Chung-Hwa Buddhist Journal (2012, 25:87-104)
Taipei: Chung-Hwa Institute of Buddhist Studies
中華佛學學報第二十五期　頁 87-104 (民國一百零一年)，臺北：中華佛學研究所
ISSN:1017-7132

The Corpus Search and Results Handling System Glossa – a Description

Janne Bondi Johannessen
The Text Laboratory, Department of linguistics and Nordic studies
University of Oslo

Abstract

The paper presents and describes Glossa, a corpus search and results handling system that has two main characteristics: It is advanced with respect to search and handling options, and it is very user-friendly. Also, it is freely downloadable. The system is suitable for monolingual and parallel corpora, and for combining different kinds of information in the search results. In the paper I show how sound, video and maps, as well as sets of double transcriptions, are presented to the Glossa user.

Keywords:

Corpus Search System, User-friendly Interface, Advanced Search, Parallel Corpora, Speech Corpora

語料庫搜尋與結果處理系統 Glossa 之說明

Janne Bondi Johannessen
奧斯陸大學

摘要
此篇文章介紹並說明語料庫搜尋與結果處理系統—Glossa，此系統有兩個主要的特性：它具有搜尋與處理選擇上的優越性，並相當考慮使用者的需要。另外，此系統可免費下載，且適用於單一語言及平行語料庫，同時可以在搜尋結果中結合不同的資訊。在此文，我將說明聲音，影像及地圖，及一套雙重抄寫如何呈現於 Glossa 的使用者。

關鍵詞：語料庫搜尋系統、人性化用戶界面、進階搜尋、平行語料庫、口語語料庫

Introduction[1]

The paper presents and describes Glossa, a corpus search and results handling system that has two main characteristics: It is advanced with respect to search and handling options, and it is very user-friendly. Also, it is freely downloadable, which means that those who have a corpus and would like it to be available on the web in a nice interface, can use Glossa. Many corpora are used with Glossa both at the University of Oslo and elsewhere. The paper is structured as follows. In section 2, I briefly describe the importance of user-friendliness. Section 3 illustrates querying with Glossa, showing options with as different texts as parallel corpora and speech corpora. That section concludes with an illustration of the indispensability of Glossa for certain types of research. The illustration shows how finding isoglosses for variation in noun morphology depend on the Glossa options of maps and parallel transcription search. Section 4 gives the technical details, including requirements on input data and a small discussion on the use of Google APIs. Section 5 concludes the paper.

Importance of User-Friendliness

There are several corpus interfaces available, see e.g. Johannessen et al. (2000), Bick (2004), Hoffmann and Evert (2006). However, they often 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, and few are completely GUI-driven.

Typically, corpus applications require queries to be formed as regular expressions in some formal language. 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.

Many corpus users find it hard to have to know the tag inventory, tag names and necessary tag abbreviations, as well as abbreviations for source texts, etc. For many potential users, these issues act as an obstacle, preventing them from making easy or efficient use of corpus 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,

1 I would like to thank the two anonymous reviewers for very good advice. I also thank my colleagues at the Text Laboratory, University of Oslo, especially Joel Priestley, Anders Nøklestad and Kristin Hagen, who are vital for the Glossa development and Lars Nygaard for his important contributions in the early development phase.

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.

Querying with Glossa

The corpus user can query the corpus by linguistic features or by non-linguistic features, or by a combination. The most common linguistic queries involve specifying a token by given attributes: word, lemma, affix or part of word (start, middle, and of word), part of speech, morphological features, syntactic functions, 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 *jump*. In figure 1 we see the same query in Glossa, with its pull-down menus. (The latter search is translated by Glossa into regular expressions.)

