A Web-Based Advanced and User Friendly System: The Oslo Corpus of Tagged Norwegian Texts

Janne Bondi Johannessen, Anders Nøklestad, Kristin Hagen

The Text Laboratory, University of Oslo
P.O. Box 1102 Blindern, N-0317 Oslo, Norway
{ j.m.b.johannessen, anders.noklestad, kristin.hagen }@ilf.uio.no

Abstract

A general purpose text corpus meant for linguists and lexicographers needs to satify quality criteria at at least four different levels. The first two criteria are fairly well established; the corpus should have a wide variety of texts and be tagged according to a fine-grained system. The last two criteria are much less widely appreciated, unfortunately. One has to do with variety of search criteria: the user should be allowed to search for any information contained in the corpus, and with any combination possible. In addition, the search results should be presented in a choice of ways. The fourth criterion has to do with accessability. It is a rather surprising fact that while user interfaces tend to be simple and self explanatory in most areas of life represented electronically, corpus interfaces are still extremely user unfriendly. In this paper, we present a corpus whose interface we have given a lot of thought, and likewise the possible search options, viz. the Oslo Corpus of Tagged Norwegian Texts.

1. Introduction

A general purpose text corpus should satisfy quality criteria at a number of levels in order to fulfill the needs of the majority of its users. The major users we take to be linguists working in academia, or to some extent in commercial enterprises, such as dictionary publishing. Although corpora exist for many languages, and new corpora are being created all the time, it is surprising how little effort is put into making them fully useful as tools. While creating the Oslo Corpus of Tagged Norwegian Texts, we found that the following criteria ought to be fulfilled:

1) Variety of corpus texts

Given that it is impossible to know beforehand what kind of questions linguists (morphologists, syntacticians, semanticists, lexicographers) will want to ask, and to what extent dictionary writers will use the corpus, it is important to ensure that the corpus has a certain size and comprises a variety of genres, and of written standards, if there are any.

2) Variety of grammatical tagging

It is vital that grammatical tagging goes beyond simple part-of-speech tagging: In many languages, there is, even within parts of speech, a lot of homonymy that could be disambiguated with a more fine-grained system of tags. Obviously, the tagging also has to be correct.

3) Variety of search options

Every feature of the corpus, marked or unmarked, should be searchable. E.g., if the corpus consists of several genres, and the texts are grammatically tagged, it should be possible to search for all occurrences of a certain grammatical feature in a certain genre.

4) Wide accessability

The corpus must be easily accessible. In our day and age, this means accessible with a simple search interface on a widely available medium. Importantly, and this is not trivial, the search interface should also be easy to use for the major user group, i.e. philologists in a broad sense. If the interface is complicated to use, it will be perceived of as unaccessible, and hence not used.

In this paper we will present the Oslo Corpus of Tagged Norwegian Texts, which we have developed in an attempt to fulfill all of the above criteria. The first two criteria have often been emphasized in the literature, and we will not discuss them here. Let us just mention that the Oslo Corpus consists of 23 million words, divided between two written standards (bokmål and nynorsk), and three main genres (newspapers and magazines, non-fiction such as laws and parliamentary reports, and fiction). The corpus has been tagged with a Constraint Grammar tagger (Karlsson et a. 1995), with a high level of accuracy - a recall of 99.3% and a precision of 95.5% (Johannessen et al 1998, Hagen et al, in print). The paper will focus on the last two criteria, which we believe are very important to guarantee success for a system. We will give some examples of unusual tagging and search options, together with some indications of why these may be useful.

2. Variety of Search Options

The search system for The Oslo Corpus of Tagged Norwegian Texts is based on the IMS Corpus Query System developed at the University of Stuttgart, with a Web interface for the Oslo Corpus of Bosnian Texts (Santos 1998), which has since been changed into a maximally user friendly system. The IMS system makes it possible to design the corpus in such a way that all tags and text source codes can be used as search criteria. Also, it offers a wide variety of options for presenting search results, such as collocations and concordances, and distribution of forms. It also makes it possible to do positive as well as negative searches. In this section, we shall focus on what kinds of search options can be useful and how the results might be presented. Although grammatical tags are perhaps the most important search alternative, we shall not say very much about those, since this is common knowledge. Instead, we shall focus on

some rather unusual categories, such as compounds, words that are not in the lexicon etc.

