

# Glossa: A multilingual, multimodal, configurable user interface

Lars Nygaard, Joel Priestley, Anders Nøklestad, Janne Bondi Johannessen

The Text Laboratory  
ILN, University of Oslo

E-mail: larsnyga@gmail.com, joeljp@gmail.com, anders.noklestad@iln.uio.no, j.b.johannessen@iln.uio.no

## Abstract

We describe a web-based corpus query system, Glossa, which combines the expressiveness of regular query languages with the user-friendliness of a graphical interface. Since corpus users are usually linguists with little interest in technical matters, we have developed a system where the user need not have any prior knowledge of the search system. Furthermore, no previous knowledge of abbreviations for metavariables such as part of speech and source text is needed. All searches are done using checkboxes, pull-down menus, or writing simple letters to make words or other strings. Querying for more than one word is simply done by adding an additional query box, and for parts of words by choosing a feature such as "start of word". The Glossa system also allows a wide range of viewing and post-processing options. Collocations can be viewed and counted in a number of ways, and be viewed as different kinds of graphical charts. Further annotation and deletion of single results for further processing is also easy. The Glossa system is already in use for a number of corpora. Corpus administrators can easily adapt the system to a wide range of corpora, including multilingual corpora and corpora with audio and video content.

## 1. Introduction

This paper describes the corpus search interface system Glossa, a system that combines high user-friendliness with advanced search, viewing and post-processing facilities. With Glossa, a corpus administrator can make a wide range of corpora available for querying with little effort. The system is available as described in section 2.

The reason for developing Glossa was the realization that if the full potential of corpus-based linguistics studies is to be achieved, corpora have to be made available for all students and researchers in language studies, not just computer experts, specialized computer linguists and corpus linguists. It is essential to create high-quality corpus interfaces. There are several corpus interfaces available, see e.g. Johannessen et al. (2000), Bick (2004), Hoffmann and Evert (2006). However, they have limitations:

- some are not network-enabled (i.e. each user has to download and manage corpora)
- some lack flexibility with regard to queries, results display and post-processing
- many are tied to a specific corpus
- few are completely GUI-driven

Typically, applications require queries to be formed as regular expressions in some formal language, such as the one defined by CWB (Corpus Workbench), see Christ (1994), Evert (2005). However, many corpus users find it difficult to learn such query languages, with their requirements for accurate use of parentheses, asterisks, percentage signs etc. Furthermore, applications often require the users to know the full tag set before querying the corpus. That is, knowledge of both tag inventory and tag names and abbreviations is necessary, as well as abbreviations of source texts, etc. Again, corpus users are often unwilling to invest the time necessary to learn

such basics, before using a corpus. For many potential users, these issues act as a threshold, preventing them from making easy or efficient use of corpora tools.

We believe that an easy-to-use, flexible graphic user interface is important for maximizing the potential of corpora in research, development and teaching. Furthermore, the interface should not presuppose full-text access to the corpora, as licence conditions may prohibit free redistribution, even if they often do allow web-based querying. Glossa satisfies these criteria.

Finally, corpus applications do not usually have advanced results management as part of their standard user-interface. Glossa comes with a comprehensive array of result processing options.

## 2. The Glossa search interface system

The Glossa search interface system uses the IMS Corpus Workbench (Christ 1994) and the relational database MySQL as a back-ends, and has a web interface that allows users to create complex queries in very simple ways, and browse, process and download result sets. Glossa supports all types of corpora, both multilingual and multimodal, with various amounts of annotation. For parsed corpora (treebanks), the similar SearchTree (Nygaard and Johannessen 2004) system can be included. All programming code is available under an open source licence.

The Glossa system is already being used with several corpora, including the following written language corpora: the Oslo Multilingual Corpus, a parallel corpus for Norwegian, English, German, French, Dutch and Portuguese (Johansson and Hofland, 1994), corpora for North Sámi and Lule Sámi, and Danish, French and Macedonian corpora, as well as several other monolingual Norwegian corpora that are either in place, or in the process of being included under the Glossa umbrella. The system is also being used with spoken language corpora, with multimodal, transcribed audio

and video options: the UPUS corpus of multicultural Norwegian, the NoTa corpus of modern Oslo dialect, and the Taus corpus of Oslo dialects from 1970s (for the latter two, see Johannessen et al. 2006, 2007 and Johannessen and Hagen, 2007).

These corpora can be accessed via the Text Laboratory's website (see references). However, it should be noted that while access to most of them will be granted on request, several of them require special permission from their owners.

## 2.1 Queries

The corpus user can specify any given token by attributes, such as:

- word
- lemma
- affix or part of word
- part of speech
- morphological features
- syntactic functions
- sentence position

Notice that these search attributes are independent of each other. Importantly, it is possible to search for part of speech without specifying a search string. All searches are done using checkboxes, pull-down menus, or writing simple letters to make words or other strings. Querying for more than one word is simply done by adding an additional query box. No regular expressions are needed to perform queries.

