GLOSSER as a Practical Application of the Semi-Bilingual Dictionary Concept

Margit Langemets

Margit Langemets heads the Department of Lexicology at the Institute of the Estonian Language, and specializes in monolingual lexicography, dictionary typology and criticism, and computer applications. She graduated in Estonian Philology from Tartu University, then taught Computational Lexicography there. She initiated the publication of the semi-bilingual dictionary in Estonia, and was among its translators. Her current projects include completing the editing and computing of the Defining Dictionary of Literary Estonian, preparing a Lexicographic Text Corpus of Estonian, and computerizing several Estonian dictionaries.

The European Union is taking an increasing interest in possible ways of cooperating with Central and Eastern Europe, and so, even before these countries become members of the EU, the real cooperation has already begun. Estonian linguistics is not an exception to this.

The enormous information flow that reaches us every day via natural language makes one feel that it would be quite easy to "lose wisdom in a mass of knowledge, lose knowledge in a mass of information, lose information in a mass of data" (a quotation from Thomas Eliot). The Fourth Framework Programme COPERNICUS 1994 financed language engineering research, reusable language resources and several pilot applications. One of the aims of this kind of joint research project should certainly be bringing together academic scholars and people from industry to prove that "everything intended to be scientific need not necessarily be slow and unsaleable" (the words of a Hungarian colleague). For the first time, Estonia had a chance to participate in five such projects. One of them was GLOSSER, on which people from Bulgaria, Estonia, the Netherlands, France and Hungary worked together for a period of two years.

The result of this language technology project is a program called GLOSSER, designed to support the processes of reading and learning to read in a foreign language. This is the prototype of a system where a computer is used as a reference, not a language teacher, and assistance is offered to advanced learners who are not afraid of "machines" and who find it exciting and useful to use the computer application beside or instead of the tedious task of thumbing through a dictionary. GLOSSER differs from an ordinary dictionary lookup program in its analysing procedure that appears right on the screen. The starting point was a question arising while reading a text: How can one find the lemma and the right sense in the dictionary while meeting several kinds of word forms in the text? GLOSSER is meant as a tool for finding an answer to this question.

System architecture connects modules for morphological analysis and disambiguation, dictionary access and corpora search with an output model. Let us have a closer look at each of them.

Morphological Analysis and Part of Speech Disambiguation

Morphological analysis is necessary if one wishes to consult an on-line dictionary. GLOSSER was fortunate in having access to the Xerox POS Disambiguator for English language which, drawing on the Morphological Analyser, picks up the correct part of speech out of all possible morphological descriptions. The theoretical base of the disambiguator is English Constraint Grammar, a theory from the late 1980s, which determines the function of the word using special rules for morphological characteristics and context. These rules are the constraints.

Disambiguation means getting rid of such morphological descriptions that do not fit the specific context of the word, while semantics is not taken into account. The disambiguator makes its decision after looking through the whole sentence. Homonymy differs in type and extent in different languages. For example, English is noted for part of speech homographs, whereas Estonian is noted for homonymy of morphological forms. (In Estonian the number of word forms is very large: on average there are 33 different forms per word. About 40% of Estonian word forms are ambiguous.)

GLOSSER is a system that examines a sentence word by word, and there is only a word-based access to the dictionary. For example, in Xerox codes the input sentence *The concert was nothing to write home about* will be analysed word by word by the disambiguator (the+AT; concert+NN; be+BEDZ; nothing+PN; to+TO; write+VB; home+NN; about+IN), although here is a multiword expression *nothing to write home about* with its dictionary definition at the end of the entry for *home* (placed after two parts of speech, several derivatives and several multiword expressions containing the headword). And how should a user know that this expression is located under *home*, but not under *nothing*? It should be the next stage of the research project to deal especially with multiword expresssions and to enable an expression-based access to the dictionary. First, the system should check the possible belonging of a word in an expression, and if the answer is yes, then, secondly, display only this part of the dictionary entry (not the whole one).[1]

The Dictionary

Reusability of lexicographic resources is a widespread trend in computational lexicography today. The only feasible option is to use an existing dictionary. For GLOSSER, the most suitable candidate was Password

mainly for its belonging to the semi-bilingual type of dictionary, but also for its appropriate size (25,000 headwords)[2]. The source language is represented by a headword, grammatical information, sense explanation and illustrative sentences. The target language is represented by brief translations for each meaning of the headword (a total of 37,000) or sub-headword (derivatives, multilingual expressions). The semi-bilingual dictionary is new and unique in Estonia, and GLOSSER was very happy to find this combination of monolingual and bilingual dictionaries in one volume.

We obtained the electronic version of the dictionary text in layout format, the so-called typographic view, which is concerned with the two-dimensional printed page. These layout codes had to be stripped and converted into a suitable format. Our task was to analyse the typographic view (the raw text format) fully, to be able to transform the text into the lexical view, i.e. lexical data as those might appear in a database, without concern for their exact textual form.

