A modernised version of the Glossa corpus search system

Anders Nøklestad[1]
anders.noklestad@iln.uio.no
Kristin Hagen[1]
kristin.hagen@iln.uio.no
Janne Bondi Johannessen[1,2]
jannebj@iln.uio.no
Michał Kosek[1]
michal.kosek@iln.uio.no
Joel Priestley[1]
joel.priestley@iln.uio.no

[1]The Text Laboratory, ILN, University of Oslo, P.O. Box 1102 Blindern, N-0317 Oslo, Norway
[2]MultiLing, University of Oslo, P.O. Box 1102 Blindern, N-0317 Oslo, Norway

Abstract

This paper presents and describes a modernised version of Glossa, a corpus search and results visualisation system with a user-friendly interface. The system is open source and can be easily installed on servers or even laptops for use with suitably prepared corpora. It handles parallel corpora as well as monolingual written and spoken corpora. For spoken corpora, the search results can be linked to audio/video, and spectrographic analysis and visualised geographical distributions can be provided. We will demonstrate the range of search options and result visualisations that Glossa provides.

1 Introduction

The paper presents and describes Glossa, a corpus search and results visualisation system. Glossa is a web application that allows a user to search multilingual (parallel) corpora as well as monolingual written and spoken corpora. It provides the user with advanced search options, but at the same time great care has been taken to make the interface as user-friendly as possible. The system supports login via eduGAIN as well as local accounts. Figure 1 shows the search interface of the Lexicographic Corpus of Norwegian Bokmål.

Glossa is open source and can be freely downloaded from GitHub. It can easily be installed on servers or even laptops for use with corpora that have been suitably prepared.

Figure 1: *Leksikografisk bokmålskorpus* (Lexicographic Corpus of Norwegian Bokmål). The metadata filtering options are located to the left, the linguistic search box is in the middle.

The version of Glossa presented in this paper is a modernised, reimplemented and improved version of the search system described in Johannessen et al. (2008). Glossa is part of the CLARINO infrastructure at the Text Laboratory, University of Oslo, and is financed by the CLARINO project. In the following months, all corpora in the Text Lab portfolio will be searchable in this new version. Already, important corpora like NoWaC, CANS, NORINT, ELENOR, and the Nordic Dialect Corpus are included.

Several alternative corpus interfaces are available, see e.g. Bick (2004), Hoffmann and Evert (2006), Meurer (2012) and Borin et. al (2012). What makes Glossa special compared to

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 251–254,
Gothenburg, Sweden, 23-24 May 2017.

other systems is its combination of user friendliness, especially with respect to approachability for non-technical users, ease of installation, and a unique set of search result visualisations that includes audio and video clips, spectrographic analysis and geographical map views.

2 Technical details

The server code in Glossa is written in Clojure, a modern dialect of Lisp that runs on the Java Virtual Machine. Likewise, the client/browser code is written in ClojureScript, a variant of Clojure that is compiled to JavaScript in order to run in the browser. Metadata pertaining to texts or speakers in a corpus is recorded in a MySQL database.

Running on the JVM enables Clojure to take advantage of the huge number of libraries available in the Java ecosystem. At the same time, Clojure syntax is extremely concise compared to that of Java, and its functional rather than imperative nature typically helps reduce bugs, especially when doing parallel processing like we do in Glossa.

Glossa is agnostic with respect to search engines, and different search engines may be used for different corpora within the same Glossa installation. Out of the box, Glossa comes with built-in support for the IMS Open Corpus Workbench (CWB, with the CQP search engine) as well as the Federated Content Search mechanism defined by the CLARIN infrastructure.

Glossa is able to take advantage of multiple CPU cores by automatically splitting a corpus into a number of parts corresponding to the number of cores on the machine. This leads to a significant reduction in search time, especially for heavy searches in large corpora. Search speeds will keep increasing as the number of cores grows, since Glossa automatically utilizes all cores.

For instance, when searching for a first person pronoun followed by a past tense verb in the 700 million words NoWaC corpus, a search directly in CWB (on a single core) returns 1,190,403 occurrences in 38 seconds (measured on the second run of the same query in order to allow CWB to take advantage of any result caching). The same search in Glossa, running on the same machine but taking advantage of all of its 8 cores, returns the same results in 12 seconds, i.e., about one third of the time. Furthermore, the first 8775 results are displayed within a couple of seconds, making the perceived search speed very high.

