Reading more into Foreign Languages

John Nerbonne
Alfa-informatica, BCN
University of Groningen
9700 AS Groningen
The Netherlands
nerbonne@let.rug.nl

Lauri Karttunen
Rank Xerox
6, Chemin de Maupertus
38240 Meylan, France
Lauri.Karttunen@grenoble.rxrc.xerox.com

Elena Paskaleva
Linguistic Modeling Lab
Bulgarian Academy of Sciences
25A, Acad. G. Bonchev St.
Sofia 1113, Bulgaria
hellen@bgcict.acad.bg

Gábor Prószéky
Morphologic
Németvölgyi út 25
H-1126 Budapest, Hungary
proszeky@morphologic.hu

Tiit Roosmaa
Computer Science
University of Tartu
2 J. Liivi St.
EE2400 Tartu, Estonia
roosmaa@math.ut.ee

Abstract

GLOSSER is designed to support reading and learning to read in a foreign language. There are four language pairs currently supported by GLOSSER: English-Bulgarian, English-Estonian, English-Hungarian and French-Dutch. The program is operational on UNIX and Windows '95 platforms, and has undergone a pilot user-study. A demonstration (in UNIX) for *Applied Natural Language Processing* emphasizes components put to novel technical uses in intelligent computer-assisted morphological analysis (ICALL), including disambiguated morphological analysis and lemmatized indexing for an aligned bilingual corpus of word examples.

1 Motivation

GLOSSER applies natural language processing techniques, especially morphological processing and corpora analysis, to technology for intelligent computer-assisted language learning (ICALL).

The project vision foresees that intermediate language learners/users of e.g., English, perhaps a native speaker of Bulgarian, might be reading on the screen, perhaps a software manual. We imagine such a user encountering an unknown word or an unfamiliar use of a known word, e.g., *reverts* as in:

"This action *reverts* the buffer to the form stored on disk"

The user can click the mouse on a word to invoke online help (following a dynamically generated hyperlink), which provides:

1. a morphological parse, separating 'revert' and 's', together with an explanation of the significance of the inflection ('s')—3rd person singular present tense;
2. the entry to the word 'revert' in a bilingual English/Bulgarian dictionary or a monolingual English one;
3. access to similar examples of the word in online bilingual corpora; and
4. an audible pronunciation. (This is included only to demonstrate further capabilities, and is available only for a small number of words.)

The example of English for Bulgarians is chosen for illustration. Software has also been developed for English/Estonian, English/Hungarian and French/Dutch.

If we assume a rudimentary level of instruction in foreign-language grammar, then a great deal of the learning required in order to read is simply vocabulary learning, which is best pursued in context (Krashen, 1989; Swaffar, Arens, and Byrnes, 1991). GLOSSER makes this as easy and accurate as possible: vocabulary is always presented in context, moreover in texts which the teacher or student may choose. Analyses, dictionary explanations and further examples are but a mouse click away.

The project has developed demonstrators as a proof of concept, and, in order to promote use, the 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 permited 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 and were a bit faster in reading.

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 langauge ability. Given our emphasis on automatic methods applicable to arbitrary texts, a spin-off in support for translations is conceivable.

(Nerbonne and Smit, 1996) provides more on the ICALL background against which GLOSSER was developed. GLOSSER distinguishes itself from many ICALL 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.

2 Technical Realization

GLOSSER is designed with four major components, which are sketched in Figure 1.

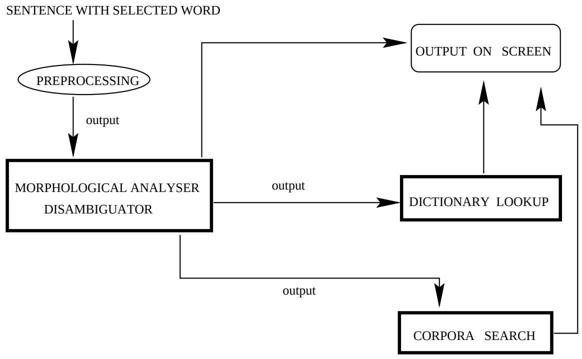


Figure 1: GLOSSER Architecture connects modules for morphological analysis and disambiguation, dictionary access, and (indexed) corpora search with an output module. The "suggestive" pronunciation module is not shown.

