Applied NLP for Language Learning

D.A.Dokter and J.Nerbonne, Alfa-Informatica, BCN
P.O.Box 716, Rijksuniversiteit Groningen, NL 9700 AS Groningen

{dokter,nerbonne}@let.rug.nl

Introduction

Glosser-R*u*G is a demonstrator developed in the GLOSSER project,[1] which applied state-of-the-art techniques in *natural language processing* (NLP) to facilitate the reading of foreign languages. Several information sources—always at the word level—were brought to bear, including morphological analysis, part-of-speech (POS) disambiguation, aligning bilingual corpora and indexing.

Glosser-R*u*G proves useful to language learners and also to users that know a foreign language a bit, but cannot read it easily or reliably. The latter group might not be trying to learn a specific language, only to cope with it. The application allows the user look up information on unfamiliar words in a straightforward and user-friendly manner. Glosser-R*u*G can therefore be viewed as an intelligent dictionary. Software has been developed for English/Estonian, English/Bulgarian, English/Hungarian and French/Dutch assistance. This paper describes the French-Dutch demonstrator.[2]

Motivation

The debate about the theoretical and practical potential of NLP is still going on. However, it is clear that significant progress has been made and that practical applications are currently feasible. Applications on the market include automatic-translation tools, limited speech-recognition, basic NL-interfaces, NL-to-SQL translation, etc. The practical applications demonstrate the reliability of current techniques, and also stimulate further research. Interestingly, practical techniques contrast with theory: modest, but reliable technology may render significant service. The GLOSSER project has focused on an application for Computer-Assisted Language Learning (CALL).

If a rudimentary level of ability in a foreign-language grammar has been attained, then a great deal of the further learning required concerns vocabulary, which is best pursued in lexical context [Mon96, Kra91]. GLOSSER makes this as easy and accurate as possible: for virtually all words that frequently occur in texts, information on the words as they are used in context is provided—in the form of the dictionary and through examples of word use in especially collected (bilingual) corpora. The project has developed demonstrators on both UNIX and Windows '95. The prototypes have

[1] COPERNICUS grant 343, 1995-97. Other partners were Rank Xerox (Grenoble), University of Tartu, the Bulgarian Academy of Science, and Morphologic (Budapest).

[2] The demonstrator for the other language pairs is described in [Glo97].

proven sufficiently robust to support reading of essentially all non-specialized texts. They have further permitted a user study. The initial results show 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 [DNSGS98]. Given our emphasis on automatic methods applicable to arbitrary texts, a spin-off in support for translations is conceivable.

Technical Realization

Glosser-R*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. All sources rely on morphological analysis and indexing techniques, which are implemented in C. Other modules, including the interface and communication are implemented in the Tcl/Tk scripting language [Ous94], ensuring easy rewriting, rapid prototyping and portability. Script language code is slow, but overall speed is still good: a single lookup of all sources of information takes approximately 2 seconds (see [Dok97a] for details), mostly in morphological analysis.

Front-end and Morphological Analysis

The front-end of Glosser-R*u*G, displayed in Figure 1 consists mainly of four separate windows. The main window (left) provides the general control, a browser (read-only editor) and three on/off-switches for controlling the specific sources of information provided. The other windows display a dictionary enrty, morphological analysis with 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. Finally, there is a separate help window.

Glosser-R*u*G's user interface tries to be helpful. First, words in the text that are currently under the cursor (available for look-up) are automatically highlighted. Second, users can make notes in the original text, to prevent the necessity of more than one lookup.

Morphological analysis/POS-disambiguation is directly informative to the user but also crucial to other processes. It is used to find the underlying lexemes (dictionary forms) of words, since in general dictionaries do not provide entries for inflected forms such as *crois, croyons, cruvait, cru* (all forms found under *croire*). The part-of-speech (verb, noun, etc.) allows the program to choose the right dictionary entry in the case of ambiguity. Finally, the corpora also make use of morphological analysis—this allows a lexeme-based index (instead of string-based) and increases the efficiency of the corpus (more examples of lexemes are found because inflected forms are found (see below on examples).

