The ASK Corpus –
a Language Learner Corpus of Norwegian as a Second Language

Kari Tenfjord, Paul Meurer, Knut Hofland

Department of Scandinavian languages and literature, University of Bergen,
Sydnesplassen 7, N-5007 Bergen
Aksis, UNIFOB, Allégaten 27, N-5007 Bergen
{knut.hofland, paul.meurer}@aksis.uib.no, kari.tenfjord@nor.uib.no

Abstract

In our paper we present the design and interface of ASK, a language learner corpus of Norwegian as a second language which contains essays collected from language tests on two different proficiency levels as well as personal data from the test takers. In addition, the corpus also contains texts and relevant personal data from native Norwegians as control data.

The texts as well as the personal data are marked up in XML according to the TEI Guidelines. In order to be able to classify "errors" in the texts, we have introduced new attributes to the TEI *corr* and *sic* tags. For each error tag, a correct form is also in the text annotation. Finally, we employ an automatic tagger developed for standard Norwegian, the "Oslo-Bergen Tagger", together with a facility for manual tag correction. As corpus query system, we are using the Corpus Workbench developed at the University of Stuttgart together with a web search interface developed at Aksis, University of Bergen. The system allows for searching for combinations of words, error types, grammatical annotation and personal data.

1. Introduction

In our paper we present the design and the interface of a Norwegian learner corpus called ASK (Andrespråkskorpus = Second Language Corpus), which contains texts in Norwegian (bokmål variant) as a second language and personal data about the learners, as well as a control corpus of texts written by native Norwegians and personal data of the informants.

2. Interdisciplinarity

The ASK corpus is an interdisciplinary project involving the *Norwegian Language Test*, the institution responsible for the official language tests for immigrants in Norway, the *Department of Culture, Language and Information Technology (Aksis)* which has the language resource competence of vital importance for establishing this electronic corpus, and the *Department of Scandinavian language and litterature* which is responsible for the second language research competence. This interdisciplinary approach is in accordance with what Granger (2002:28) recommends for future corpus design and research.

3. The Texts

The texts are written essays from two different tests of Norwegian as a second language measuring language performance at two different levels (compared to the level description given in the Common European Framework of Reference for Languages: level B1, Threshold level, and level B2, Vantage level). The basic criterion for selecting texts for the corpus is the mother tongue of the learner. The languages chosen are German, Dutch, English, Spanish, Russian, Polish, Bosnian-Croatian-Serbian, Albanian, Vietnamese and Somali. The corpus will contain 100 texts on each test level from those language groups. There are however problems in obtaining as many as 100 texts on each level for two of the language groups.

4. The Personal Data

Among the personal data included are country of origin, mother tongue, age, sex, education, duration of stay in Norway, the extent of formal instruction received, degree of contact with native Norwegians etc. To be in compliance with the requirements of the Norwegian Data Inspectorate, ASK has to make sure that the learner's identity may not be deducible from the texts or personal data. Therefore, names, places and dates (among others) had to be anonymized.

5. The Control Corpus

In addition to the texts and personal data from language learners of Norwegian, we have been collecting data from native Norwegians writing comparable essays and supply the relevant personal data. At the end, we will have 100 texts on each level.

6. Learner Language and Second Language Aquisition Studies

The main aim of our language learner corpus is to create an electronic searchable data base which may function as a tool for doing research on second language acquisition. The corpus represents a novel opportunity to do quantitative research on much larger samples than what has been possible previously. It is also a possible tool for more explorative studies or for generating hypotheses which can be tested either on our corpus or on other data sources. A language learner corpus can be designed in many different ways; the design may be guided by special research interests or by accessibility of example data. For us, the archive of the Norwegian Language Test was a great opportunity to easily collect a large amount of homogeneous data, both textual and personal. The informants' mother tongue (L1) was our basic criterion for selecting texts for the corpus; a second criterion was typological variation between the different language groups. Thus our corpus design is to a certain extent theoretically guided, that is, guided by a research interest in second language acquisition studies in general; on the other hand, the influence of the research interests of the Norwegian research community in particular, namely the question of the L1 (Norwegian) influence on the acquisition process, cannot be ignored. But the corpus annotation itself is theory-neutral.

We use the term "error codes" in our annotation. This term is a technical one; it only refers to differences between the learner language and the standard Norwegian written language norm. It must not be interpreted as a theoretical stand in relation to what are the inherent properties of the learner languages.

7. The "Error Codes"

The texts and the personal data are marked up in XML according to the TEI Guidelines. To be able to classify errors in the text we introduced three new attributes to the TEI *corr* and *sic* tags (see below). For each error tag a correct form is also annotated in the text. During the process of developing the error tag system, we arrived at the conclusion that it was best to use a very simple set of tags in order to avoid inconsistencies in the error coding, as well as to avoid that the coding involves learner language analysis. To compensate for the simple coding system, the texts are grammatically tagged using an automatic tagger developed for standard Norwegian, the "Oslo-Bergen tagger".