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

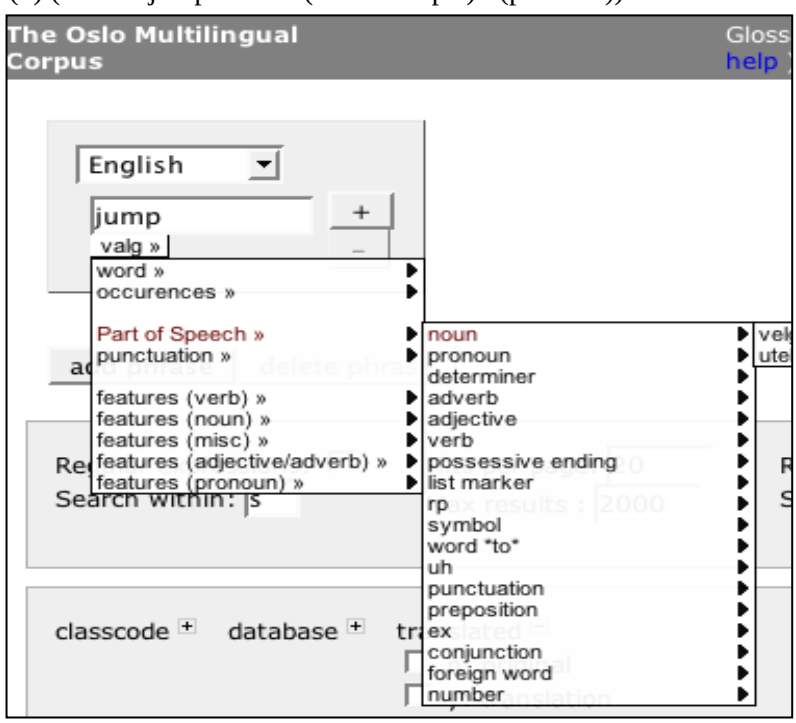

Figure 1: Querying Glossa using linguistic specifications.

All searches are done using checkboxes, pull-down menus, or writing simple letters to make words or other strings.

The querying in figure 1 is a monolingual search. In figure 2 we see how a query can address more than one language in a parallel translation corpus. The user has indicated that (s)he wishes to get all hits where the English text contains *jump* followed by a preposition, and the Norwegian translation equivalent contains *hopp*.

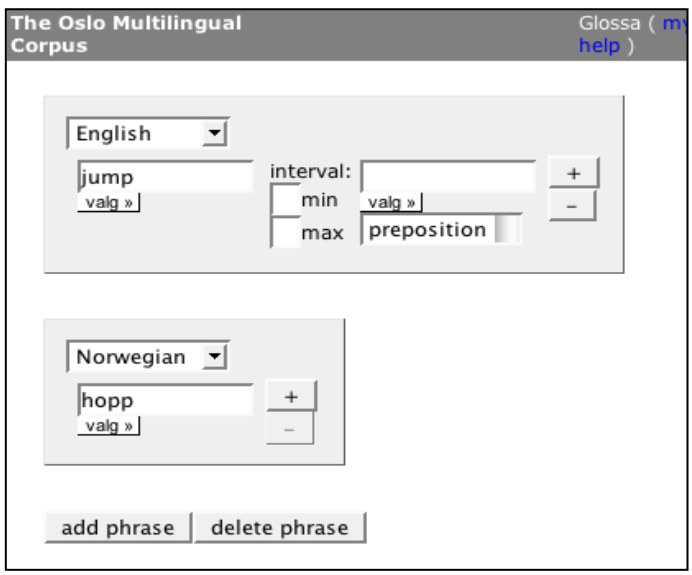

Figure 2: Querying for a parallel search.

The parallel search in figure 2 is translated to a regular expression by the system, presented in (2), and the search results are presented in figure 3. Without the interface, the users would themselves have to write this regular expression.

(2) "([((word="jump" %c))][((pos="prep"))]) :OMC4_NO ([((word="hopp.*"%c))]) ;"

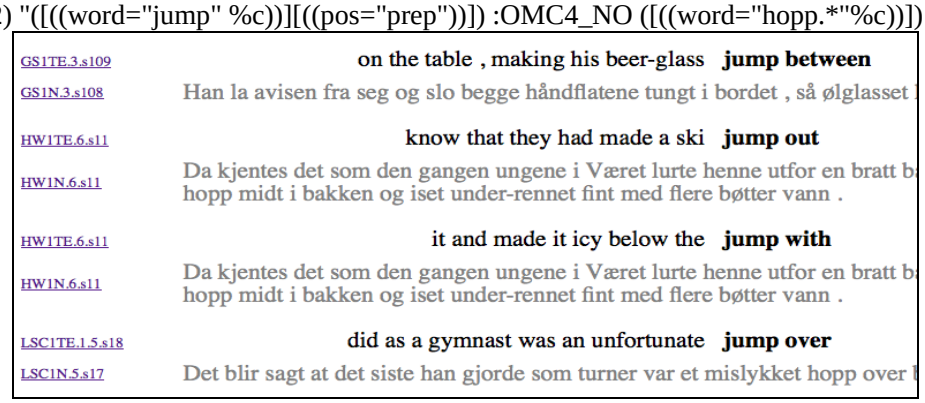

Figure 3: Some results from a parallel corpus query.