2.1. Grammatical tags

In order to give the user a wide selection of search options, we decided to tag the corpus in more ways than is common. Let us first mention, however, that the morphosyntactic tagging includes a wide range of tags in a very fine-grained system. There are approximately 200 tags distinguishing between e.g. the following types of words:

Word	Tag
1	det quant sg
hver 'each'	det quant fem sg
fem 'five'	det quant pl
alles 'everybody's'	det quant pl gen
en 'a'	det quant masc sg
ens 'one's'	det quant masc sg gen
et 'a'	det quant neut sg

Table 1. Some morphosyntactic tags

We have also tagged the corpus with syntactic tags, following the CG dependency grammatical conventions:

@<ADV	ADVERBIAL modifying a word on its left
@<DET	DETERMINER modifying a word on its left
@<P-UTFYLL	PREPOSITIONAL COMPLEMENT modifying a preposition on its left
@I-OBJ	INDIRECT OBJECT
@SUBJ	SUBJECT

Table 2. Some syntactic tags

2.2. Compounds

We have chosen to tag the corpus in some rather unconventional ways in addition to the grammatical ones. We believe that any information that is available in the corpus should in principle be available for searching.

One such piece of information is productively formed compound words. Norwegian is a language where compounds play an important role, both as a way of creating new words for new concepts or translations, and as a way of varying one's language. For example, a (football) *keeper* is a *målmann* (goal man), and instead of saying the Norwegian equivalent of "the paper for the LREC proceedings", one might say *LREC-artikkelen* or *konferanseartikkelen* or *rapportartikkelen.* Since compounding is a very productive process in the language, it is obvious that most compounds cannot be in the dictionaries, and hence not in the tagger's lexicon. As a result, any tagger for Norwegian needs a compound analyzer. Instead of tagging the words analyzed by the compound analyzer anonymously with grammatical tags only, we also mark them as "compounds", making them searchable. This makes new research methods possible, such as measuring an aspect of the creativity of different authors, simply by counting the number of compounds in their texts. Also, such productively created words give an idea of the main theme of a text, a fact that could be made useful for automatic document summarizers, pointing to yet another possible use of a richly tagged corpus. Below is an example of the results of a compound search in a novel.

Compound word (Norwegian)	English translation
Pariser-kommunenes	Paris communes
arbeiderbataljonene	worker battalions
papirflagg	paper flags
Trekkspillåt	accordion tune
kaffemelk	coffee milk
lampeskjæret	lamp shine
Sacre-Coeur-kirken	Sacre Coeur church
militærmaskiner	military machines
jakkeopslaget	jacket collar
firti-timers-dagen	fourty hour day
klasseaksjon	class action
maktutvidelse	power expansion
likbyen	dead body town
kirkegårdsveien	church yard road
dødsmuren	death wall

Table 3. Some compounds in Nordahl Grieg: *Spansk sommer.*

2.3 Non-standard words and grammar

While developing the grammatical tagger, we realized that a lot of the recurring words (spellings), inflections and derivations that are in common use are actually non-standard, whether it is because they are old-fashioned or because they have never entered the official norm. Since any tagger performs best if the individual words are analyzed correctly, we included a lot of non-standard words in the lexicon used for the tagger. Like we did for the compounds, we decided to mark these words especially, to be able to retrieve them if so desired.

Non-standard	Standard	English translation
billedet	bildet	the picture
bragte	brakte	brought
syv	sju	seven
idag	i dag	today
kolonihave	kolonihage	allotment garden
efter	etter	after
melkefarvede	melkefargede	milk coloured
hverken	verken	neither
ennu	ennå	still
mellemgulvet	mellomgulvet	diaphragm
hugget	hogget	chopped
stenen	steinen	the stone

Table 4. Some non-standard spellings and inflections in Nordahl Grieg: *Spansk sommer.*

This kind of tagging is of course extremely useful for dictionary makers and language councils that draw guidelines for language use. In some languages, like Norwegian, non-standard words (spellings and grammar) belong to a finite number of subgroups, such as radical forms, conservative, forms, youth jargon etc. Doing a search for non-standard words will quickly reveal the style of an author. (The orthography of the above author is old-fashioned by today's standard.)

2.4 Unrecognized words

During the process of developing the tagger, we found that a not insignificant number of words remained unanalyzed by the tagger. Using the same idea as before, we tagged these words as unanalyzed, making them possible to retrieve.

