CALL for the Nordic Languages

COPENHAGEN STUDIES IN LANGUAGE

30

Samfundslitteratur

ASK — A Computer Learner Corpus[1]

Kari Tenfjord
University of Bergen
tenfjord@lili.uib.no

Abstract (Norwegian)
Artikkelen presenterer et prosjekt under arbeid, etableringen av et elektronisk, søkbart innlærerkorpus med norsk som andrespråk som kan koble språklige og personlige data. Målet er at korpuset kan tjene som en kilde for forskning på andrespråkstilegnelse, men det har også potensiale som et CALL-instrument. Prosjektet er tverrfaglig og involverer tre ulike miljø: testing, andrespråksforskning og språkressurser. ASK-korpuset har en demonstrator på følgende adresse:
`http://decentius.aksis.uib.no/corpus/askdemohome.html`

Abstract (English)
This article presents an ongoing project of establishing an electronic, searchable learner corpus of Norwegian as a second language with links between linguistic data and personal data. The aim of the project is that the corpus can serve as a resource for second language acquisition research, but it has also potential qualities as a CALL instrument. The project involves competence from three different milieus: language testing, second language research, and language resources. A demonstrator of the ASK corpus is to be found on:
`http://decentius.aksis.uib.no/corpus/askdemohome.html`

0. Introduction

The ASK project is in the process of establishing an electronic, searchable corpus of Norwegian as a second language with links between linguistic data and personal data, which can serve as a resource for second language acquisition research. The corpus also has potential qualities as a CALL instrument. In this article, we will present the overall design and the interface

[1] ASK is an abbreviation of the Norwegian name "Andrespråkskorpus" (Second Language Corpus).

of this language learner corpus and give a brief discussion of some theoretical problems concerned with data collected from second language learners.

1. The Goal of the ASK Project

The main aim of building this corpus is to strengthen the possibility of doing research on second language acquisition (SLA). The corpus will make it possible to test hypotheses generated from earlier studies in Norwegian as a second language, as well as more general hypotheses of SLA. The corpus may also be a rich source for developing new hypotheses of lexical, grammatical and textual features of written SLA, as well as hypotheses of individual and external factors influencing the acquisition process.

2. Interlanguage

The field of SLA research in Norway started out in the early 1980s. Since that time, we have been paying a lot of attention to the question of what kind of language it is that the learners produce, and consequently, what the characteristics of our object of study are. When Selinker (1972) introduced the term "Interlanguage" for the language a learner produces while he/she is in the process of acquiring a second language, he regarded the learning process as a creative process where the learner creates his/her individual language, which has its own rules, and which differs in part from both L1 and the target language. In accordance with these ideas, it is controversial to speak about "error" in learner languages. Still, the term error is widely used. And the phenomenon it refers to is important for understanding second language acquisition. Corder (1967), in his seminar paper "The Significance of Learner's Errors", states that errors "provide to the researcher evidence of how language is learned or acquired, what strategies or procedures the learner is employing in his discovery of the language". And perhaps the most important aspect, according to Corder (1967), is that they are indispensable to the learner, because making errors "is a way the learner has of testing his hypotheses about the nature of the language he is learning". This positive view of errors in learner languages is important to bear in mind, but we must also pay attention to the fact that if errors are the only source for an understanding of the acquisition process, they are an incomplete source. Concerning the terminology, it is preferable to use terms that do not support the rather negative impression that learner languages are erroneous varieties of the target language, and several suggestions have been put forth for alternative, more positive, terms for the phenomenon, for example: deviation, learner solution and transitional form.

There is, however, a theoretical problem connected with errors in the analysis of interlanguages that is more basic than the problem of choosing the most suitable term for the phenomenon under discussion. This has to do with the "correction" or the "translation" of an error into the target language. Lattey (1982) asks the following question: "What is the 'same thing' in interlinguistic comparison?" When a structure in the interlanguage is deviant or different from the target language norm, it is not obvious what should be regarded as 'the same thing' in the interlanguage and in the target language. You cannot always be sure of the intended meaning in an interlanguage structure. I will come back to this problem when presenting the coding procedure in the ASK project.

