GLOSSER-RuG: in Support of Reading

John Nerbonne and Petra Smit
Vakgroep Alfa-informatica
Rijksuniversiteit Groningen
nerbonne@let.rug.nl, smit@let.rug.nl

Abstract

This paper reports on ongoing work on a CALL system to facilitate foreign language learning: GLOSSER-RuG. The system is particularly dependent on advanced morphological analysis. Following a brief introduction to the project, the paper describes the architecture of GLOSSER-RuG. Then we describe in detail the main components/modules that are part of the implemented prototype. Finally, implementation issues and details involving the user interfaces of the tool are discussed. We outline the design of an integrated system to support the reading of French text by Dutch speakers.

1 Introduction

This paper reports on our ongoing research towards a computer-assisted language learning (CALL) tool, GLOSSER-RuG. After only several months, a first prototype was operational. This demonstrates that useful language-learning and language-assistance systems are presently within reach—given the availability of key components such as morphological analysis software and online dictionaries. In the case of GLOSSER-RuG, this was morphological analysis software made available by Rank Xerox, Grenoble (Chanod and Tapanainen 1995; Daniel Bauer and Zaenen 1995) and an online French-Dutch dictionary provided by Van Dale Lexicographie (VanDale 1993). The system integrates previously existing software modules, and supplies the minimal additional ones together with interfaces in order to support the reading of French text by Dutch speakers.

Following a brief introduction to and motivation for the project, the paper describes the architecture of GLOSSER-RuG. We describe the main components/modules that are part of this prototype, including implementation and the user-interface.

1.1 Motivation

(Zaenen and Nunberg 1995) notes that, even as fully automatic machine translation has receded as a reasonable mid-term goal for natural language processing, several goals have emerged which are less ambitious, but useful and attainable. These focus less on eliminating language barriers and more on assisting people in learning and understanding the wide range of languages in current use. It is still the case that language differences form a substantial barrier to the free flow of ideas and technologies: ideas are effectively only accessible only to those in command of the language they are expressed in. But since an ever increasing number of people encounter texts electronically, automated methods of language processing may be brought to bear on this problem. GLOSSER-RuG is designed to help people who know a bit of French but cannot read it quickly or reliably. It allows a native Dutch person to learn more about French morphology, it removes the tedious task of thumbing through the dictionary and it gives examples from corpora.

GLOSSER-RuG may also be contrasted with more traditional computer-assisted language learning (CALL) software (Last 1992) which has focused primary on providing exercises, answer keys, and links to grammar explanations. GLOSSER-RuG on the other hand, focuses on providing assistance to novice readers — whether these are activeley involved in educational programs or not, and the focus is clearly on the level of *word*, including the grammatical information associated with inflectional endings. We therefore regard traditional CALL software as complementary in purpose.

2 Design

We envision a user of intermediate level in French (school level, not university level). While the user reads a text, s/he can select with a mouse an unknown or unfamiliar word. The program makes available:

- the internal structure of the word, incl. the grammatical information encoded in morphology
- the dictionary entry of the word in a bilingual French-Dutch dictionary; and
- other examples of the word from corpora

A user-interface 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 French morphology and 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 provided 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 studies indicate 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 were further concerned to use texts from a variety of genres, and we attempted (with very limited success) to find bilingual French-Dutch texts. To-date we have only the bible and the treaty of Maastricht in bilingual form.

2.1 Morphological Analysis

As we have seen, morphological analysis is necessary if one wishes to access an online dictionary. Since large coverage analysis packages represent very major development efforts, GLOSSER-RuG was fortunate in having access to *Locolex*, a state-of-the-art system provided by Rank Xerox.

Some examples of its analyses:

- **vont** as `aller+IndP+PL+P3+FinV`;
- **bien** as `bien+Masc+SG+Noun`, and `bien+Adv`; and
- **chats** as `chat+Masc+PL+Noun`.

The information from the morphological parse enables a dictionary lookup and the grammatical information is directly useful to readers. But there are also examples of words which could have different grammatical meanings.

2.2 Dictionary

GLOSSER-RuG was likewise fortunate in obtaining the use of an online version of the VanDale dictionary *Hedendaags Frans*. VanDale is the premier publisher of Dutch dictionaries.

In *Hedendaags Frans*, for example, the word **baiser** could be a noun as well as a verb and contains therefore the following information (the actual data structures are different, and confidential).

entry 1
<LEMMA> baiser
<GRAM> masculine noun
<TRANS> kus [a kiss]
...