The examples we have seen up to this point are ones that query linguistic features. The corpus user can also filter the searches by non-linguistic features, on the same query page. Here the choices are hidden in clickable boxes that appear when the little plus-sign is expanded. This is shown in figure 4.

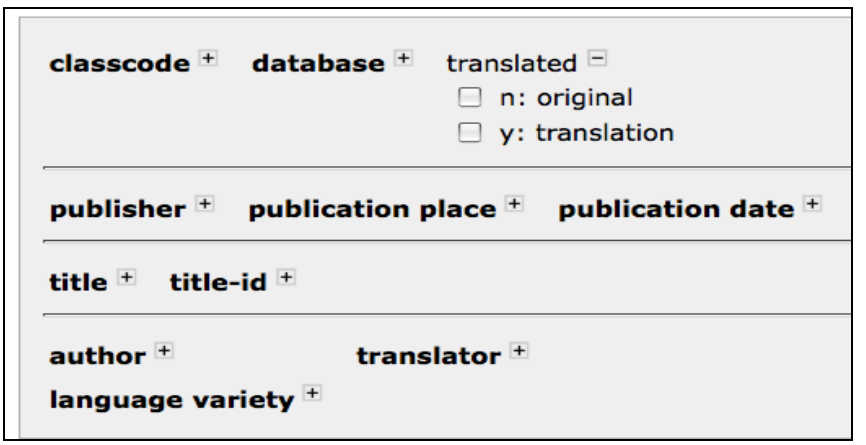

Figure 4: The non-linguistic features that can be used for filtering the search.

The Nordic Dialect Corpus (Johannessen et al. 2009) is a corpus that contains a lot of information: the speech of five languages, hundreds of dialects, tagging, at least one transcription for each dialect, but sometimes two, information on informants, like age and sex, type of recording, recording year etc. But querying the corpus is no more complicated than with simpler corpora.

Figure 5 illustrates how to search for a suffix. The suffix *–um* in the Övdalian dialect of Sweden has this suffix for two functions: dative plural on nouns and 1st person plural on verbs. It is an interesting search option since the Övdalian dialect is rapidly changing, and one does not know whether the various inflections are still used. (They are no longer used in standard Swedish.) Especially use of the dative suffix is clearly rapidly losing ground. Since *–um* is a non-standard suffix, there is no standard orthography for handling it, and consequently, a standard orthographic search is not viable. The user must specify that the search should be performed in the phonetic transcription (the option at the bottom of the pull-down menu). The standard orthography simply leaves out the suffix altogether, so there is no way this could be used (see figure 7 for the results of the search, illustrating also the difference between the two types of transcription).

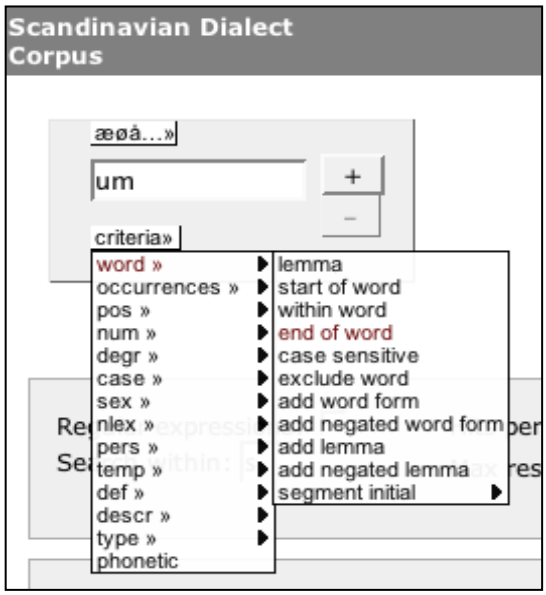

Figure 5: A simple query for the suffix *–um*.

Figure 6 illustrates many of the non-linguistic variables that can also be used to limit the search. In addition to those regarding informants, there are also some other choices that deal with the presentation of the result (top of figure 6). I will mention particularly the option of choosing one or two or both types of transcription. This option is irrelevant of whether the researcher originally searched in the phonetic or orthographic transcription.