GLOSSER was fortunate in having state-of-the-art software for morphological analysis and POS-disambiguation from Rank Xerox Research Centre: Locolex. An example analysis is shown in Figure 1, middle right window. *Locolex* incorporates a stochastic POS tagger for disambiguation. In case *Locolex* disambiguates incorrectly (quite infrequently), the user may specify an alternative morphological analysis, which is then looked up in the dictionary and examples index.

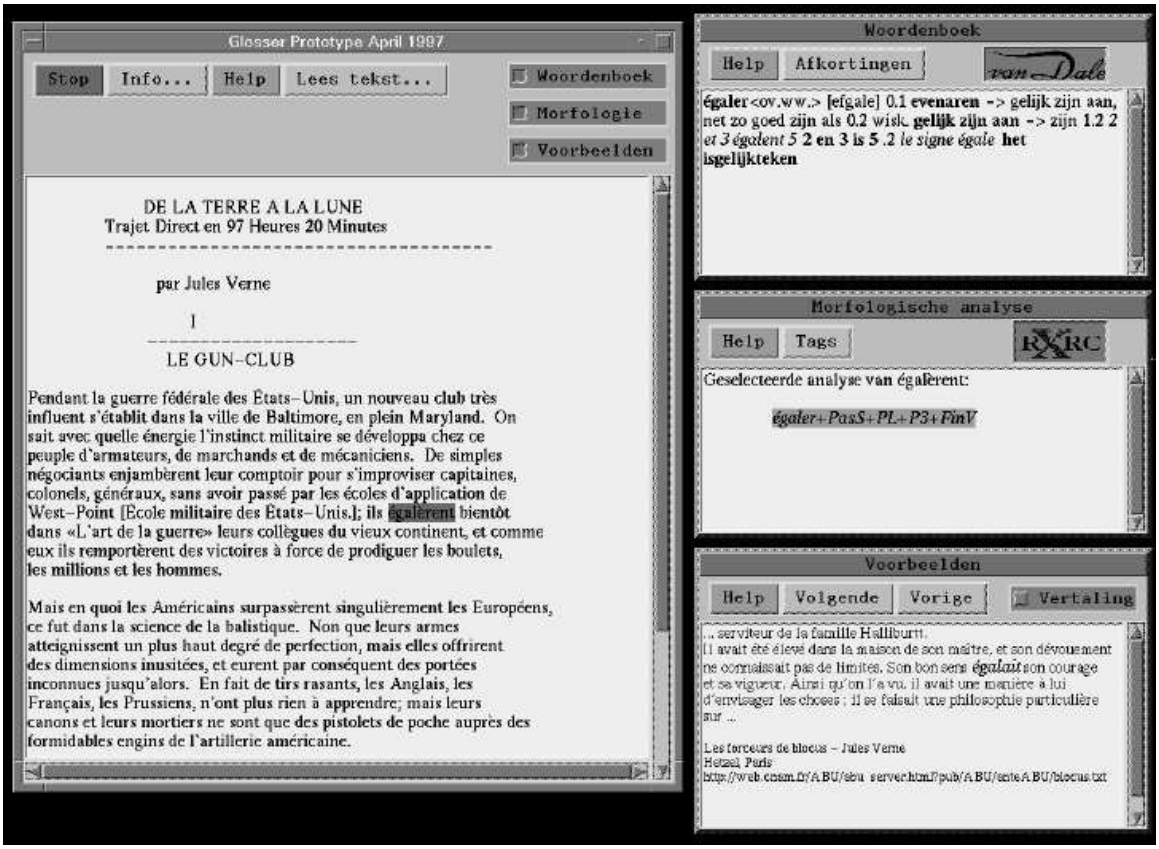

Figure 1: *GLOSSER's front-end, omitting the Help window.*

Dictionary and Examples

Glosser-R*u*G was likewise fortunate in obtaining the Van Dale dictionary *Hedendaags Frans* [vD93]. Figure 1, upper right window illustrates the front-end of the dictionary within Glosser-R*u*G. For dictionary lookup lexemes and POS are used (as generated by the morphological analysis). The POS allows improves the accuracy of dictionary lookup, which otherwise suffers under grammatical ambiguity.