Word	Why not recognized
læll	dialect
«schönt»	German
seg.2	misprint
mellom30	misprint
ship	English
forskninng	spelling error
bedrifer	spelling error
komb.	nonstandard abbreviation
tøfft	spelling error
fouls	English
èn	wrong diacritic
ungpian	dialect

Table 5. Some words not recognized in the newspaper *Adresseavisa.*

This is useful for several purposes. First, of course, it reveals words that we might consider to include in the lexicon for new tagging purposes, or that might be considered by dictionary writers. Second, it gives language councils more material to work with with respect to kinds of mistakes that people make in standard orthography and inflection. Third, it can be used to detect the extent to which words and phrases from foreign languages are used in different texts.

2.5 Simple and combined search options

We think that it is important to be able to search for any category or feature in the corpus. One should think this would be a matter of course, but it turns out to be less common than one might think.

For example, the SARA Windows client for the well-known British National Corpus does not allow selecting part of speech for a search query "without specifying the word to which it is attached" (Reference Guide to the SARA Windows Client: 3.5). This of course makes it impossible to ask for even a simple query like "verb followed by preposition". Similarly, the search pattern for the on-line corpora at the Linguistic Data Consortium do not allow searches for parts of speech: "[I]f the overall pattern does not contain any specific words, the program will refuse to search for it. Thus the following are illegal patterns: V, "V NNP", "VB NN", "N+ V V" (from the LDC web page *Searching LDC text corpora*).

In the Oslo Corpus, it is possible to search for any combination of words and morphosyntactic tags, a fact which gives the user a better chance of finding answers to his or her questions.

2.6 Quantity

Although what is most important is quality, with respect to correctness of tagging and of the texts generally, quantity is also an important variable. There are small, good quality corpora available (e.g. the 1 million word ICE-GB corpus for English), but it is often necessary to have a bigger corpus. For many types of lexical studies, a small corpus simply will not give enough occurrences of each item to be useful. Many grammatical phenomena are also too marginal to be studied within a small corpus only. The Oslo Corpus consists of two variants of Norwegian, with 23 million words altogether, which we consider to be a minimum.

Quantity is also relevant with respect to search results. Some corpus systems actually limit the amount of hits that are available to the user (e.g., the Swedish Parole Corpus gives a maximum of 1000 hits), which may be a problem for some kinds of studies.

2.7 Presentation of search results

The IMS program that we use has many virtues when it comes to presentation of the search results. The results can be shown as concordances, collocations, and with counts for the distribution of variable search criteria. We have added choices of sorting the results with respect to source or alphabetical order for the search criterion, or for the words to its right or left. This may be very useful for example when the context of a word is important for the study in question. Also we give the user the option of seeing the tags of the search word or of all the words in the context, an option which is actually very rarely available. We know of no other system where this is possible.

3. Wide Accessability

The fourth criterion - having to do with accessability and, in particular, user friendliness - we find particularly important. Judging from many other text corpora, we seem, surprisingly, to be rather alone in this respect. We know of only one corpus search interface where user friendliness has been taken seriously to the extent that it is easy and self-explanatory, viz. the Zürich Web Interface for the British National Corpus.

3.1 A user friendly interface

In the Oslo Corpus we have developed further the IMS CQP system, combining the flexibility of regular expressions with the boxes and links offered by HTML. The result is an advanced search system of clickable boxes, which makes it possible for any linguist to make advanced searches in an easy and intuitive way, without knowing anything about regular expressions - or about the tag names. The system allows searching for three words, strings, categories or a combination of these, with any number of words in between. For each wordstring, the user can specify by clicking in boxes whether she wants that particular wordform, or whether the given string is a