3. Interdisciplinarity

There are three different milieus involved in the ASK project. The *Norwegian Language Test* (Norsk språktest) is the institution that is responsible for the two official language tests for migrants in Norway. The written responses to the tests have been collected together with personal data about the test takers. The *Department of Culture, Language and Information Technology* (Aksis) has language resource competence that is of vital importance for establishing an electronic corpus. Researchers at the *Department of Scandinavian language and literature* hold the second language research competence. This interdisciplinarity is in accordance with what Granger (2002:28) recommends for future corpus design and research.

4. The Data

The data are of two different kinds, textual data, which give information about language proficiency, and data concerning personal variables, such as mother tongue, sex, age, and age at arrival in Norway, which make it possible to do statistical analysis of variables that affect the acquisition process. SLA research has been based on various types of sources: written or spoken language production, discourse or narratives of some kind, introspective data, elicited data, and experimental data. In spite of this kind of diversity, the research data are often too limited or too heterogeneous to base generalizations on. The data in the ASK corpus will, to a certain degree, represent new possibilities for research on Norwegian SLA, since it contains both written and personal data from a high number of informants. It could, for example, serve as a new and rich source for doing quantitative studies.

4.1 The Textual Data

The textual data consist of essays collected from the archive of the Norwegian language test, which are written responses from migrants who have taken a test in the Norwegian language. From this archive, we have collected data from test takers. The written performance of the test takers have been assessed to be at or above certain levels of proficiency, for *Språkprøven i norsk* (here: Test 1) at B1 (Threshold level) and for *Test i norsk – høyere nivå* (here: Test 2) at B2 (Vantage level) in accordance with the Common European Framework of Reference for Languages.

The texts are essays and may be expository texts, narrative texts, or texts that contain elements from both types. The essays collected from Test 1 contain about 240 words averagely, while the average of Test 2 is about 450 words. The corpus will contain 1000 essays from each test level with a total of about 600.000 words.

4.2 Criteria for Text Selection

The basic criterion for selecting texts for the corpus is the mother tongue of the learner. One of the variables, which have been most widely discussed in the area of SLA, is whether the mother tongue (L1) has any effect on second language acquisition, and if so, in what way it affects language learning. Today, there appears to be a widespread agreement among SLA researchers that L1 affects the learning process in some way, but the field of SLA is facing methodological problems in testing hypotheses concerning the role of the mother tongue. Isolating the factor "mother tongue" from other factors, which influence language learning, is perhaps not possible. The most promising methodological approach today is to do statistical analysis of the language produced by learners with different mother tongues while keeping other factors alike for the learners. This methodology will be possible to use when doing research based on the ASK corpus.

A second criterion for text selection is that of variation in language typology. This criterion competed, however, with another: the number of texts from learners of different L1's. We decided that, in order to have enough data for statistical analysis of L1 influence on SLA, we need to have 100 texts written by learners of the same L1. This has as a consequence a somewhat limited typological variation. It has not been possible to find as many as a hundred texts from the two different language tests in all the source languages that we would have chosen, if we could choose freely. The languages chosen are the following: German, Dutch, English, Spanish, Russian, Polish, Serbo-Croatian, Albanian, Vietnamese and Somali[2].

4.3 The Personal Data

In connection with taking the tests, the learners fill out a form with personal information that may influence the language learning process, for example, mother tongue, country of origin, sex, age, what kind of Norwegian courses they have taken, education, amount of contact with native Norwegians, etc. The personal data will be processed in accordance with the quite strict Norwegian law that protects people from any violation of their right to privacy. The Norwegian Data Inspectorate has given instruction on the procedure ASK has to undertake. There must be no way that *a* learner's text or personal data may form the basis of identification of the individual. When transcribing and coding the texts, we have to insure the learner's anonymity, and for this reason, we have developed special codes for personal information that otherwise might lead to identification of the learner.

[2] There are still problems in obtaining as many as 100 texts on each level for Vietnamese and Somali learners.

4.4 The Control Corpus

We are now in the process of collecting both textual and personal data from native Norwegians. The aim is that 100 informants will take each of the two tests. The natives must, to some degree, reflect the individual variation among the migrants. We have, therefore, chosen informants from groups where we expect a variation in age, sex, and educational background (for example choirs and sports clubs).