The core modules provide the information noted in Section 1, (1-3): morphology, bilingual dictionary entry, and examples from use. A fourth (user-interface and display) module controls interaction with the user and formats the information provided. Among other things, it allows the range of information to be tailored to individual preference.

The usefulness of the first two sorts of information is evident. We chose to include the third sort as well because corpora seemed likely to be valuable in providing examples more concretely and certainly more extensively than other sources. They may provide a sense of collocation or even nuances of meaning.

The realization of these design goals required extensive knowledge bases about morphology and the lexicon.

- Most crucially, the morphological knowledge base provides the link between the inflected forms found in texts and the "citation forms" found in dictionaries (Sproat, 1992). LEMMATIZATION recovers citation forms from inflected forms and is a primary task of morphological analysis. A substantial morphological knowledge base is likewise necessary if one is to provide information about the grammatical significance of morphological information.

  The only effective means of providing such a knowledge base is through morphological analysis software. Even if one could imagine storing all the inflected forms of a language such as French, the information associated with those forms is available today only from analysis software. The software is needed to create the store of information.

  Even apart from this: people occasionally create new words. Analysis programs can provide information about these, since most are formed according to very general and regular morphological processes.
- Obviously, the quality of the online dictionary is absolutely essential. The only feasible option is to use an existing dictionary. Our investigative user study indicates that the dictionary is the most important factor in user satisfaction.
- The essential design questions vis-à-vis the corpus were (i) how large must the corpus be in order to guarantee a high expectation that the most frequent words would be found; and (ii) what sort of access techniques are needed on a corpus of the requisite size—given that access must succeed within at most a very few seconds.

  We tried to use texts from a variety of genres, and we attempted (with some limited success) to find bilingual English-Bulgarian, English-Estonian and French-Dutch texts.

2.1 Morphological Analysis

As we have seen, morphological analysis is necessary if one wishes to access an online dictionary. Since broad-coverage analysis packages represent very major development efforts, GLOSSER was fortunate in having use of *Locolex*, a state-of-the-art system from Rank Xerox (Bauer, Segond, and Zaenen, 1995).

A French example analysis (from Figure 2):

- **atteignissent** as `atteindre+SubjI+PL+P3+FinV`;

The semi-regular form is recognized as a subjunctive, third-person plural finite form of the verb *atteindre*. The information about the stem (lemma) from the morphological parse enables a dictionary lookup, and the grammatical information is directly useful. Note that, in contrast to commercially available systems, the information is generated automatically—so that it is available on-line for any text.

But there are also examples of words which could 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 looed up in the dictionary and for which corpora examples are sought.

2.2 Dictionary

GLOSSER was likewise fortunate in obtaining the use of good online dictionaries: the Van Dale dictionary *Hedendaags Frans* (van Dale, 1993) is used for French-Dutch, and the Kernermann semi-bilingual dictionaries are used for mapping English to Bulgarian, Estonian, and Hungarian. Only the Estonian version is complete. Although there are no paper versions of the latter available, (Kernermann Publishing, 1993) demonstrates the basic concept for English-Finnish.

2.3 Corpus

We have relied on other projects, the ECI and MULTEXT for bilingual corpora, although this has involved some work in (re)aligning the texts.

The results of disambiguation and morphological analysis serve not only as input to dictionary lookup but also to corpus search. The current implementation of this search uses a LEXEME-based index for rapid and varied access to the corpus.

In order to determine the size of corpus needed, we experimented with a frequency list of the 10,000 most frequent word forms. A corpus of 2 MB contained 85% of these, and a corpus of 6 MB 100%. Our goal is 100% coverage of the words (lemmata) found in the 30,000-word dictionaries, and 100% coverage of the most frequent 20,000 words. The current corpus size is 8 MB.

As the corpus grows, the time for incremental search likwise 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. This was implemented and is described in (van Slooten, 1995). The indexed lookup is most satisfactory—not only has the absolute time dropped an order of magnitude, but the time appears to be constant when corpus size is varied between 1 and 10 MB.

Lexeme-based search looks not only for further occurrences of the same string, but also for inflectional variants of the word. If the selected word is **livre+Masc+SG+Noun**, the search should find other tokens of this and also tokens of the plural form **livres**. This is made possible by lemmatizing the entire corpus in a preprocessing step, and retaining the results in an index of lemmata. It is clear that this improves the chance of finding examples of a given lexeme immensely.