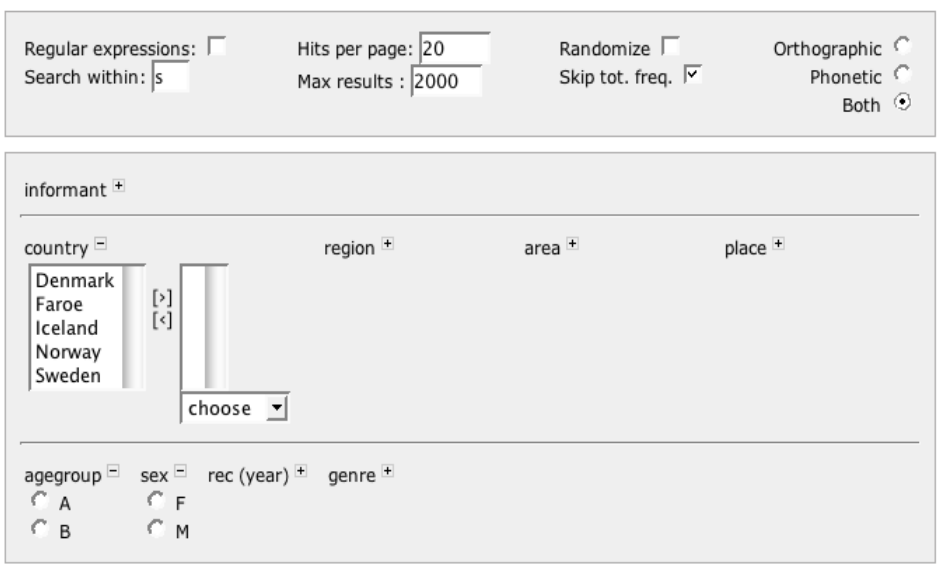

Figure 6: Non-linguistic search options in the Nordic Dialect Corpus.

In figure 7 some of the search results for the suffix query for *–um* are displayed, and we see how the two transcriptions complement each other.

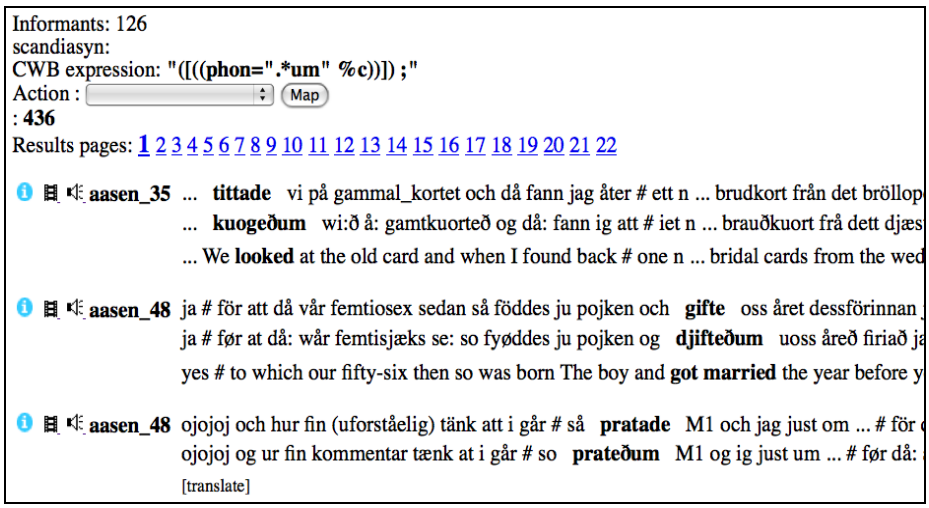

Figure 7: Three different displays of search results: two types of transcription (orthographic and phonetic) plus an English translation – in that order.

Without being able to search in the phonetic transcription we would not have been able to find these suffixes. Without the orthographic transcription a non-expert dialect speaker would not have been able to understand the phonetic transcription, given how far it is from the standard. We would like to point to the fact that the displayed results are translated to English by using a Google Translate API. This has to be done for each concordance line separately, and is a service to less proficient, Nordic language speakers.

Figure 8 shows what the results window looks like when the film icon button next to a result line is pushed. The video and audio give exactly the same segment as the text line in the results list.