To provide a rich selection of examples, a large and varied corpus was needed, including instance, literary, technical, political, and other prose. *Bilingual* texts were particularly attractive. GLOSSER relied partly on specialized corpus projects, such as the ECI [ECI] and MULTEXT [MUL] for bilingual corpora. Also, a tool was developed for aligning bilingual corpora [PM98].

The current corpus size for Glosser-R*u*G is 5 MB in monolingual, 3 MB in bilingual text (that is, the French text), including 16,701 different lexemes. The texts are indexed by determining the lemmata and POS of the individual words using the same morphological software described above. An index ([Dok97b]) links lemmata to full, possibly inflected forms. Lexeme-based indexing indexes inflectional variants to a single lexeme (dictionary form). A search for examples of *croire* turns up the nearly 100 inflected variants. 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 2. If the example has been found in a bilingual text, the user can 'pop-up' the translation from the examples window.

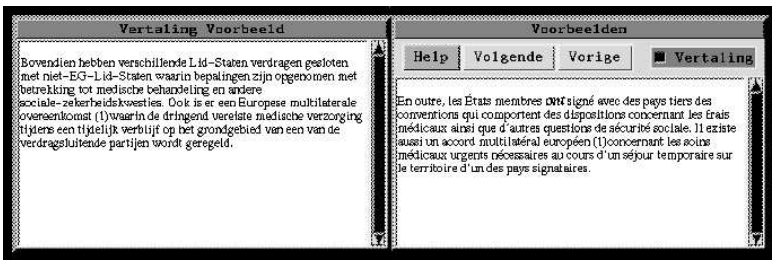

Figure 2: *A bilingual example in Glosser-RuG.*

References

[DNSGS98] Duco Dokter, John Nerbonne, Lily Schurcks-Grozeva, and Petra Smit. Glosser-RuG: A user study. In Sake Jager, John Nerbonne, and Arthur van Essen, editors, *Language Teaching and Language Technology*, pages 169–178, Lisse, to appear, 1998. Swets and Zeitlinger.

[Dok97a] Duco Dokter. Glosser-RuG, Prototype December 1996. Techreport, Alfa-Informatica, University of Groningen, 1997.

[Dok97b] Duco Dokter. Indexing corpora for glosser. Techreport, Alfa-Informatica, University of Groningen, 1997.

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

[Glo97] Glosser. Glosser, final report. Final project report, Alfa-Informatica, University of Groningen, 1997.

[Kra91] Gösta Krantz. *Learning Vocabulary in a Foreign Language.* PhD thesis, Göteborg, 1991.

[Mon96] Jan-Arjen Mondria. *Vocabulaireverwerving in het vreemdetalenonderwijs: De effecten van context en raden op de retentie.* PhD thesis, University of Groningen, Groningen, The Netherlands, 1996.

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

[Ous94] J. K. Ousterhout. *TCL and the TK Toolkit.* Addison-Wesley, Reading, Mass., 1994.

[PM98] Elena Paskaleva and Stoyan Mihov. Second language acquisition from aligned corpora. In Sake Jager, John Nerbonne, and Arthur van Essen, editors, *Language Teaching and Language Technology*, pages 43–52, Lisse, to appear, 1998. Swets and Zeitlinger.

[vD93] van Dale. *Handwoordenboek Frans-Nederlands + Prisma, 2e druk.* Van Dale Lexicografie b.v., 1993.