The combination of general TEI tags, specially developed error attributes and the automatic grammatical tagger provides a corpus with reliable tagging and very flexible querying possibilities when the corpus is put into a query system.

The coding categories we have developed in ASK can be divided into five types. They are based on differences between the language learner texts and a possible reconstruction of the texts in accordance with target language norms:

Lexical codes:

W	wrong word
ORT	orthographic error
PART	overcompounding
SPL	oversplitting
DER	deviant derivational affix used
CAP	deviant letter case (upper/lower)
FL	Non-Norwegian word

Morphological codes:

F	deviant selection of morphosyntactic category
INFL	deviant paradigm selection, but interpreted to be in accordance with the morphosyntactical form in Norwegian

Syntactical codes:

M	word or phrase missing
R	word or phrase redundant
O	word or phrase order

The deviation category O has the following subcategories:

INV	non-application of subject/verb inversion
OINV	application of subject/verb inversion in in-appropriate contexts
MCA	incorrect position for main clause adverbial
SCA	incorrect position for subsidiary clause adverbial

Punctuation codes:

PUNC	wrong selection of punctuation mark
PUNCM	punctuation mark missing
PUNCR	punctuation mark redundant

Unidentified error:

X	impossible to interpret the writer's intention with the passage)

The coding categories F, CAP and PUNC have the following subcategories:

AGR	"agreement errors," i.e. errors following logically from, and triggered by, previous errors, the agreement itself being in accordance with the target language norm

8. Preparation of Texts

The hand written essays are keyed in with some basic mark-up using a standard XML editor, Oxygen. In the next phase the error tagging takes place. We use a stripped down version of the DTD to minimize the list of pop-up suggestions for tag names and attribute names in Oxygen. We use the *sic* tag for marking up errors and have added two attributes to the TEI version of this tag. The new attribute *type* holds the main error categories. For some categories we use a second attribute named *desc* to encode the subcategory. The correct form of the word or the phrase is given in the standard TEI *corr* attribute. *Sic* tags can be used recursively to mark up more than one error in a word or phrase. Oxygen is also used to validate the XML file.

To facilitate proofreading we have devised a set of transformations of the XML file to a presentation format (in HTML) which is viewed in a standard web browser. These transformations are done by running XSLT scripts at the server.

Another tool helps checking the consistency of the use of error codes in the texts. This tool is a special web based concordance where it is possible to select a subset of the files by giving mother tongue, date when the text was made, name of person doing the mark-up etc. together with a word or the name of a error type. From lines in this concordance a link is activated to open the actual text file in the XML editor.

9. The Coding Procedure

One of the main challenges in doing analysis on learner language is to interpret the text and decide what the learner intended to express. There may be different alternative reconstructions of an error, and sometimes a thorough reading of the whole text is necessary to decide what reconstruction is the most reasonable. But this decision is only a proposal which may help the researcher who uses our corpus to study second language acquisition. Hence the ability to view parallel sentences (see below) is of special interest both for those doing the correction annotation and for researchers using the corpus for text analysis, since it displays a synopsis of original and reconstructed text in a user friendly way.

10. The System Architecture

The ASK corpus system is designed as a client-server application with a web-based user interface.

As underlying corpus query system we are using Corpus Workbench (CWB), a corpus engine developed at IMS (University of Stuttgart) (Christ 1994), whereas the remaining parts of the system are developed at Aksis (University of Bergen) and implemented in (Allegro) Common Lisp. Common Lisp and CWB communicate via CWB's socket protocol (Cqi); web pages are generated as XML and converted to HTML on the server side using XSLT style sheets.

When a text (as XML file) is added to the corpus system, several derived files are generated: a grammatically tagged version of the text, in which the grammatical annotations (including sentence boundaries) are added as additional XML elements (see below for details); a corrected version of the text; and a grammatically tagged corrected version. (In addition, CWB input files suitable for index building are generated.) The corrected version is constructed by (recursively) replacing words or phrases contained in *sic* elements with the content of the *sic*'s *corr* attribute (but keeping the error codes). It is straight-forward to grammatically tag the corrected version since it is supposed to represent standard Norwegian. CWB indexes are built from both the original (annotated) and the corrected versions and serve as input for CWB's sentence alignment algorithm; as a result, the original and the corrected texts are searchable as parallel corpora.

Among the attributes indexed are the obvious ones: word, lemma, morphosyntactic tags, error codes, document id and relevant information from the document header, but in addition, we also index the byte offsets of the occurrences of the indexed word (and the elements it is contained in) in all four of the previously described files. Indexing those file positions makes it easy to link a hit in a corpus search to its (narrower or wider) contexts in any of the four files.