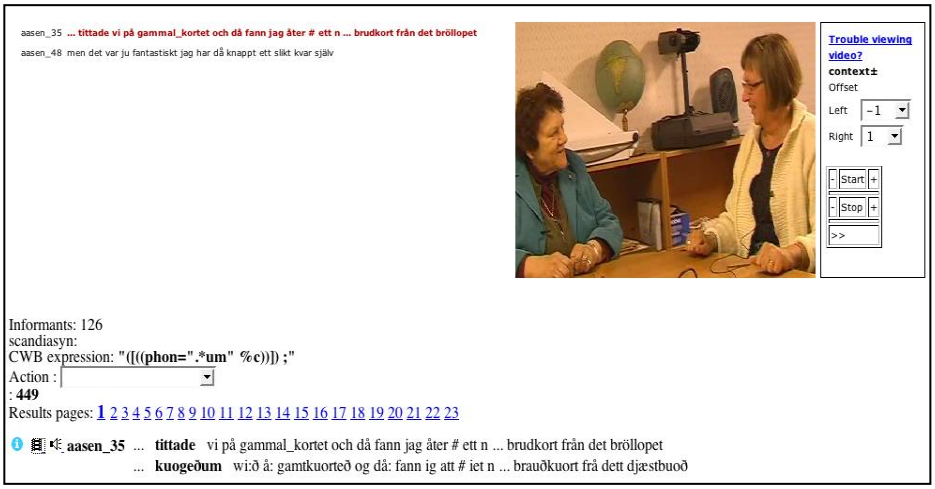

Figure 8: Search results with audio and video.

In addition to the many search options, there are also various options for handling the results. The Action menu visible in figure 7 and 8 gives a large selection of choices, for example: sorting on matching phrases, bibliographic information or arbitrary points in the context, counting matched phrases, downloading result sets 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, frequency count of all hits.

In figure 9, we have simply asked for a count of the results from a search on *jump* as first part of a word. This option gives the researcher a very nice overview of the words of the resulting search concordance. Here, the case-sensitive option has been chosen, thereby distinguishing *jump* from *Jump*. This is a choice the user has to make before displaying the result of the count.

occurences	match
88	**jumped**
64	**jump**
30	**jumping**
27	**jumps**
15	**jumper**
4	**Jumping**
2	**jumpers**
2	**jumpy**
2	**Jumping-Off**
1	**jumped-up**
1	**Jump**
1	**jump-racing**
1	**jumped...**

Figure 9: Word count

The word count can also be represented by a pie chart or a histogram, among other things, as illustrated in figures 10 and 11:

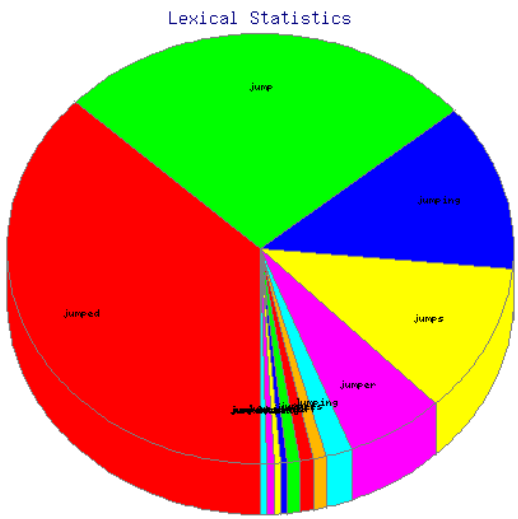

Figure 10: Frequency displayed as a pie chart.

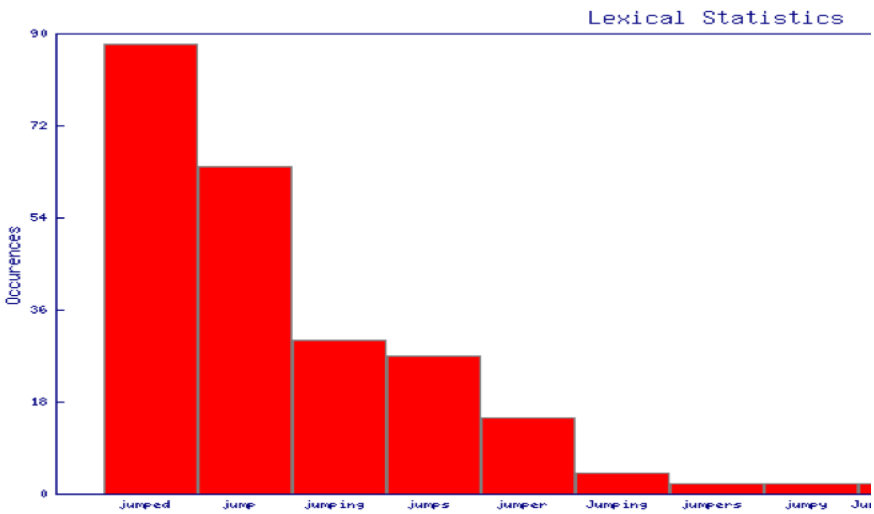

Figure 11: Frequency displayed as a histogram.