3 Querying with Glossa

A corpus user can query the corpus for linguistic features or non-linguistic features, or a combination. Glossa offers three different search interfaces, ranging from a simple Google-like search box for simple word or phrase queries, via an extended view that allows complex, grammatical queries, to a CQP query view that allows the user to specify the CQP query expression directly, potentially taking advantage of all the sophisticated options that the CQP search engine provides, see figure 1.

The most common linguistic queries involve specifying a token by given attributes: word, lemma, start or end of word, part of speech, morphological features, and sentence position. These queries can always be done in a user-friendly way.

In (1) we exemplify what a search using a search language of regular expressions would be like, in order to search for a plural noun starting with the letter sequence *dag*. In figure 2 we see the same query in *Extended Search* in Glossa. *Noun plural* is chosen from the box in figure 3. (Example 1 is translated by Glossa into regular expressions.)

(1) [word="dag.*" %c & ((pos="noun" & num="pl"))]

All searches are done using checkboxes, pull-down menus, or writing simple letters to make words or other strings. Lists of metadata categories are conveniently located to the left of the search results, allowing the results to be gradually filtered through successive selections of metadata values (see figure 4a).

Figure 2: Extended Search in Glossa.

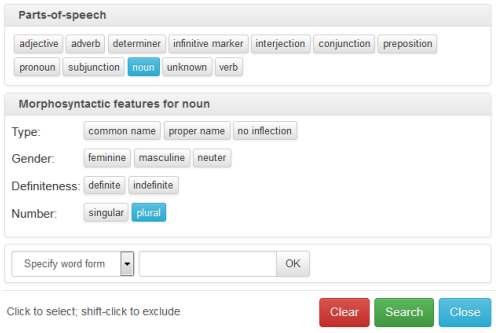

Figure 3: Choosing parts of speech in Glossa.

4 Result visualisations in Glossa

The default result visualisation is in the form of a concordance as in figure 5, but results can also be shown as frequencies as in figure 4b or as locations on a geographical map as in figure 6. The results can also be downloaded as Excel or CSV files.

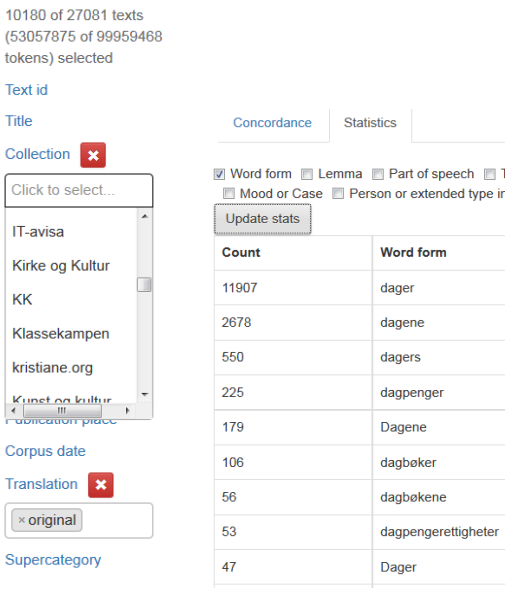

Figure 4: a) Filtering metadata in the left menu. b) Results shown as frequency list.

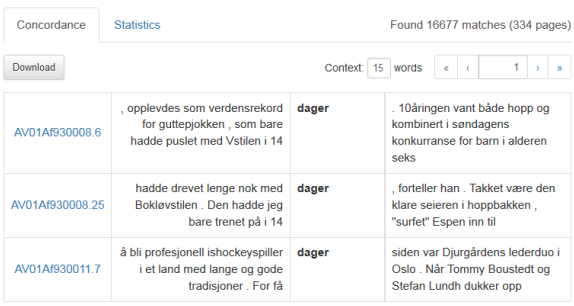

Figure 5: Results shown as concordance.

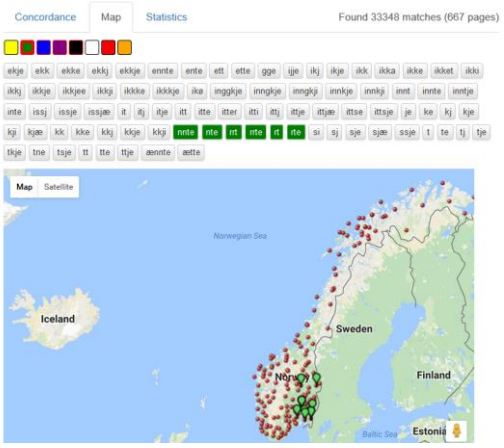

Figure 6: Geographical distribution of phonetic forms.

