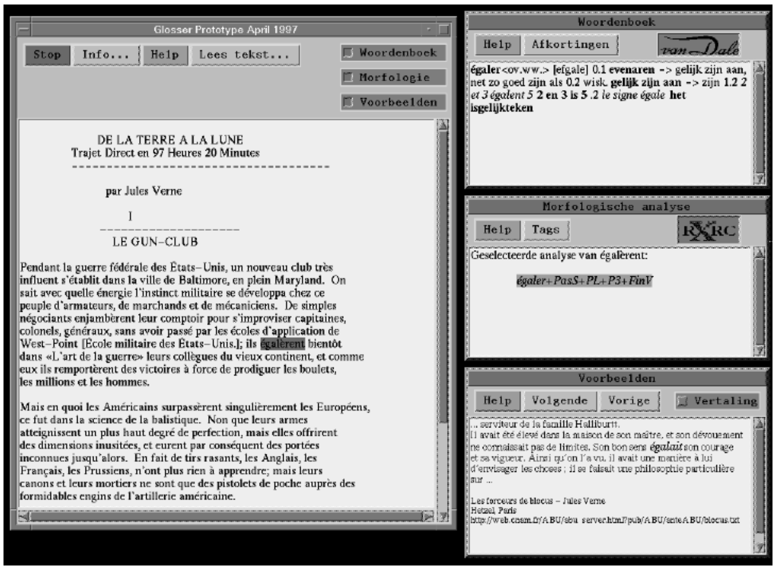

Figure 1: *The front-end as it is displayed on the screen. The window providing help is omitted*

3 Technical Realization

Glosser-R<u>u</u>G is implemented on the UNIX platform, and facilitates the reading of French texts for Dutch speaking students. Four sources of information are available on words: morphological analysis, POS-disambiguation, a dictionary and examples of word use in especially collected corpora. The current demonstrator is implemented completely in the Tcl/Tk scripting language [6], ensuring easy rewriting of parts of the program, rapid prototyping and portability of the source. Also, although the use of a scripting language obviously slows down processing in relation to compiled code, lookup speed is still sufficiently fast: a single lookup including all sources of information takes approximately 2 seconds (see [2] for details). Most of this time is consumed by morphological analysis, which is an external, compiled program.

3.1 Front-end

The front-end of Glosser-R<u>u</u>G, displayed in Figure 1 consists mainly of four separate windows. The main window that is popped up when Glosser-R<u>u</u>G is started provides the general control of the application, a read-only editor and three on/off-switches for controlling the specific sources of information that are to be used for word-lookup. The other windows are used for display of the three different sources: a dictionary, morphological analysis/ POS-disambiguation, and examples. The window providing examples actually consists of two separate windows, one for display of the example, the other for the aligned translation. Apart from these main windows, there is a separate window that provides help on the different components of the application, in a hypertext fashion.

The sources used for a single lookup can be specified by the user at any stage during a session. Other features (buttons) include a hypertext oriented help function that is shown in a separate window and a pop-up menu that shows the files in the current directory. Clicking on a file loads it into the editor to be processed.

3.2 Morphological Analysis

Morphological analysis/POS-disambiguation is directly informative to the user but also indispensable. It is used to find the underlying lexemes of words as they appear in the text, since in general dictionaries do not provide entries for inflected forms. Also, the part-of-speech that the word constitutes in the text is used to be able to choose the right entry in the case of homographs in the dictionary and examples. Furthermore, an index of the corpora used to provide examples also exploits lexemes as entries, to ensure that a wide variety of examples is provided (see subsection 3.4. GLOSSER was

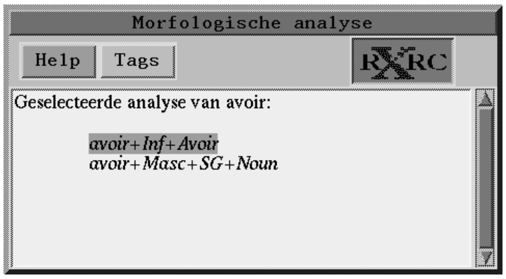

Figure 2: *Morphological analysis.*

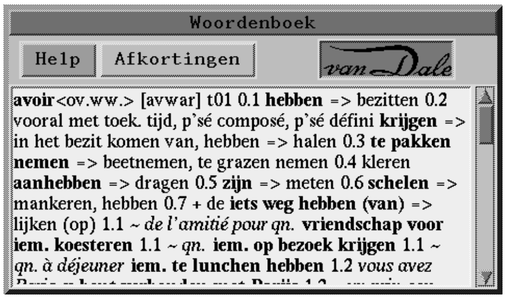

Figure 3: *A dictionary entry in Glosser-RuG.*

fortunate to be able to use a state-of-the-art morphological analysis/POS-disambiguation package from Rank Xerox Research Centre: Locolex. An example analysis is shown in Figure 2.