The Action menu also gives the possibility of showing collocation data, as in figure 12.

Left context				Right context			
ngram	**rank**	**AM**	**occ**	**ngram**	**rank**	**AM**	**occ**
and **	3	0.1622	21	** up	1	0.2305	31
he **	5	0.1111	14	** .	2	0.1762	23
, **	6	0.0960	12	** ,	4	0.1406	18
i **	6	0.0960	12	** to	6	0.0960	12
the **	7	0.0884	11	** and	6	0.0960	12
to **	8	0.0806	10	** out	7	0.0884	11
she **	8	0.0806	10	** into	9	0.0729	9
ski **	9	0.0729	9	** over	10	0.0650	8
a **	11	0.0571	7	** in	10	0.0650	8
would **	13	0.0412	5	** off	11	0.0571	7
from **	15	0.0249	3	** on	12	0.0492	6
little **	15	0.0249	3	** down	12	0.0492	6
his **	15	0.0249	3	** from	13	0.0412	5
him **	15	0.0249	3	** at	13	0.0412	5
. **	15	0.0249	3	** onto	14	0.0331	4
have **	15	0.0249	3	** with	15	0.0249	3
all **	15	0.0249	3	** ashore	15	0.0249	3
my **	15	0.0249	3	** or	15	0.0249	3
go **	16	0.0167	2	** the	15	0.0249	3
they **	16	0.0167	2	** jockeys	16	0.0167	2

Figure 12: Collocations

As mentioned, Glossa is continuously being developed and is getting new features. I have not shown all the options that can be had with this corpus search and results handling system, but I would like to mention one of the newest additions to the system; that of showing maps for each concordance line. Thus, if we make a search for some feature that is distributed geographically, a map display is very useful.

I choose to present a final example that illustrates how useful the Glossa options are for linguistic research, by the overall research question of isoglosses for noun morphology. A topic that has interested Norwegian dialectologists over many years is the distribution of the various noun suffixes. While detailed maps were drawn for the noun morphology in the mid 1900s, it is expected that the situation differs now, but it is costly to do a full dialect survey only for this topic. With the Nordic Dialect Corpus available in Glossa, a simple search for a specific, common noun such as *ungene* 'the children. MASCULINE, will give the desired results revealing the geographical distribution of this plural definite suffix within seconds. There are 568 hits, and the results page shows each form of the noun as in figure 13, and the geographical distribution on a Google map, as in figure 14. It should be mentioned that I could also have chosen to search for just the string *–ene* 'plural definite suffix', but have chosen not to do so here, since that would have given hits for all three genders (neuter and feminine as well). Since many dialects distinguish the plural definite suffix according to gender, I would have gotten many more forms, which I do not find useful for this illustrative example.

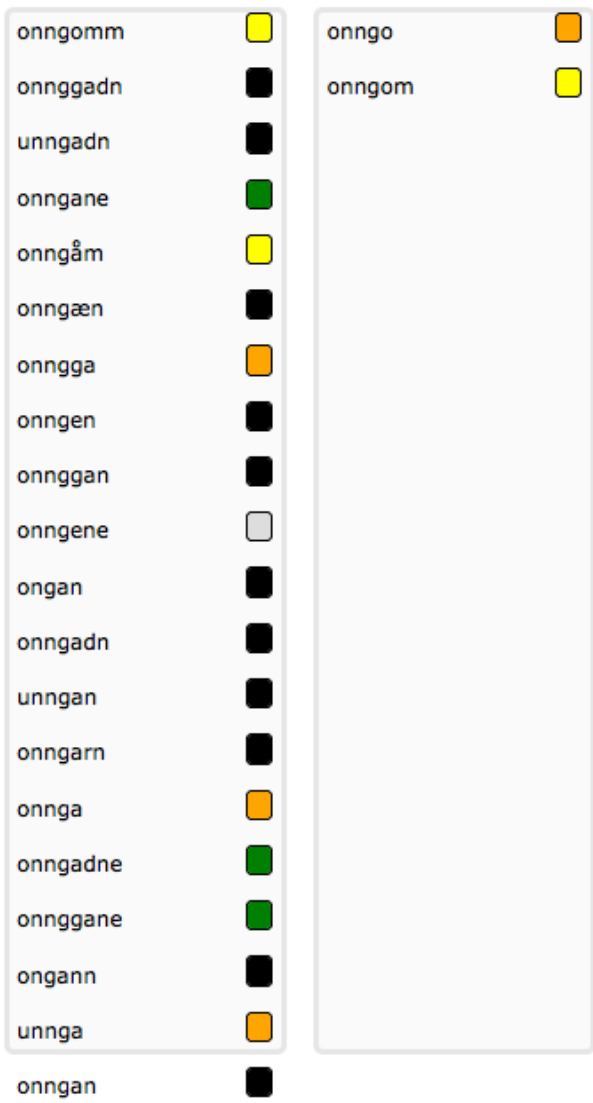

Figure 13: The full range of pronunciations of the word *ungene* 'the children in the Nordic Dialect Corpus, transcribed in a traditional Norwegian system.