The list of headwords was sent to Xerox for testing. Prof. L. Karttunen made a network from the list and checked them against the English transducer Xerox supplied for GLOSSER. About 400 headwords were not recognized by the analyzer and needed to be added to the system. For example a) British spelling for words that are in the morphological analyzer only with American spelling (*apologise, ardour*, etc.); b) French words in their original French orthography (*café, cortège,* etc.); and c) words that are not found in the American Heritage Dictionary (*casuarina, dhoti, kampung, rambutan*, etc.). The latter words originate from several other local editions. The Estonian version was supplemented with 'kroon', 'sprotid', etc.

Usually, the microstructure of a dictionary is hierarchic and, depending on type, rather complex. The conservative form of a traditional printed dictionary, because of its implicit information, is satisfactory for users, but not for various computer systems, which require information types to be set out explicitly.

For encoding the text of the printed dictionary TEI Guidelines were consulted. The encoding format has to adhere to the rigorous principles of traditional dictionaries and present them in such a way as to facilitate dictionary reusability and automatic processing. The Guidelines use the Standard Generalised Markup Language (SGML) to define their encoding scheme. It provides for a formal definition in terms of elements and attributes, and rules governing their appearance in a text[3]. A dictionary is seen as a linear text stream interspersed with markup. The tags provide an indication of the content of the fields they delimit. Each of the information fields has an opening marker <..> and an end marker </..>. One field can embed another.

- **Entry**
  <entry> contains a reasonably well-structured dictionary entry. <hom> and <sub> mark the sub-division of entries into part-of-speech homographs and sub-headwords (derived words, multiword expressions, idioms), respectively. The tags serve to group information relating to each component. The attribute (type=xref) marks a cross-reference. Entries comprise several constituent parts (form, sense, usage, etc), each providing a different type of information about the word treated. Exceptional cases are still characteristic of lexical data. Information of the same kind can appear at different levels in the same entry.

- **Form**
  <form> is the first item in an entry. <en> gives the print-form of the headword. Orthography and stress of a single element lexeme are not separated. <pr> contains the pronunciation(s) of the word. Other information, such as variant or alternative, abbreviated and full forms, inflected forms illustrating the inflectional pattern of the headword, orthographic form(s) for displaying gender contrasts, negative word or use, and comparison, is presented by additional attributes (type=var / abbr / full / infl / gen / neg / comp). Collocations (*T-shirt* under *T*) of the headword to show multiple-word lexical items are also marked.

- **Grammatical Information**
  For encoding, the tag <gr> is used to group all grammatical information about a lexical item. Usually it consists of <pos> for indicating the part(s) of speech. But in addition there might be <colloc type=prep> for prepositions and <nr> for the grammatical number associated with a form. The 'plural' specification may apply either to a) the inflected forms provided (*passers-by*), or b) the headword itself (*noun plural*).

- **Sense Information**
  <sense> stands for for semantic description. It groups information (forms, grammatical information, usage, translation(s), etc) about the given sense of a word. Attribute (n='..') indicates the sense number. <df> contains the text of the definition. Definitions describe the meaning of some lexical item, most often of the headword of the entry, while in some cases they describe examples. <ee> contains the Estonian translation text. Definitions and translations are usually accompanied by examples. <ex> contains an example text with at least one occurrence of the word form, used in the sense being described, in it. Examples may still contain other elements, eg, <pr>, <usg>, <pos>, etc.

- **Usage Information**
  Most dictionaries provide restrictive labels and phrases indicating the usage (<usg>) of given words or particular senses. Attributes help to define usage more precisely. A distinction is made between a definition and an additional descriptive phrase (type=hint; *eg of horses*; *in football, hockey, etc*). Geographic area, national or regional use (type=geo) is marked in some cases (*whisky* .. Irish and American *whiskey*). Not much is told about regional style (type=reg), but there are style labels, such as 'formal', 'informal', 'offensive', 'rare', 'unkind', etc

- **Cross References**
  These refer the reader to additional information elsewhere in the dictionary. They may be free-standing within an entry. The metalanguage remains untagged (labels 'same as', 'see', 'see below', etc). <xr> groups the information relating to a cross-reference (a phrase or sentence). <ptr target='..'> defines a pointer to another location. Cross reference-like cases occurring occasionally in definition texts remain untagged.

- **Notes**
  Notes about usage, grammar, etc, may be placed within an entry. <note> contains a note or annotation. Notes may give extra information about form (*as part of a word; with capital*) or grammar *(in questions, negative sentences, etc; placed after a noun*).

- **Related Entries**
  These are included in many dictionaries for direct derivatives or inflected forms of the headword, or for compounds, phrases, collocations, and idioms . <re> contains a degenerate entry embedded inside a larger entry. It is often of reduced form, consisting mainly of nouns without any sense information.