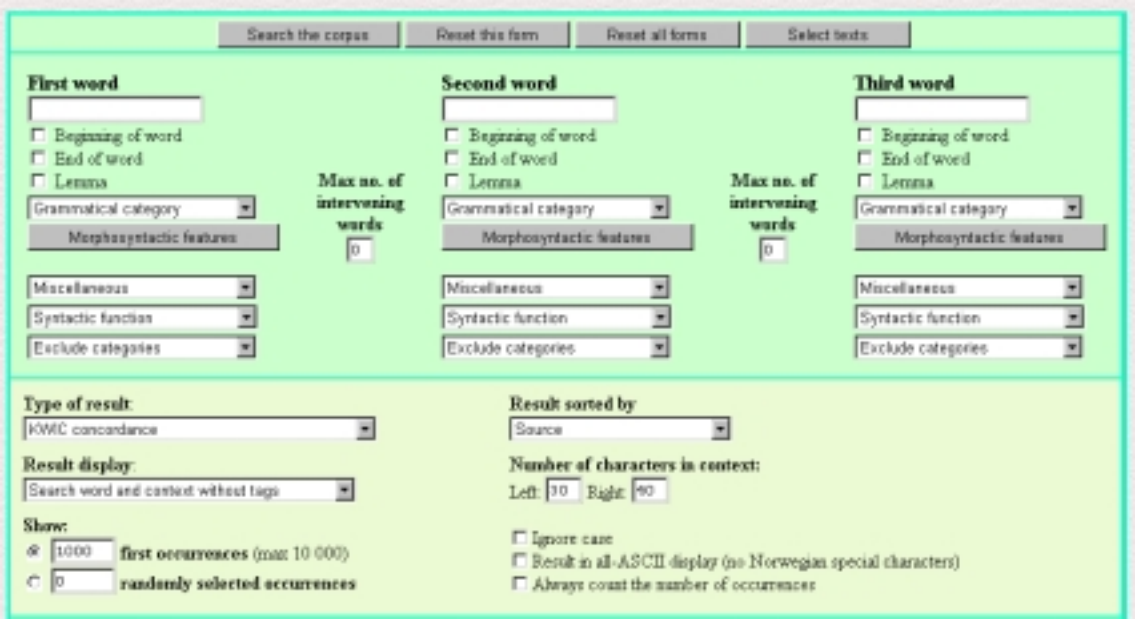

Figure 1. The web interface of the Oslo Corpus.

prefix, suffix or a lemma. It is of course also possible to specify any grammatical category or feature with or without linking the search to a particular string. Clicking in boxes is also the way to choose text categories or particular texts, or to specify how the results should be presented.

Let us explain in a little more detail what it looks like. The user may want to search for a sequence of two words with up to five words in between. In our system, she simply writes each word in the appropriate boxes, and writes the number five in another box. The alternative, which is still the most common in other corpora, is to write a regular expression:

Norwegian	English
[tagg=".*\"glad\".*"] []{0,5} [tagg=".*\"lingvist\".*"]	[tag=".*\"happy\".*"] []{0,5} [tag=".*\"linguist\".*"]

Table 6. Regular expression for the sequence *happy...linguist*.

For most linguists, this kind of search is not straightforward. But this example is still fairly simple: With a little bit of training, it's possible for anybody to use regular expressions. However, once the user needs to search for grammatical categories, things get worse. And this time, experienced programmers and technical linguist novices are equally unhappy. Unless one has studied the tagset used in a particular corpus in detail, it is impossible to know exactly which grammatical categories are used and what the tags look like. This is where the option of clicking in boxes really makes a difference. Consider searching for a particular word (*høy* 'high'), specified as being an adjective, with the features definite and positive. Here, there are three tags that are necessary to know about:

Norwegian	English
[tagg=".*\"høy\".*" & tagg=".* adj.*" &tagg=".* pos.*" &tagg=".* be.*"]	[tag=".*\"high\".*" &tag=".* adj.*" &tagg=".* pos.*" &tagg=".* def.*"]

Table 7. Regular expression for the sequence high (adjective, positive grade, definite form).

In the Oslo Corpus, the user simply writes the search word in a box, then selects a grammatical category from a list of categories, and grammatical features from a list of features.

Searching in a subpart of the corpus is another problem where it does not help to be an experienced writer of regular expressions. Each text in the corpus has its own code, and again, unless one has studied the whole list, it is impossible to know exactly what code a given text has. Consider for example a user that wants to search in all the newspapers, plus one particular government report and one particular novel. Three codes are, then, necessary to know. Thus the whole search expression would be as follows:

Norwegian	English
[word="høy" & tagg=".* adj.*" & tagg=".* pos.*" & tagg=".* be.*" &(src="AV.*" |src="SA/NO94/13" |src="SK/AlKa/01")]	[word="high" & tag=".* adj.*" & tag=".* pos.*" & tag=".* def.*" &(src="NEWS.*" |src="NO-F/NO94/13" |src="FIC/AlKa/01")]

Table 8. Regular expression for the sequence *high (adjective, positive grade, definite form),* in a subpart of the corpus; all newspapers, the government report No 13 from 1994, and the novel *Gaia* by Karsten Alnæs.