10.1. Tagging Erroneous Text with a Tagger for the Standard Language

In general, it is problematic to use a tagger written for the standard language on learners' texts with their high frequency of orthographic, morphological and syntactic errors. However, the tagger we are using (the Oslo-Bergen tagger) is based on the Constraint Grammar formalism and as such it is rather robust; it does not simply give up on ungrammatical input, but rather returns to a large extent acceptable output, although the error rate will be higher and the degree of disambiguation lower than on standardized input.

(It should be noted that although the Oslo-Bergen tagger annotates both on the morphological (part of speech, morphosyntactic features) and on the syntactic level (syntactic functions like subj, obj, finite verb, pp etc., and dependent-head relations), we largely disregard the syntactic annotations since they are less reliable than the morphological tags.)

We have implemented a couple of strategies to improve the quality of the grammatical tagging and to make its shortcomings less severe.

10.2. Correction of Orthographic Errors

Among the errors categorized in the ASK project, the most problematic ones, from the tagger's point of view, are orthographic errors (which in general are tagged as unknown words). But since orthographic corrections are provided by the annotators in the *corr* attribute, we simply hand those to the tagger instead of the original words. Thus, we end up with the original erroneous words annotated with the tags of their corrections. This leads to a twofold gain: on one hand, the erroneous words themselves are searchable by their (intended) morphological features, and on the other hand, the rules of the CG tagger see sensible context when disambiguating readings of neighboring words.

10.3. Manual Correction of Other Error Types

Obviously, not all error types lend themselves to such a straightforward automatical treatment. Therefore, we have implemented a mechanism and an interface for manual correction of tagging errors.

In the interface, on can go sequentially through the sentences matching a given corpus query and correct their tagging in several ways:

– non-fully disambiguated words can be disambiguated further;

– wrongly disambiguated words can be disambiguated manually starting from their full set of readings;

– words missing from the tagger lexicon (and thus either wrongly analyzed as a compound or tagged as unknown) can be added, together with their morphosyntax; in this case, the whole sentence should be tagged anew, the tagger then finds those words in the lexicon;

– wrongly split compounds (lexical error SPL) can be joined.

Once a sentence has been edited, all relevant files (i.e. the derived XML file containing the grammatical annotations and the CWB input files for index generation) are patched with data from the edited sentence. CWB queries will reflect the grammatical annotation of the edited files only after index regeneration (which is reasonably fast), but otherwise, the CWB index and the files are still in sync.

10.4. Parallel Corpus of Tagged Corrected Texts

In addition to the original texts, we also tag the corrected texts, whose automatic tagging is very reliable since the corrected texts represent standard Norwegian, the language the tagger grammar was designed for. Both corpora are linked together as (CWB) parallel corpora and thus can be searched in parallel. Query results, too, can be displayed as listings of sentence pairs.

10.5. Querying and Results Display

We have implemented two querying modes in the system: a menu-driven interface for composing simple queries, and a textual "expert" mode where queries can be formulated in CWB's powerful query language.

Search results can be displayed either as traditional KWIC-konkordances, as pairs of matching sentences from the original and the corrected corpus, together with relevant attributes (each sentence containing one search hit), and as sentences visualized using XSLT style sheets that highlight different aspects of the text.

In addition, collocations and various types of statistical information can be generated, although the possibilities are still rather limited and need improvement.

11. Applicability of the System to Other Languages

Although the ASK system has been developed with Norwegian learner texts in mind, it is only the tagger actually used in our system (the Oslo-Bergen tagger) which ties it to Norwegian. The system can be used without a tagger and is then easily adapted to other languages. We have in cooperation with the University of Ljubljana, Department for Slovenistics, developed a tiny learner corpus for Slovene as a proof of concept. On the other hand, the system benefits much from the possibility to semi-automatically supply grammatical information, and it would be a feasible task to include taggers for other languages if needed.

12. References

Christ, Oli (1994). *A modular and flexible architecture for an integrated corpus query system. COMPLEX'94, Budapest.*

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge Press.

Granger, Sylviane (2002). A Bird's-eye view of learner corpus research. In S. Granger, J. Hung and S. Petch-Tyson (eds.), *Computer Learner Corpora, second Language Acquisition and Foreign Language Teaching.* John Benjamins, 3–33.

Tenfjord, Kari (2004). ASK – A Computer Learner Corpus. In Peter Juel Henriksen (ed.), *Call for the Nordic Languages. Tools and Methods for Computer Assisted Language Learning.* Copenhagen Studies in Language 30, Samfundslitteratur, 147-158.

[IMS Corpus Worbench]. *http://www.ims.uni-stuttgart.de /projekte/CorpusWorkbench/*

[Oslo-Bergen Tagger]. *http://decentius.aksis.uib.no/cl/cgp /cgp/obt.html*