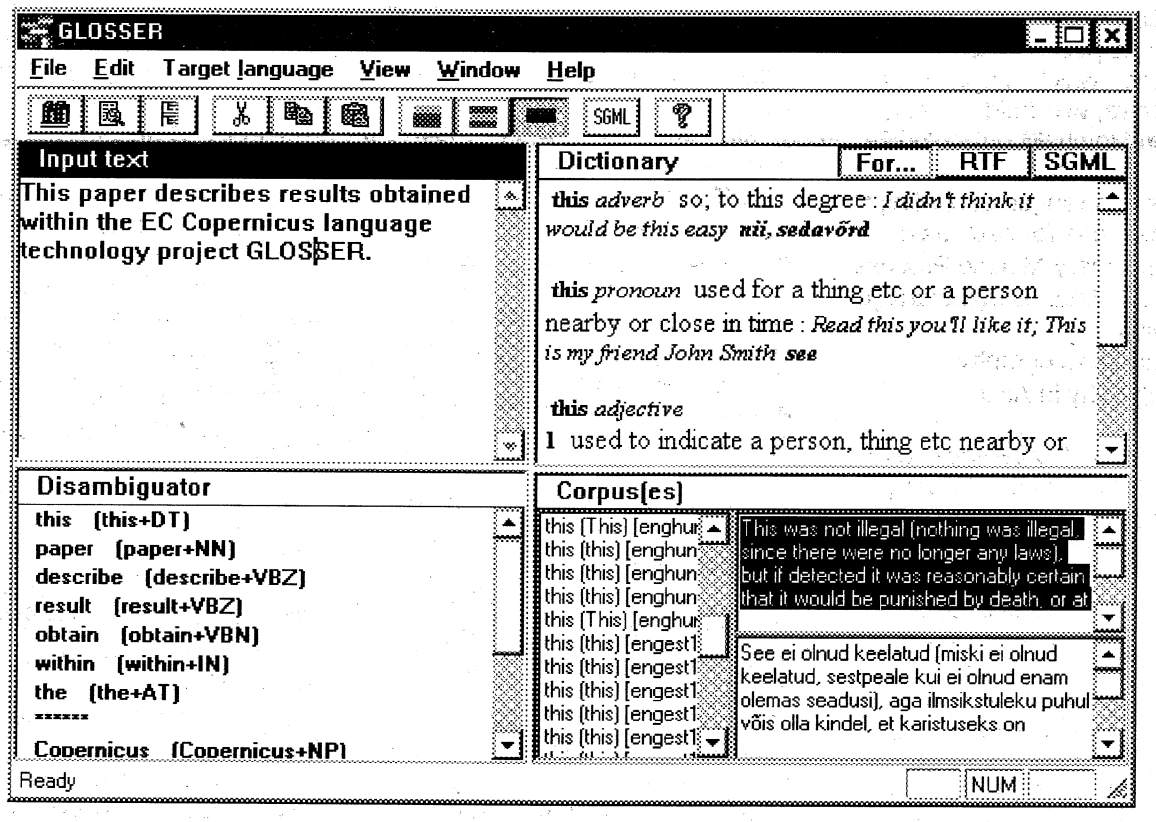
Text Corpus

The results of morphological analysis also serve as input to corpus search. Lexeme-based search looks for further occurrences of the same string. Up to now GLOSSER has relied on another EC project for bilingual corpora, which involved work in (re)aligning the texts. The corpus should be big enough to cover the 10,000 most frequent words, i.e. ca 5MB.

User Interface

The system is supported by Unix and Windows environments. The main window consists of four child windows: the Input Window, the Disambiguator Window, the Dictionary Window and the Corpus Window. Input may be typed in or read in from elsewhere in the computer memory. After marking the text in the input window, one can either analyse it or look it up in the dictionary, or get different examples of aligned sentences with their translations of the word.

Figure Main Window:



♦ **Notes**

1 In the dictionary project COMPASS (finished in 1996) all multiword expressions were coded.

2 PASSWORD Inglise-eesti seletav sõnaraamat English Dictionary for Speakers of Estonian, 1995. Kernerman Semi-Bilingual Dictionaries. Tallinn, TEA Language Center Ltd. 855 pp.

3 C. M. Sperberg-McQueen, L. Burnard (eds), Guidelines for Electronic Text Encoding and Interchange (TEI.P3). Chicago, Oxford, 1994

♦ **References**

Nerbonne J. and P. Smit, GLOSSER-RuG: in Support of Reading. Working paper of Vakgroep Alfa-informatica, Rjiksuniversiteit Groningen.

Roosmaa T. and G. Proszecky, GLOSSER - Using Language Technology Tools for Reading Texts in a Foreign Language. Final report of GLOSSER.

Viks Ü. Arvutuslingvistika hea aasta. Keel ja Kirjandus, 1997/1, pp. 54-58.

Viks Ü. 1984 Sõnavormide homonüümia eesti keeles. Keel ja irjandus, 1984/2, pp. 97-104.