5. Explicit Design Criteria

One of the explicit design criteria is that ASK will not be a random collection of heterogeneous learner data. The archive of the Norwegian Language Test is a unique source for building a corpus. In addition to data selection criteria, the texts are collected from the same test situation for every test taker. They have the same amount of time for taking the test, they take the test under the same conditions, the tests are scored by sensors with the same kind of training, and the collection of personal data has been done in connection with the test situation. The learning context is, of course, not quite the same for every individual, and the individuals differ in many ways. But since we have coded information of important personal variables, we can control differences both in learning context and in learners' background. We are convinced that, so far, the corpus passes the test that Granger (2002:9) put forth: "The usefulness of a learner corpus is directly proportional to the care that has been exerted in controlling and encoding the variables".

6. Tagging

The texts as well as the personal data are marked up in XML according to the TEI Guidelines (Text Encoding Initiative). 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 grammatical tagger developed for standard Norwegian, "The Oslo-Bergen tagger".

6.1 Error Tags and Corrections

Error tagging has been established as a standard procedure in learner language corpora, a consequence of the fact that this kind of corpora needs its own techniques:

> *"... computer learner corpora quite naturally call for their own techniques of analysis, [...] such as error tagging, which are specially designed to cater for the anomalous nature of learner language." Granger (2002:18)*

The techniques that are being developed must, however, be in accordance with modern theories of SLA. The theoretical basis for using the term "error" and "error code" is not quite clear, nor have possible negative consequences of this terminology been extensively discussed. This has perhaps to do with the fact that the field of language learner corpora has its roots in corpus linguistics, and it may be an example of what Granger (2002:28) has in mind, when she questions whether corpus linguistics and SLA specialists have met in learner corpus research. It is, of course, the responsibility of the SLA research field to insure that analytical categories, for example, error codes are in accordance with the theoretical foundation of SLA.

So far we have chosen to use the terms "error" and "error coding" in the ASK project. We will, however, emphasize that these are technical terms with no theoretical foundation in any theory of learner language.

Now I turn to the most basic problem when dealing with interlanguage texts: how to decide on "error" type and how to "correct errors" or produce a normalized Norwegian version of the deviant parts of the interlanguage. Some of the error coding and corrections are trivial processes. But in many cases, we make decisions that are based on subjective interpretation of the language in question. This is an unavoidable problem in SLA research, and it will exist independently of the research context, outside or inside the field of learner corpora. In order to cope with this problem, we have:

a) instructed those who do the "error coding" to choose the correction that demands the fewest changes of the learner text

b) developed a relatively simple set of error codes (the more complex the error coding system, the greater is the chance of inconsistency in the coding)
c) developed an error coding manual which contains, in addition to the error codes, a collection of examples
d) we will carry out reliability tests of the coding. In spite of these procedures, we will, nevertheless, not be able to avoid the problem of subjectivity in the interpretation of the texts, and researchers who want to use the ASK corpus will need to be aware of and critical to the coding of error types and their corrections

The error codes we have developed in ASK can be divided into five types:

a) Lexical:
W (wrong word chosen)
ORT (orthographic error)
PART (deviant partition)
SPL (splitting of compounds)
DER (deviant affixation)
CAP (deviant capitilization)
FL (word from other languages than Norwegian)

b) Morphological:
F (deviant morphosyntactical form)
INFL (deviant formation of a morphosyntactical form)

c) Syntactical:
M (word missing)
R (redundancy of word or phrase leading to an ungrammatical or un-idiomatic structure)
O (deviant word order)
INV (inversion missing)
OINV (inversion in structures which do not require inversion)
MCA (wrong order of sentence adverbial in main clause)
SCA (wrong order of sentence adverbial in subordinate clause)

d) Punctuation:

PUNC (wrong punctuation)

PUNCM (punctuation is missing)