The corpus users themselves choose which words to group together by way of a colour code. Here I have chosen to distinguish between three types: the full two-syllable suffix *–ane* (green [editor's note: download PDF for color reproductions]), the apocoped one-syllable suffix *–an* (black), the short non-nasal suffix *–a* (orange), and the dative suffix consisting of a rounded vowel and a bilabial consonant *–om* (yellow). In figure 14 the geographical distribution of these types is clearly displayed, and the isoglosses easy to see.

Figure 14: Map with the distribution of the plural definite suffix.

The map shown that the full suffix *–ane* (green markers) is commonly used in the south and west parts of southern Norway. The apocoped suffix *–an* (black) is used in all of north Norway and the middle part of south Norway down to the coast. The short suffix *–a* (orange) is mainly found in the eastern part of south Norway and in one place in north Norway. The latter could be evidence of an immigrant group that came to this area in the 1700s–1800s from the eastern valleys of south Norway, a fact that, even today, is clearly reflected in the language. The dative suffix *–om* (yellow) is only found in a few places in the northern part of south Norway. Dative case is slowly dying in Norway, just as it is on the other side of the border (recall the discussion on the similar Övdalian dative case suffix *–um*).

The last suffix search with the resulting map illustrates two important features of Glossa. If it had not had the possibility of searching for aligned phonetic transcription variants via an orthographic search, finding so many versions of the suffix would have been nearly impossible. With only access to orthographic transcription, no variation would have been found, and conversely, with access only to phonetic transcription, a

comprehensive search would have required detailed knowledge of all the dialect forms, a near-impossible requirement. The second important feature for this search is the map. Without the visual illustration, the isoglosses would have been hard to spot – with so many places and so many linguistic forms.

Technical Details

Glossa (Nygaard 2007, Johannessen et al. 2008) is implemented partly through new programming and partly with other reusable resources. The corpus search part is performed with the IMS Corpus Workbench (CWB, Christ 1994), and the meta information is put into a relational MySQL database. Although the web interface is simple, it allows users to create complex queries in very simple ways, browse, process, download result sets etc. Glossa supports all types of corpora, both multilingual and multimodal, with various amounts and kinds of annotation. The statistics options are implemented with the Ngram package (Pedersen 2008). Google Translate and Google Maps are used for added value of display of search results. All in all, as indicated, Glossa combines several features and functions, and makes them all available in the same user interface. We know of no other interface that combines so many options. We have used existing APIs for the programs mentioned here, but have not developed additional ones.

The use of Google APIs for translation and map functions deserves a comment. Google is a commercial company with whom one does not communicate directly. They offer good quality programs via APIs for free and have therefore been a valuable choice for us. For example, we could have gotten Norwegian electronic maps free, from a different company, as a university institution. However, we needed maps for many countries in Northern Europe, while our institution only had agreements with one company for Norwegian maps. Thus, Google's free service turned out to be our only option. Their API covers a lot and their functionality is good. A problem with Google, however, is that they as a commercial company change their terms of service along the way. Thus, the translation option that we have described here, was provided free of charge, but now has to be paid for. Using Google thus makes some of the modules less predictable in the long run.

When it comes to formats, Glossa needs texts to be in the format required by the CWB, i.e., tab-separated text with XML tags. Glossa uses the XML tags for structural (i.e., not about individual words) information such as sentence ID and time codes (for audio and video files). If input texts come with TEI or other XML markup, information from these tags will be extracted and inserted into the MySQL database. For Glossa to be able to communicate with the third-party services (for example maps) and to link the corpus text directly to audio and video, the corpus must have markup that includes latitude/longitude coordinates and time codes, respectively. Grammatical tags are part of the input tab-separated text, and must be mapped to the menu-structure of the search-

interface. Mapping for TreeTagger for some languages is included in the system. But any tag set and values can be imported. Glossa itself requires simply text in tab-separated format and a MySQL database for metadata (extra-linguistic information on informants or text sources). The programming languages used are Perl, Ruby, PHP and JavaScript. CWB allows Unicode.