In the Oslo Corpus, the user does not have to write any regular expressions. We saw above how writing and clicking in boxes is all that's needed to choose grammatical categories. Specifying the desired texts for a subpart of the corpus is equally easy, this is done by selecting text from the menus corresponding to the three genres of the corpus (newspapers and magazines, non-fiction and fiction).

Let us mention that, in spite of what has been said above, there might be cases where the interface with clickable boxes is not flexible enough. We have, for example, limited the number of words in a search string to three. If one desires four words, the click interface cannot be used. We therefore offer an interface for regular expressions as well, with the full range of search options available by the CQP system. However, even then one can make use of the click interface, in order to avoid the problem of tag names discussed above: Since the click interface translates every search into regular expressions that are shown in the search result, the user can start by making a simple search, and copy the regular expression that is given for that search into the other interface, and then add the few extra bits that are needed. This way the user does not have to create the whole regular expression from scratch.

3.2 Medium

Availability also has to do with accessability. There is no medium that compares with the Internet in this respect. The Oslo Corpus has a web interface, available for a range of different versions of Internet browsers. Some of the owners of the texts in the corpus prefer non-commercial use only. Every user therefore has to declare that they will use the corpus in accord with our guide lines, before they get a user account.

4. Conclusion

A tagged corpus with texts from many sources contains a lot of information. It is important that all this information can be retrieved, i.e. that it is searchable. We find the IMS CQP system very suitable in this respect for retrieval and presentation of search results. But it is also important that during the process of tagging, there might be information that could be useful for researchers. This information ought also to be available later. For this reason the Oslo Corpus includes tags like "productively formed compound" and "unrecognized word".

Creating a sizable and tagged corpus takes a lot of time and effort. It is therefore vital that it is widely used. To this end, two kinds of availability are important. The first kind has to do with the prior knowledge that the users have - no formal training should be required. If the corpus has a complicated interface, many of its prospective users - linguists and lexicographers - will keep away from it. We therefore believe that a simple interface is essential. A system based on simple boxes and menus with alternative choices presented on screen is preferable to a system where the user is required to write regular expressions. The second kind of availability has to do with the medium in which the corpus is presented. The more accessible the corpus is, the more people will use it. A system usable on the Internet is far more accessible than on any other medium.

We believe that our system is very flexible, and at the same time very easy to use. The comments from users since we made the corpus public in September 1999 support our impression. The Oslo Corpus now has users in 15 countries, a number that is not unimpressive given the fact that the Norwegian language is not spoken anywhere else than in Norway.

5. References

Johannessen, J.B. and H. Hauglin, 1998. An Automatic Analysis of Norwegian Compounds. In T. Haukioja (ed.), *Papers from the 16th Scandinavian Conference of Linguistics*, Turku/Åbo, Finland, 209-220.

Hagen, K., J.B. Johannessen, and A. Nøklestad, 2000. A Constraint-Based Tagger for Norwegian. Paper presented at the Scandinavian Conference of Linguistics, Odense, Danmark. In print.

Karlsson, F., A. Voutilainen, J. Heikkilä, and A. Anttila, 1995. *Constraint Grammar*. Mouton de Gruyter.

Santos, D., 1998. Providing access to language resouces through the World Wide Web: the Oslo Corpus of Bosnian Texts. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada (eds.), *Proceedings from the First International Conference on Language Resource and Evaluation*, Granada, Spain, 475-481.

The British National Corpus, Reference Guide to the SARA Windows Client:
http://info.ox.ac.uk/bnc/getting/chap4.htm

The British National Corpus, The Zürich Interface:
http://escorp.unizh.ch/cgi-bunbnc2/BNCquery.pl

The ICE-GB Corpus:
http://www.ucl.ac.uk/english-usage/ice-gb/

The IMS Corpus Work Bench:
http://www.ims.uni-stuttgart.de/CorpusToolbox

The Linguistic Data Consortium:
http://www.ldc.upenn.edu/lol/textreadme.html

The Oslo Corpus of Bosnian Texts:
http://www.tekstlab.uio.no/Bosnian/Corpus.html

The Oslo Corpus of Tagged Norwegian Texts:
http://www.tekstlab.uio.no/norsk/bokmaal/
http://www.tekstlab.uio.no/norsk/nynorsk/
http://www.tekstlab.uio.no/norsk/bokmaal/english.html.

The Swedish Parole Corpus:
http://spraakdata.gu.se/lb/parole_org.html