2.4 User Interface

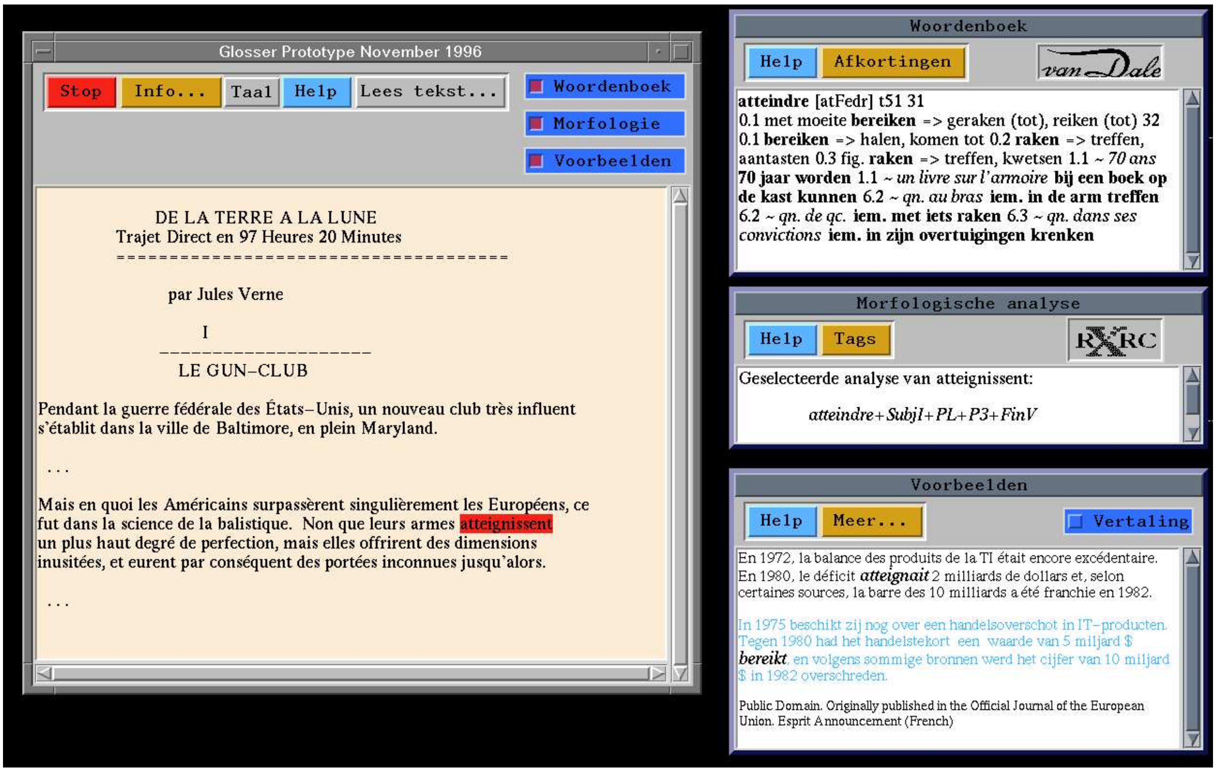
The text the user is reading is displayed in the main window. Each of the three sorts of information is displayed in separate windows: MORPHOLOGY, the results of morphological analysis; DICTIONARY, the French-Dutch dictionary entry; and EXAMPLES, the examples of the word found in corpora search. See Figure 2 for details.

3 Using Glosser

A pilot study involving 20 university-level students of French was conducted in Feb. 1996. Half of the students used GLOSSER, and the other half a paper version of the same dictionary and all read the same text and answered questions tested text comprehension and satisfaction. The time needed for the task was also measured. The results of this pilot study were encouraging: although the level of student was too high (Dutch foreign language students have a high level of proficiency), so that no differnces in comprehension were noted, the GLOSSER users were faster, and reported enjoying the experience and interested in using the system further. We have just completed a more careful replication with more students at a lower level of French proficiency, and the predictions of the pilot are borne out: there are very significant differences in speed, insignificant advantages in comprehension, and high overall satisfaction (Dokter et al., to appear 1997).

