

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

* Kaldera språkteknologi AS, Oslo: <http://kaldera.no>

- an easy-to-use, flexible graphic user interface
- does not presuppose full-text access to the corpora
- has advanced results management as part of its standard user interface

- - multilingual corpora
 - multimodal, with various amounts of annotation.
- The Glossa system is already being used with several corpora.
 - For written language:
the Oslo Multilingual Corpus (a parallel corpus for Norwegian, English, German, French, Dutch and Portuguese), corpora for North Sámi and Lule Sámi, and Danish, French and Macedonian corpora, several other monolingual Norwegian corpora.
 - For 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 older Oslo dialects.

























CWB expression: "([(**word**="jump" %c & (**number**="pl"))]) ;"

The screenshot shows the Morphological Analyzer interface. At the top, the word "English" is selected in a dropdown menu. Below it, there is a checkbox labeled "optional alignment" which is currently unchecked. The main input field contains the word "jump". To the right of the input field are two buttons: a plus sign (+) and a minus sign (-). Below the input field, a list of morphological features is displayed: "valg »", "word »", "occurrences »", "Part of Speech »", "Morphology (verb) »", "types (punctuation) »", "morphology (noun) »", "morphology (misc) »", "features (adjective/adverb) »", and "features (opponent) »". To the right of this list, a table shows the corresponding morphological information for each feature: "common noun", "rural", "proper noun", and "singular". At the bottom right, there are two buttons: "velg" and "utlekk".

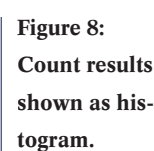
In multilingual corpora, one or more search phrases can be created for the aligned corpus. The graphical interface makes it easy to create queries.

Figure 6 shows a collocation table selected in the action menu. The collocations option in the action menu comes with several sub-options, as provided by the Ngram Statistics Package (Pedersen 2008).

Bilingual query results are presented in the same way as monolingual results. Fig. 7 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. The results can be saved for further annotation, processing and refinement.

Finally, results can be counted in a number of ways. In fig. 8 are results from the query “all occurrences of the Norwegian lemma *bil* ‘car’ ” in the corpus. (For nouns there are four inflections: sg indef, sg def, pl indef, pl def; *bil*, *bilen*, *biler*, *bilene*)



- A front-end PHP file.
- JavaScript files, both general and specific to the individual corpora. The JavaScript is both generated and hard-coded, and is used for building the interface menu structures.
- Configuration files, declaring corpus structure, attributes, tags, metadata, etc.
- MySQL database tables containing corpus metadata, associated with the individual texts within the corpora.
- The IMS Corpus Workbench and corpus files.
- A collection of back-end CGI scripts, where most of the work is done.
- Various Perl modules used by the CGI scripts.

All programming code is available under an open source licence. All components are freely available. The IMS Corpus Workbench is available for both Solaris and Linux platforms.

Glossa supports the creation of bibliographic databases of arbitrary complexity. This enables queries to be limited according to criteria such as • linguistic register • geography • time • author • age • sex • social background.

Glossa has been used extensively with speech corpora of transcribed audio and video material. The time codes generated during transcription allow accurate video streaming of the segments matching query hits. We have used the QuickTime format for streaming video content.



Figure 9: Audio/video panel, used with speech corpus.

The Glossa Manual. The Text Laboratory.
www.hf.uio.no/tekstlab/English/glossa.html




Figure 1. With Glossa querying corpora is fun.

The corpus user can specify any given token by attributes. Importantly, all attributes are independently searchable.

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

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

The results of queries in the Glossa system are presented as traditional KWIC lists, and, if available, with audio/video, aligned text and annotation of tokens. The attributes associated with the hit can be viewed by passing the mouse over the words (fig. 5). Notice the action option on top and the clickable bibliographical information link on the left.

Action: 

Hits found: **200**(max)

Results pages: **[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#)**

[ABI:1_s21](#) Now in his sixties, **he felt** himself to be unchanged

[ABITN:1_s21](#) Selv om han **lemma: sixty** ene , følte han seg ufor

pos: n

[ABI:1_s69](#) Sent to London **type: c** boy to live with h

which he did **number: pl** s an apartment house

[ABITN:1_s20](#) Han var blitt sendt til London som en skremt guttur

[ABITN:1_s71](#) mustungsgården i Compayne Gardens , som han ikke

[ABI:1_s192](#) And their **boy** doing so well : a miracle .

[ABITN:1_s198](#) Og gutten deres som gjorde det så bra : et mirakel .

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



UNIVERSITY OF OSLO
FACULTY OF HUMANITIES