With spoken corpora, search results can be linked to audio and video files (figure 7), and spectrographic analysis of the sound can be displayed (figure 8).

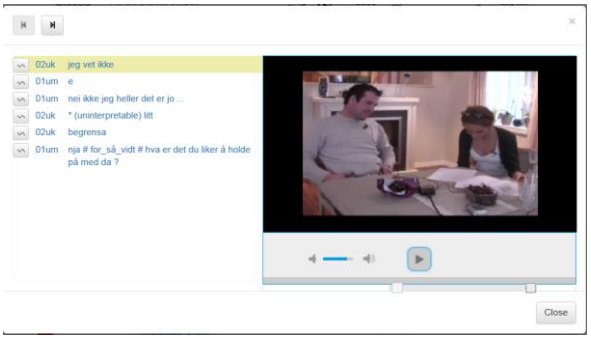

Figure 7: Video of search result with transcription.

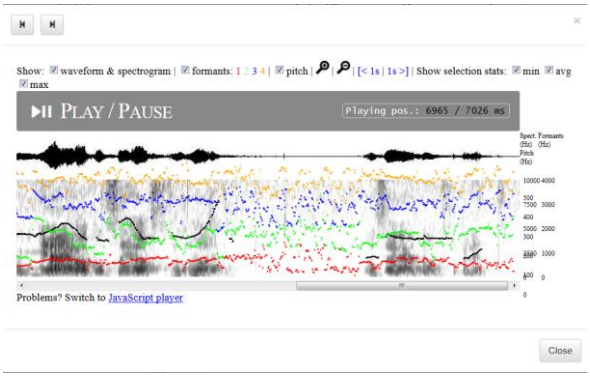

Figure 8: Spectrogram view of search result.

5 Future work

In the future we plan to implement search in syntactic annotations. We also plan to include more result views such as collocations, syntactic structures, and topic models.

References

Eckhard Bick. 2004. Corpuseye: Et Brugervenligt Webinterface for Grammatisk Opmærkede Korpora. Peter Widell and Mette Kunøe (eds). *Møde om Udforskningen af Dansk Sprog, Proceedings.* Denmark: Århus University. 46-57.

Lars Borin, Markus Forsberg and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*. Istanbul: ELRA, pages 474–478.

Sebastian Hoffmann and Evert, Stefan. 2006. Bncweb (cqp-edition): The Marriage of two Corpus Tools. S. Braun, K. Kohn, and J. Mukherjee (eds). *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, volume 3 of English Corpus Linguistics*. Frankfurt am Main: Peter Lang. 177 - 195.

Janne Bondi Johannessen, Lars Nygaard, Joel Priestley, Anders Nøklestad. 2008. Glossa: a Multilingual, Multimodal, Configurable User Interface. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).* Paris: European Language Resources Association (ELRA).

Paul Meurer. 2012. Corpuscle – a new corpus management platform for annotated corpora. In: Gisle Andersen (ed.). *Exploring Newspaper Language: Using the Web to Create and Investigate a large corpus of modern Norwegian, Studies in Corpus Linguistics 49*, John Benjamins, 2012.

Web sites

CANS (Corpus of Norwegian-American Speech): http://tekstlab.uio.no/norskiamerika/english/index.html

CLARIN federated content search: https://www.clarin.eu/content/federated-content-search-clarin-fcs

CLARINO: http://clarin.b.uib.no/

Clojure: https://clojure.org/

ELENOR: http://www.hf.uio.no/ilos/studier/ressurser/elenor/index.html

Glossa on GitHub: https://github.com/textlab/cglossa

IMS Open Corpus Workbench: http://cwb.sourceforge.net/

Leksikografisk bokmålskorpus:https://tekstlab.uio.no/glossa2/?corpus=bokmal

MySql: https://www.mysql.com/

Nordic Dialect Corpus: http://www.tekstlab.uio.no/nota/scandiasyn/index.html

NORINT: http://www.hf.uio.no/iln/english/about/organization/text-laboratory/projects/norint/index.html

NoWaC (Norwegian Web as Corpus): http://www.hf.uio.no/iln/om/organisasjon/tekstlab/prosjekter/nowac/index.html