In general, a large number of words can have different grammatical meanings. *Locolex* incorporates a stochastic POS tagger which it employs to disambiguate. In case *Locolex* is wrong (which is possible, but quite unlikely), the user is free to specify an alternative morphological analysis, which is then looked up in the dictionary and examples index.

3.3 Dictionary

Glosser-RuG was likewise fortunate in obtaining the Van Dale dictionary *Hedendaags Frans* [9]. Figure 3 illustrates the front-end of the dictionary within Glosser-RuG.

For dictionary lookup lemmata are used, as generated by the morphological analysis, as well as the POS of the word as determined by Locolex. The latter feature implies that the correct entry is found, for words with multiple entries (due to different possible syntactic categories), that is, in accordance with the POS the word is tagged with.

3.4 Examples

To provide users with a rich variety of examples, a large corpus was needed, consisting of a number of different kinds of texts, for instance, literature, technical, political, etc. Also, as many *bilingual* examples as possible had to be provided, to ensure easy understandability of how words can be used. For the corpus the GLOSSER project has relied on specialized corpus projects, such as the ECI [4] and MULTEXT [11] for bilingual corpora, although some work in (re)aligning the texts was needed.

In order to determine the size of corpus needed, experiments have been conducted with a frequency list of the $10,000$ most frequent word forms in French. A corpus of 2 MB contained 85% of these, and a corpus of 6 MB 100% [12]. The goal for GLOSSER is 100% coverage of the words (lemmata) found in the $30,000$-word dictionaries, and 100% coverage of the most frequent $20,000$ words for each language involved. The current corpus size for Glosser-RuG is 5 MB in monolingual, 3 MB in bilingual text (that is, the French text), accounting for 16701 different lexemes.

As the corpus grows, the time for incremental search likewise grows linearly. When the average search time grew to several seconds (on a 70 MIPS UNIX server), it became apparent that some sort of indexing was needed. The texts were then indexed by determining the lemmata and POS of the individual words using the same morphological analysis method as described in section 3.2, and

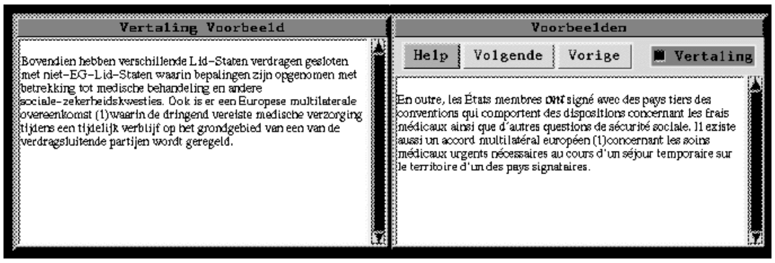

Figure 4: *A bilingual example in Glosser-RuG.*

creating an index of N occurrences of each lemma and POS thus found [3]. The index thus provides a link between the lemmata and the full, possibly inflected forms. Lexeme-based indexing provides not only further occurrences of the same string, but also for inflectional variants of the word. If the selected word is **livre+Masc+SG+Noun** for instance, the search should find other tokens of this and also tokens of the plural form **livres**. It is clear that this improves the chance of finding examples of a given lexeme immensely. Examples are displayed, with a reference to the source (if available), in the "examples" window, as shown in figure 4. If the example has been found in a bilingual text, the user can 'pop-up' the translation from the examples window.

4 Functionality

This section describes the general features implemented in Glosser-R*u*G, constituting much of the functionality of the application. Most features have been developed in accordance with user demands, as derived from practical use.

4.1 Automatic Word Selection

One of the main problems in user-application interaction is the robustness of the code in relation to user actions. In general, it is impossible to account for all possible input. Therefore, especially in applications that rely for a large part on how comfortable and trusting a user feels with it, it is essential that the user can only interact with the application on a very basic level, that is, within a clearly predefined set of actions. The main interaction with Glosser-R*u*G is the selection of words that the user wants to look up. A very restricted set of character strings in texts constitute words that can actually be found in dictionaries. It is considered useless to look up numbers, punctuation, half-words, names etc. Also, user-determined selection requires sensitive specifications on the part of this user, such as marking the start of the selection, marking the end etc. For this reason, control was removed from the user and selection is completely automated. If the mouse cursor is on a string in the text that is defined as a word in the application, this word is highlighted. The only thing a user has to do (and *can* do), is simply click with the mouse causing the word to be looked up.