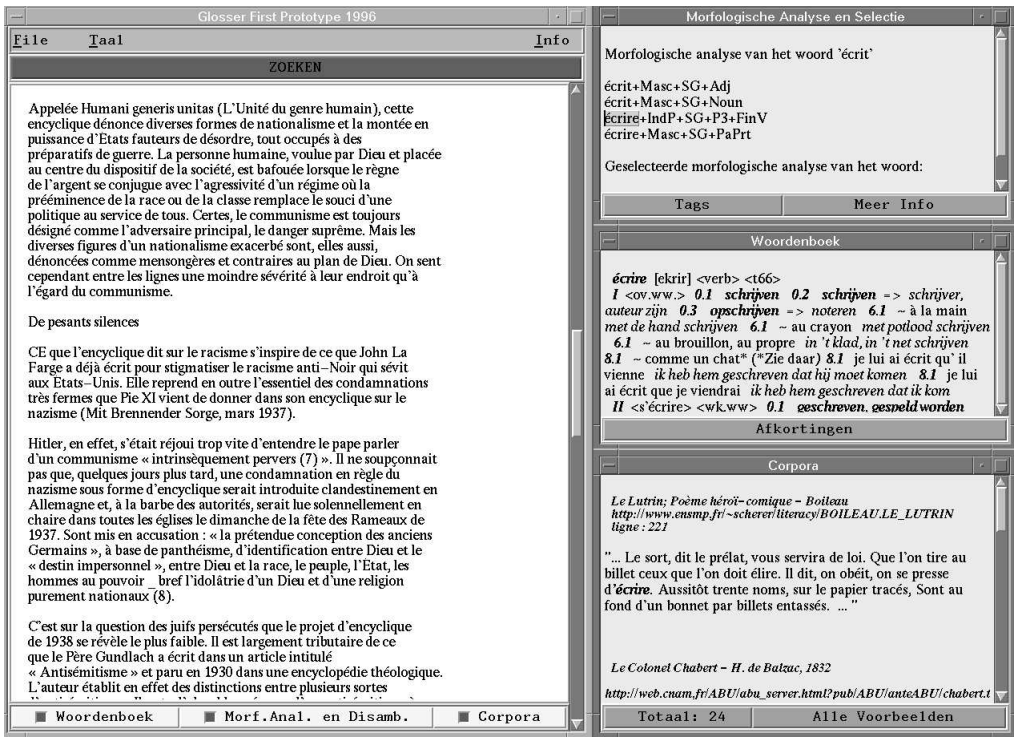

Figure 1: User Interface GLOSSER-RuG. On the left is a text, on the right, from the top are windows for morphological analysis, dictionary, and further examples.

entry 2
<LEMMA> baiser
<GRAM> transitive verb
<TRANS> kussen [to kiss]
...

Cases like these suggest a potentially crippling problem for the GLOSSER-RuG concept: if words are in general ambiguous, then providing morphological analyses for them may be too tiresome to be of genuine use to language learners. A long list of potential analyses is potentially of very little use. Since indeed most words are multiply ambiguous, a problem looms.

2.3 Disambiguation

The solution to this problem is disambiguation: to find the right entry in the dictionary, a part-of-speech (POS) disambiguator is applied before morphological analysis in order to obtain the contextually most plausible morphological analysis.

For example in the sentence *Bon, donne-moi un baiser* 'Good, give me a kiss', the disambiguator should return a tag for the word **baiser** indicating [masculine] noun and in the sentence *Il ne peut pas baiser* 'He can't kiss' the word **baiser** should be assigned with a tag indicating verb [infinitive]. The combination of POS disambiguator and morphological analysis suffice to provide the contextually most likely analysis nearly all the time. Stochastic POS disambiguation is implemented in the Rank Xerox *Locolex* package.

2.4 Corpus

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 only string matching to find further tokens. Our design calls for LEXEME-based search however, and a preliminary version of this has also been implemented.

In order to determine the size of corpus needed, we experimented with a frequency list of the 10,000 most frequent words. A corpus of 2 MB contained 85% of these, and a corpus of 6 MB 100%. Our goal is 100% coverage of the words found in *Hedendaagse Frans*, and 100% coverage of the most frequent 20,000 words, and we are close to it. 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.

2.5 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 1 for an example.

In case the disambiguator / morphological-analyser cannot decide which analysis is more likely, the user is allowed to select which he is interested in (this feature toggles for users who prefer fewer choices).