Configuration of the interface and the mapping from corpus data to menus and search options is achieved using a set of corpus-specific configuration files. Search results can be exported to several formats, such as tab separated and comma separated text, Excel etc.

Glossa is freely downloadable on a GPL licence from GitHub, and is undergoing regular development and improvement in close contact between users and developers. Some installation support can be given upon request. The Glossa package includes scripts that convert written texts in TEI formats, as well as spoken language in Transcriber-XML, into a full corpus and database.

There are many types of corpora that use Glossa (speech corpora, parallel, written corpora, and monolingual written corpora). For a list, please consult the end of the paper.

Conclusion

The paper describes some features of the corpus search and results handling system Glossa, developed at the Text Laboratory, UiO. We have seen that the basic search system is the same for any kind of corpus, but that specific features (audio, audio or translated texts) will give various additions to the usability. Glossa is currently used for monolingual and multilingual, parallel written corpora and for speech corpora with audio and video.

The Glossa system is freely downloadable (see web site below) and some support can be given for corpus installation.

References

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

Christ, Oli. 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. *Complex'* '94. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences.

Evert, Stefan. 2005. *The CQP Query Language Tutorial*. Germany: Institute for Natural Language Processing, University of Stuttgart. (*http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial*)

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

Johannessen, Janne Bondi; Nøklestad, Anders; Hagen, Kristin. 2000. A Web-Based Advanced and User-Friendly System: The Oslo Corpus of Tagged Norwegian Texts. *Second International Conference on Language Resources and Evaluation. Proceedings.*

Johannessen, Janne Bondi; Nygaard, Lars; Priestley, Joel; Nøklestad, Anders. 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).

Johannessen, Janne Bondi; Priestley, Joel; Hagen, Kristin; Åfarli, Tor Anders; Vangsnes, Øystein Alexander. 2009. The Nordic Dialect Corpus - an Advanced Research Tool. Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. *NEALT Proceedings Series Volume 4*. Eds. Jokinen, Kristiina and Bick, Eckhard Bick. Denmark: Northern European Association for Language Technology.

Nygaard, Lars. 2007. *The Glossa Manual.* Norway: The Text Laboratory.

Pedersen, Ted. 2008. *Ngram Statistics Package*. (*http://www.d.umn.edu/~tpederse*)

Corpora that use Glossa

Big Brother Corpus (Speech), Norwegian: *http://www.tekstlab.uio.no/nota/bigbrother/*
The European Parliamentary Comparable and Parallel Corpora (ECPC) (under development): *http://www.ecpc.uji.es/EN/home.php?language=en*
Lexiographical Bokmål Corpus: *http://www.hf.uio.no/iln/forskning/samlingene/bokmal/index.html#bokmalskorpus*
Lule Sámi Corpus: *http://giellatekno.uit.no/doc/lang/corp/corpus-smj.html*
Macedonian Text Corpus: *http://www.tekstlab.uio.no/glossa/html/index_dev.php?corpus=mak*
Mörkuð íslensk málheild (Icelandic Corpus): *http://mim.hi.is/*
Nordic Dialect Corpus (Speech): *http://www.tekstlab.uio.no/nota/scandiasyn/*
North Sámi Corpus: *http://giellatekno.uit.no/doc/lang/corp/corpus-sme.html*
NoTa Oslo Speech Corpus: *http://www.tekstlab.uio.no/nota/oslo/*
Oslo Multilingual Corpus: *http://www.hf.uio.no/ilos/OMC/*
Ruija Speech Corpus of Kven: *http://www.hf.uio.no/iln/tjenester/kunnskap/sprak/korpus/talespraksk orpus/ruija/index.html*
RUN Parallel Corpus: *http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/run/corpus/*
TAUS Speech Corpus of Norwegian: *http://www.tekstlab.uio.no/nota/taus/index.html*
UPUS Speech Corpus Multiethnic Norwegian: *http://www.hf.uio.no/iln/forskning/prosjekter/upus/*

Other Web Sites

GitHub: *https://github.com/*

Glossa: *http://www.hf.uio.no/tekstlab/glossa.html*

Google Translate: *http://translate.google.com*

IMS Corpus Workbench: *http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/*

MySQL: *http://www.mysql.com*

Open Source: *http://www.opensource.org*

Text Laboratory: *http://www.hf.uio.no/tekstlab/*