4 Conclusions

GLOSSER was developed with the philosophy of exploiting available NLP technology wherever possible. Morphological analysis (lemmatization) is robust and accurate and more than up to the task



DE LA TERRE A LA LUNE
Trajet Direct en 97 Heures 20 Minutes
=====================================

par Jules Verne

I

LE GUN-CLUB

Pendant la guerre fédérale des États-Unis, un nouveau club très influent s'établit dans la ville de Baltimore, en plein Maryland.

...

Mais en quoi les Américains surpassèrent singulièrement les Européens, ce fut dans la science de la balistique. Non que leurs armes atteignissent un plus haut degré de perfection, mais elles offrirent des dimensions inusitées, et eurent par conséquent des portées inconnues jusqu'alors.

...

atteindre [atFedr] t51 31
0.1 met moeite **bereiken** => geraken (tot), reiken (tot) 32
0.1 **bereiken** => halen, komen tot 0.2 **raken** => treffen, aantasten 0.3 fig. **raken** => treffen, kwetsen 1.1 ~ *70 ans* **70 jaar worden** 1.1 ~ *un livre sur l'armoire* **bij een boek op de kast kunnen** 6.2 ~ *qn. au bras* **iem. in de arm treffen** 6.2 ~ *qn. de qc.* **iem. met iets raken** 6.3 ~ *qn. dans ses convictions* **iem. in zijn overtuigingen krenken**

Geselecteerde analyse van atteignissent:

atteindre+SubjI+PL+P3+FinV

En 1972, la balance des produits de la TI était encore excédentaire. En 1980, le déficit ***atteignait*** 2 milliards de dollars et, selon certaines sources, la barre des 10 milliards a été franchie en 1982.

In 1975 beschikt zij nog over een handelsoverschot in IT-producten. Tegen 1980 had het handelstekort een waarde van 5 miljard $ ***bereikt***, en volgens sommige bronnen werd het cijfer van 10 miljard $ in 1982 overschreden.

Public Domain. Originally published in the Official Journal of the European Union. Esprit Announcement (French)

Figure 2: USER INTERFACE GLOSSER. On the left is a text in which in which information for the word *atteignissent* has been requested; on the right, from the top are windows for dictionary (Van Dale), morphological analysis (Rank Xerox) and examples in bilingual corpora.

of supporting instructional software. The text processing techniques employed in GLOSSER are not exotic, and likely robust enough to support quick access to corpora on the order of 10 MB in size.

5 Acknowledgements

The Copernicus program of the European commission supports the GLOSSER project in grant 343. The authors are site coordinators; the project has been conducted by them and other members, including Mariana Damova, Duco Dokter, Margit Langemets, Auke van Slooten, Petra Smit, Maria Stambolieva, Tarmo Vaino and Ülle Viks. Valuable criticism has come from Poul Andersen, Susan Armstrong and Serge Yablonsky.

References

Bauer, Daniel, Frederique Segond, and Annie Zaenen. 1995. LOCOLEX: Translation rolls off your tongue. In *Proceedings of the conference of the ACH-ALLC'95*, Santa Barbara, USA.

Dokter, Duco, John Nerbonne, Lily Schurcks-Grozeva, and Petra Smit. to appear, 1997. Glosser-R*u*G: A user study. In Arthur van Essen, Sake Jager, and John Nerbonne, editors, *Language Teaching and Language Technology*. to appear in proceedings of conference to be held 28-9 Apr 97.

Kernermann Publishing. 1993. *Password: English Dictionary for Speakers of Finnish*. Porvoo, Finnland: Kernermann.

Krashen, S. D. 1989. We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *Modern Language Journal*, 73(4):440–464.

Nerbonne, John and Petra Smit. 1996. GLOSSER-RuG—in support of reading. In *Proc. of COLING '96*, pages 830–35, Copenhagen.

Sproat, Richard. 1992. *Morphology and Computation*. Cambridge: MIT Press.

Swaffar, Janet, Katherine Arens, and Heidi Byrnes. 1991. *Reading for Meaning : an Integrated Approach to Language Learning*. Englewood Cliffs, N.Y.: Prentice Hall.

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

van Slooten, Auke. 1995. Searching and quoting examples of word-usage in french language corpus. Technical report, Alfa-Informatica, Rijksuniversiteit Groningen.