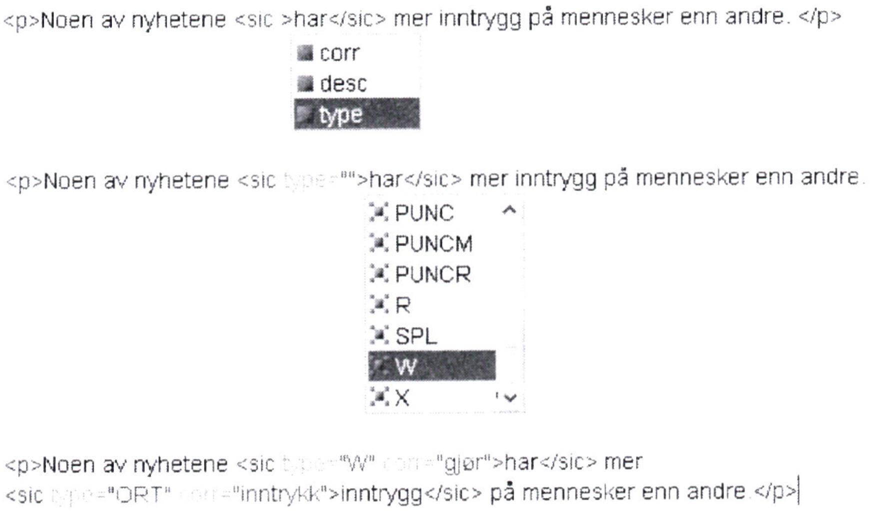
PUNCR (punctuation is redundant)

e) Unidentified error:

X (interpretation is impossible)

The error types F, CAP and PUNC have the following sub type:

AGR (error caused by a previous error)



```
<p>Noen av nyhetene <sic >har</sic> mer inntrygg på mennesker enn andre. </p>
```

```
<p>Noen av nyhetene <sic type="">har</sic> mer inntrygg på mennesker enn andre.
```

```
<p>Noen av nyhetene <sic type="W" corr="gjør">har</sic> mer
<sic type="ORT" corr="inntrykk">inntrygg</sic> på mennesker enn andre.</p>
```

Figure 1: The error coding editor

6.2 Automatic Tagging

"The Oslo-Bergen Tagger", an automatic tagger developed for standard Norwegian, is used in addition to the manual coding of errors. In general, it is problematic to use a tagger written for a standard language on learners' texts with their high frequency of orthographic, morphological, and syntactic deviations. Since the tagger works on a lexicon, the automatic tagger in our corpus works on the corpus corrected for orthographic errors (the error code ORT), and the automatic tagging can also be controlled and edited manually by a function "search and edit".

7. The Query System

The combination of general TEI tags, specially developed error attributes, and the automatic grammatical tagger has the potentials of a corpus with reliable tagging and very flexible querying possibilities. As corpus query system, we are using Corpus Workbench, a corpus engine developed at IMS (University of Stuttgart) together with a web search interface developed at Aksis (University of Bergen). The system allows searching for combinations of words, error types, grammatical annotation, and personal data.

8. Search Results

Search results can be displayed either as traditional KWIC-concordances, 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, which are visualized by using user definable (XSLT) style sheets that highlight different aspects of the text. In addition to this, collocations and various types of statistical information can be generated.