With pedagogical software there is a danger of assuming too much expertise on the part of users. In GLOSSER-RuG this danger could take the form of displaying further unknown words in either the dictionary or the examples windows. To obviate this at least partially, both of these windows have been made sensitive to GLOSSER-RuG's search. Thus, if, e.g., corpus search turns up examples with further unknown words, these may be submitted to GLOSSER-RuG for analysis, look-up and examples.[1]

2.6 Summary of Design

The prototype was designed to consist of the following modules: a a disambiguator, morphological analyser, a dictionary lookup and a corpora search as shown on the next page. Corpus lemmatization and indexation based on lemma are done off-line. In the next section we will illustrate these modules in more detail.

3 A session with GLOSSER-RuG

The present section steps through the various modules in order to illustrate the system more concretely and in order to motivate some further design decisions.

[1] This is a point at which input from traditional language pedagogy could be very useful—especially reading material that has been screened and edited to be accessible to a particular level.

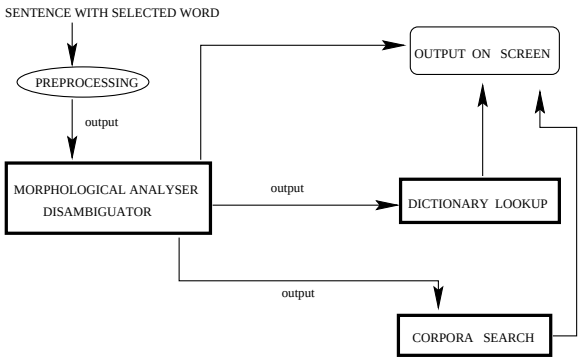

Figure 2: ARCHITECTURE GLOSSER-RuG.

3.1 An Example

When the user selects a word in a text for example **écrit** in the sentence:

...La colère était **écrit** sur son visage...

3.2 Preprocessing

The program must first extract from the text the sentence in which the word occurs. It does this on the basis of punctuation, paying special attention to the occurrence of abbreviations (*e.g., .e., P.J.*) and titles (e.g., *dr., mm.* etc.).

3.3 The morphological analyser

After this so-called preprocessing, the morphological analyser is called to get the morphological information of the selected word, i.e. the lexeme and possible tags according to result of the morphological analysis.

Figure 3: THE MORPHOLOGICAL ANALYSIS RANK XEROX *Locolex*.

As the example shows the morphological analyser gives four possible [grammatical] readings of the selected word and two base forms [lexemes]. It should be noted that the preprocessing phase isn't necessary for the morphological analyser.

3.4 Disambiguator

As mentioned in the previous section the morphological analysis information might not be enough to get the right entry in the dictionary. In this example there are many possible base forms of the selected word, namely:

entry 1
<LEMMA> écrit
<GRAM> masculine noun
<TRANS> geschrift
...
entry 2
<LEMMA> écrire
<GRAM> verb
<TRANS> schrijven
...
[abbreviated]

In order to get the right entry, in this case **entry 1**, one has to consider te whole sentence. Research on POS-tagging has proved it to be a good method to disambiguate a sentence. The disambiguator assigns every word of the sentence a tag. In this example the disambiguator chooses the `écrire+Masc+SG+PaPrt` reading as the most likely one, as shown in Figure 3.

3.5 Dictionary Lookup

After disambiguation the lexeme with the most likely tag is used to get the right entry of the selected word in the dictionary.

Figure 4: The dictionary Lookup
Van Dale *Hedendaags Frans*.

The dictionary lookup process is straightforward. The exact structure of the dictionary source files is confidential, but it is well-structured, and allows uncomplicated access. The right file is opened and searched until a match with the lexeme occurs. If this is the case the information of this lexeme is printed in pretty form on the screen. In the case the user reads a French word in the dictionary output and wants to get the dictionary entry of this particular word, s/he can select this word in the dictionary output and after a push on the search button the selected word is morphological analysed and, if possible, disambiguated and with the lexeme another dictionary lookup will taken place and the information found will be placed in another dictionary window on the screen.

3.6 Dealing with Inaccuracies

Although the disambiguator is very accurate, it doesn't always assigns the right tag to a word. Consider for example the sentence

> Je pense que tu as l'**as** de pique [I think you've got the ace of spades]

According to the morphological analyser the selected word **as** has two base forms namely *avoir*, indicating a verb [`avoir+ INDP+SG+P2+Avoir`]- and *as*, indicating a noun [`as+Masc+INVPL+NOUN`]. To choose the right base form, one consults the disambiguator, but it selects the 'verb' tag instead of the wanted 'noun' tag. In this case the dictionary lookup module will fetch the wrong entry, namely of *avoir*. In order to get the right entry, namely *as*, it is possible for the alert user of GLOSSER-RuG to override the decision of the disambiguator. The user can select the other ('wanted') tag, push the search button, and accordingly get the right dictionary entry and corpora examples on the sceen.