As an example, consider how a query for a word beginning with the sequence "jump", specified to be a noun and plural, would look like with a regular expression:

`(word="jump.*"%c&(number="pl")&(pos="n"))`

Figure 1. A regular expression in the language defined by CWB, for a plural noun starting with the letter sequence *jump*.

With the Glossa system, the user instead writes the word *jump* into a box, and chooses any appropriate additional constraints by pulling down menus, removing the need to memorize complex tag-sets. Attributes can be negated and arbitrarily combined. Two or more tokens can be combined to form a phrasal search, where the number of unspecified positions between the tokens can be specified.

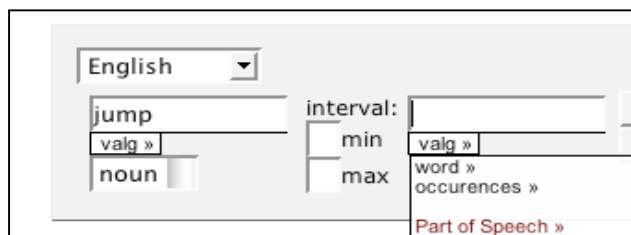


Figure 2. Query for a specific word (*jump*) limited to the noun occurrences of this word.

In multilingual corpora, one or more search phrases can be created for the aligned corpus. The graphical interface makes it easy to create queries. See figure 3.

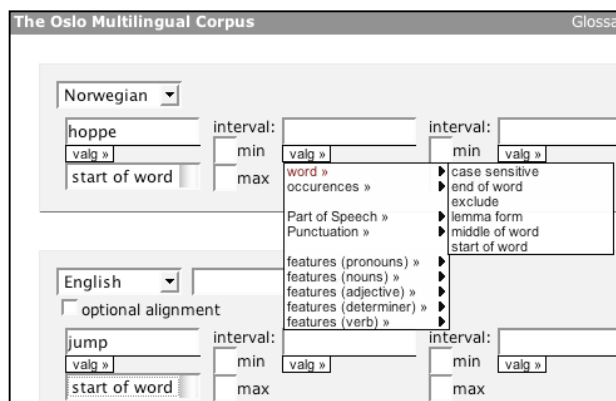


Figure 3: A simple query for a bilingual search.

## 2.2 Results

The results of queries in the Glossa system are presented as traditional KWIC lists. If available, the results can also be presented as audio/video, aligned text and annotation of tokens.

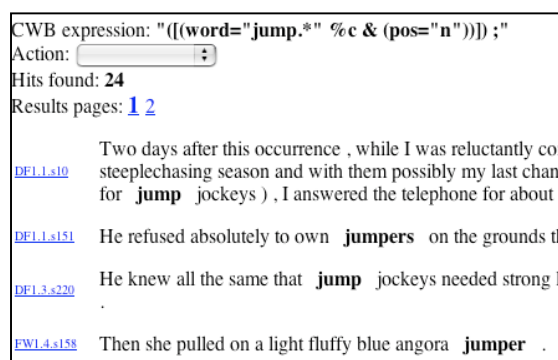


Figure 4: Part of the results window. Notice the action option on top and the clickable bibliographical information link on the left of each result.

The attributes associated with the hit, but not displayed in the results, can be viewed by passing the mouse over the words, as seen in figure 5.

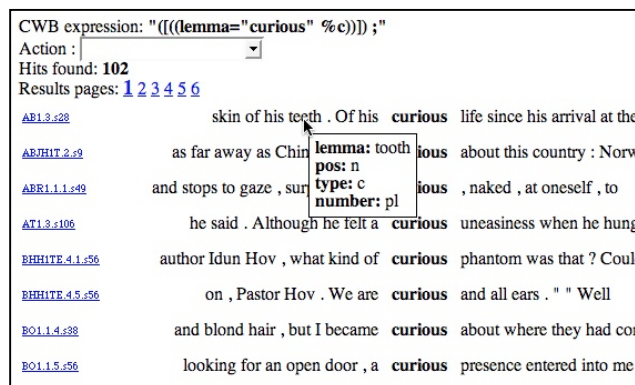


Figure 5: Results window, demonstrating the mouse roll-over effect.

A transparent, modular set-up has been created to allow integration of further post-processing and browsing options (collocation analysis etc.). The results can be post-processed in a number of ways, including:

- sorting on matching phrases, bibliographic information or arbitrary points in the context
- counting matched phrases
- downloading result set in various formats, e.g. tab separated values and Excel spreadsheets
- collocation analysis
- co-occurrence analysis
- user-defined annotation
- singling out individual hits or whole results file for saving or deletion
- viewing with regards to metadata distribution

Figure 6 shows a collocation table selected in the action menu on the results page. The collocations option comes with several sub-options, as provided by the Ngram Statistics Package (Pedersen 2008). The user can choose between bigrams and trigrams, and between measures such as Dice Coefficient and Log Likelihood-ration.