4.2 Real 'Glossing'; Making Notes

Typically, readers make notes of words in foreign text that have been looked up in a dictionary, in text comprehension. Some beginning readers first look up all unfamiliar words, in order to combine translations in a next stage. Glosser-R*u*G therefore offers the possibility of adding translations within the text. A user can click on any translation provided by the dictionary, which is than automatically inserted in the text, directly following the word that was looked up. Clicking on an inserted translation removes it.

4.3 String-Based Word-Sense Disambiguation

A very primitive notion of word-sense disambiguation is implemented in the application: in general, a dictionary offers a range of examples of lexical contexts in which the entry may occur, and in which context it is translated in a specific way. For example, the entry *mondial* in the Van Dale dictionary provides the example *guerre mondiale*, translated as *wereldoorlog* (world-war). Glosser-R*u*G exploits this feature. Whenever a lexical context is found in the text that is also provided in the dictionary, the example in the dictionary is highlighted. This reduces the time needed by the user for selection of the correct translation.

4.4 Help

Glosser-R*u*G includes a hypertext help-function, that can be popped up from different windows of the application, to show immediate context-sensitive help. The user can of course also browse through the help text by clicking on hypertext links. This function, in combination with the limited interaction between the user and the program, makes Glosser-R*u*G sufficiently self-explanatory, such that users or learners are not distracted from the main function of the program by technical matters.

5 Conclusions

Glosser-R*u*G was developed with the philosophy of exploiting available NLP technology wherever possible. Demonstrators like Glosser-R*u*G show that valuable tools for *communicative* CALL are feasible given the current state of technology.

Acknowledgements

Glosser-R*u*G is part of the GLOSSER project, COPERNICUS grant 343. We would like to thank Lily Grozeva-Schürcks from the University of Groningen and Felice Portier of the *Centre Culturel Français* and her students for valuable criticism on previous prototypes. We also like to thank Petra Smit for her work for GLOSSER.

References

[1] L.Schurcks-Grozeva D.A.Dokter, J.Nerbonne and P.Smit. Glosser-R*u*G; a User Study. In *Language Teaching and Language Technology*, Groningen, The Netherlands, 1997.

[2] D.A. Dokter. Glosser-R*u*G, prototype december 1996. Techreport, Alfa-informatica, University of Groningen, Groningen, The Netherlands, 1997.

[3] D.A. Dokter. Indexing corpora for glosser. Techreport, Alfa-informatica, University of Groningen, Groningen, The Netherlands, 1997.

[4] ECI. *European Corpus Initiative (ECI) Multilingual Corpus I.* http://www.elsnet.org/resources/eciCorpus.html.

[5] Glosser. Glosser, final report. Final project report, Alfa-informatica, University of Groningen, Groningen, The Netherlands, 1997.

[6] J.K.Ousterhout. *Tcl and the Tk Toolkit.* Addison-Wesley, Reading, MA, USA, 1994.

[7] J.Nerbonne and P.Smit. Glosser-R*u*G: in Support of Reading. In *COLING96*, Groningen, The Netherlands, 1996.

[8] G. Krantz. *Learning Vocabulary in a Foreign Language; A Study in Reading Strategies.* PhD thesis, University of Göteborg, Göteborg, Sweden, 1990.

[9] Van Dale Lexicografie. *Handwoordenboek Frans-Nederlands, 2e druk.* VanDale Lexicografie b.v., Utrecht, 1993.

[10] J.-A. Mondria. *Vocabulaireverwerving in het vreemde-talenonderwijs; De effecten van context en raden op de retentie.* PhD thesis, University of Groningen, Groningen, The Netherlands, 1996.

[11] MULTEXT. *Multilingual Text Tools and Corpora.* http://www.lpl.univ-aix.fr/projects/multext/.

[12] A. van Slooten. Searching and quoting examples of word-usage in french language corpus. Techreport, Alfa-informatica, University of Groningen, Groningen, The Netherlands, 1996.

[13] M. Warschauer. Computer-assisted language learning: An introduction. In S. Fotos, editor, *Multimedia language teaching*, pages 3–20. Logos International, Tokyo, 1996.