3.7 Corpora Search

The selected word and its lexeme form also the input for the Corpora Search module. This component uses indexed files (van Slooten 1995). The index is set up in two parts. The first part is an index to generate a key for every word. This index is used for all files in the corpus[2]. This key is then used in the second part where for every file in the corpus two extra index files are generated. These files contain information about the position of words by their key in the corpus file up to a certain maximum (e.g. 50) of occurrences. As the index consists of two parts, so does the lookup. The first part is to get all the keys of words starting with a particular string from the first index. Then these keys can be used to search in the second index, one index file for each corpus, for occurrences of the word denoted by these keys. If the Corpora Search Module has as input **écrit** [the selected word] and **écrire** [the base form] the following examples (a.o.) will be found:

[2]The corpora text are collected from different sides on the WWW.

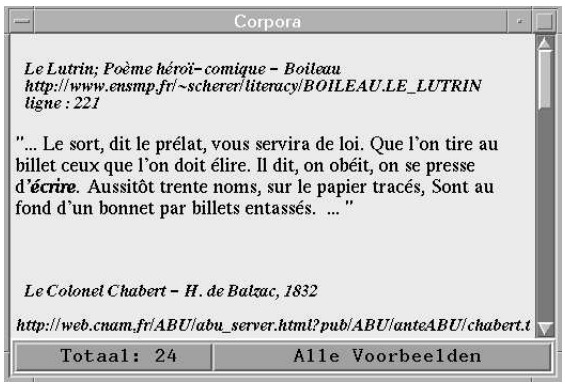

Figure 5: Some Corpora Examples.

As in the Dictionary window it is also possible to select another French word in the Corpora output and push the Search button. The morphological analysis and disambiguation of this selected word and the dictionary entry will accordingly be displayed in the relevant windows.

4 Final Remarks

The intergration of existing morphological processing tools has led to a powerful CALL tool. The tool provides a dictionary lookup, it gives examples from corpora and displayes morphological information, all on-line. Other languages could be easily implemented in the overall skeleton of GLOSSER-RuG. Although development of the prototype GLOSSER-RuG is still ongoing, these first results look very promising. The prototype was sufficiently advanced in February for Groningen communications students to conduct an investigative user study. Although we'll report on this seperately, it indicated user interest.In the near future we're planning to index the corpora on basis of lexemes. Later we wish to extend the software with for example a teaching and diagnosting module so that the tool matures to real CALL software.

5 Acknowledgements

The work is supported by grant number 343 to the University of Groningen from the EU Copernicus program. The Copernicus partners consulted on a common design, in particular Lauri Karttunen (Rank Xerox, Grenoble, France), Elena Paskaleva (Linguistic Modelling Laboratory Bulgarian Academy of Sciences, Bulgaria), Gabor Proszeky (MorphoLogic, Hungary), Tiit Roosmaa (Tartu University, Estonia), Maria Stambolieva (Institute of Bulgarian Language, Bulgarian Academy of Sciences), and Ülle Viks (Institute of the Estonian Language, Estonia). Auke van Slooten designed and programmed the corpus indexing and search routines. Lauri Karttunen (Grenoble) advized on the use of morphology and Gertjan van Noord (Groningen) on TCL/TK.

References

Jean-Pierre Chanod and Pasi Tapanainen. 1995. Creating a tagset, lexicon and guesser for a french tagger. In *Proceedings of the ACL SIGDAT workshop on "From Texts To Tags: Issues In Multilingual Language Analysis"*, pages 158–64, University College Dublin, Ireland.

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

R. Last. 1992. Computers and language learning: Past, present - and future? In C. Butler, editor, *Computers and Written Texts*, pages 227–245, Oxford: Blackwell.

Richard Sproat. 1992. *Morphology and Computation.* MIT Press.

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

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

Annie Zaenen and Geoff Nunberg. 1995. Communication technology, linguistic technology and the multilingual individual. In Toine Andernach, Mark Moll, and Anton Nijholt, editors, *CLIN V: Papers from the Fifth CLIN Meeting*, pages 1–12, Enschede. Taaluitgeverij.