Left context			Right context		
ngram	rank	AM occ	ngram	rank	AM occ
the **	1	0.4516 7	** .	2	0.4000 6
my **	3	0.2222 3	** place	4	0.1538 2
angora **	5	0.0800 1	** and	4	0.1538 2
a **	5	0.0800 1	** to	4	0.1538 2
for **	5	0.0800 1	** jockeys	4	0.1538 2
, **	5	0.0800 1	** off	4	0.1538 2
daughter **	5	0.0800 1	** from	5	0.0800 1
bloodstained **	5	0.0800 1	** in	5	0.0800 1
little **	5	0.0800 1	** which	5	0.0800 1
fawn **	5	0.0800 1	** around	5	0.0800 1

Figure 6: Collocations of nouns beginning with *jump-*. Collocation is one of the choices in the action menu on top of the results page (figure 3).

Bilingual query results are shown in figure 7. They are presented in the same way and have the same options as monolingual results.

CWB expression: "(((word="hoppe.\*" %c))) :OMC3\_EN (((word="

Action: [v]

Hits found: 86

Results pages: [1](#) [2](#) [3](#) [4](#) [5](#)

[AHITN.2.2.4353](#) Leende **hoppet** de uti og svømte i land med jolla på slep .

[AH1.2.3.235](#) Laughing , they jumped overboard and swam ashore , pulling

[AHIT.2.2.4353](#) Leende **hoppet** de uti og svømte i land med jolla på slep .

[AH1.2.3.235](#) Laughing , they jumped overboard and swam ashore , pulling

[ATITN.3.139](#) Hadde den visst det , ville den aldri ha **hoppet** inn i bilen .

[AT1.3.139](#) If he ' d known that , he never would have jumped into the car

Figure 7: Results from a bilingual query.

Figure 8 shows how the results page is enriched with a checkbox for each result, after the user has chosen the delete option in the action menu. This enables the user to remove individual hits.

Delete selection

[ABJTH.3.451](#) De ble oppfordret til å **gå** tur , for tante hadde ikke kjent kuld

[AB1.3.450](#) Walks were encouraged , for Aunt had not felt the cold , and th

[ABIT.3.451](#) De ble oppfordret til å **gå** tur , for tante hadde ikke kjent kuld

[AB1.3.450](#) Walks were encouraged , for Aunt had not felt the cold , and th

[ABJHT.3.3.48](#) Den blir ikke feiret , som i enkelte andre land , med militærpara

[ABJHT.3.3.49](#) It is n't celebrated , as independence is in other lands , with mili

[ABJHT.3.3.49](#) Det skjer i Oslo , hvor barnetoget alltid **går** forbi Slottet hvor

[ABJHT.3.3.49](#) In Oslo the children 's parade always walks in front of the palad

[ABJHT.3.9.47](#) De **går** tur med barna .

[ABJHT.3.3.4](#) Norwegian men bend over backwards to be fair : they will stay or by-play games so very apparent in other countries ; children

[ABRITN.1.1.190](#) Den lange spaserturen ned bakken søndag morgen ; Bestemor

Figure 8. Results enriched with a box for ticking off hits that should be deleted.

In addition to the option of deleting unwanted results, Glossa also offers the option of annotating results. These results can likewise be saved and further processed at a later stage. This can be seen in figure 9.

Save annotations

ev [v]

002  **002** (minv?) du da # rasisme (uforståelig) ... egentlig jeg har ikke følt så mye ras

ev [v]

003  **003** de er vennen din de gir fa- egentlig de sier " jævla svarting kor

ev [v]

002  **002** forhold til \* det er veldig komplisert egentlig jeg veit egentlig ikke v- m j

ev [v]

003  **003** hvor aktiv du er i samtalen det egentlig n \* mm \* mm mm (trekke

ev [v]

004  **004** var ikke noe # mer enn det egentlig jeg kunne jo selvfølgelig øn

Figure 9: Annotated results.

Finally, results can be counted in a number of ways. Our final examples illustrate results from the query "all occurrences of the Norwegian lemma bil 'car'" in the corpus. The action Count was then chosen for the resulting hits, giving as a result the number of occurrences of each inflected form of the word. (For nouns there are four inflections: singular indefinite, singular definite, plural indefinite and plural definite – *bil, bilen, biler, bilene*.) The results are shown as two charts in figure 10 and 11.

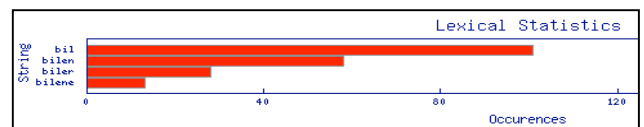


Figure 10: Count results shown as histogram.

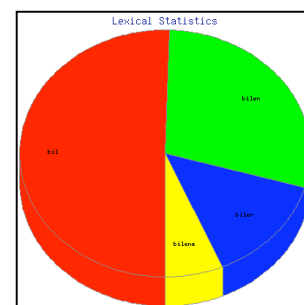


Figure 11: Count results shown as pie chart.