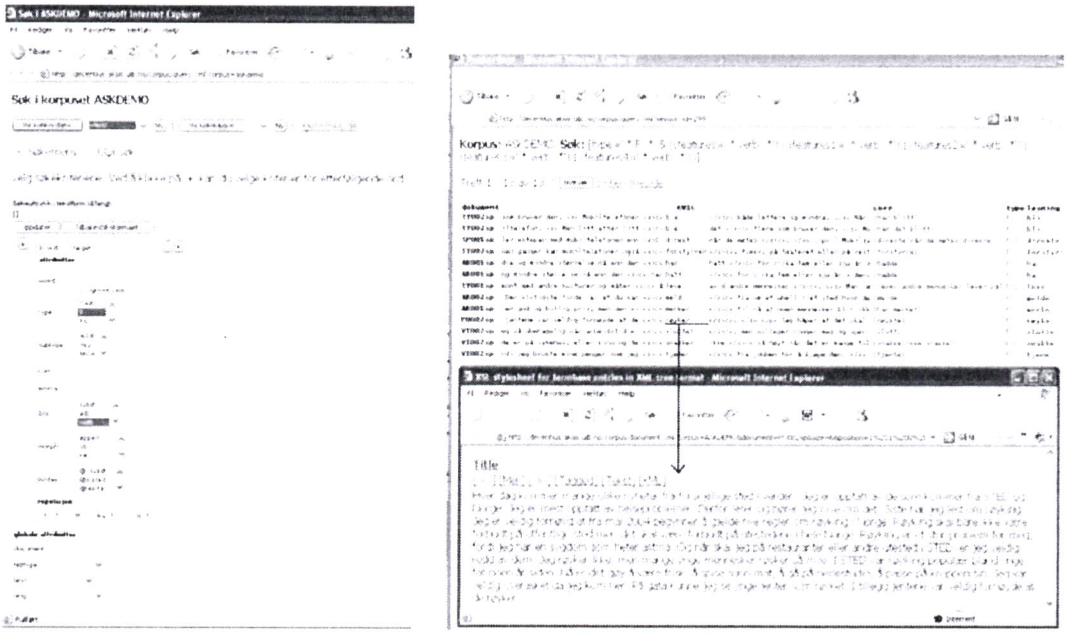

Søk i korpuset ASKDEMO - Microsoft Internet Explorer
Søk i korpuset ASKDEMO
Korpus:
Søk:
XSL stylesheet for lexeme articles in XML tree format - Microsoft Internet Explorer
Title

Figure 2: Concordance where the search is error type F combined with the word class 'verb'

11. Some Concluding Remarks

When evaluating the ASK corpus, one important success criterion will be that researchers, students, and teachers in the SL classroom use the corpus. It is important that the search interface is transparent, and that the user manual is pedagogical. So, please visit the ASK project and the demonstrator of the search system, and we are, of course, happy for any comments:

`http://spraktek.aksis.uib.no/projects/ask`

`http://decentius.aksis.uib.no/corpus/askdemo-home.html`

References

Corder, S. P. 1967. The significance of learner's errors. *International Review of Applied Linguistics* 5. 161-169.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge University Press.

Granger, S. 2002. A bird's-eye view of learner corpus research. In *Computer learner Corpora, second Language Acquisition and Foreign Language teaching.* Granger, S. & Hung, J. & Petch-Tyson, S. (eds). Benjamins, J. 3-33.

Hunston, S. 2002. *Corpora in Applied Linguistics.* Cambridge University Press.

Selinker, L. 1972. interlanguage. *International Review of Applied Linguistics* 10. 209-231.

9. ASK as Database and Methodological Instrument for SLA Research

ASK is still an ongoing project. It will be finished by the end of 2005, and there is still a lot of work left to do. But so far, it is promising as a research tool that creates new possibilities for SLA research. It is the interdisciplinary nature of the project that creates the new possibilities as Granger (2002:4) notes: "The area of linguistic enquiry known as learner corpus research, ... has created an important link between the two previously disparate fields of corpus linguistics and foreign/second language research." The field of learner corpora is a relatively new branch of corpus linguistic, and it may serve as a powerful methodology for research in the field of SLA.

10. ASK as a Potentially CALL Instrument

ASK is both a learner corpus and a parallel corpus in the sense that it is possible to search in the learners' original text as well as in a "translated" or corrected text. Since all the texts are coded for "errors" and each error code contain a proposed reconstruction or translation, it is possible to compare a learner text with a text translated in accordance with a Norwegian norm for written texts. We will also have the possibility to search and compare learner texts with the texts written of native Norwegians (the control corpus). Hunston (2002:15) describes a parallel corpus in the following way: "Two (or more) corpora in different languages, each containing texts that have been translated from one language into the other." The ASK corpus contains three parallel corpora:

1) interlanguage texts
2) reconstructed interlanguage texts
3) texts written in Norwegian by native Norwegians.

The corpus may, thus, be a rich resource for doing different kinds of comparison tasks. Comparing language varieties is in concordance with modern teaching programs, i.e. it is methodology of "consciousness raising" that may help the student "noticing the gap".

»Copenhagen Studies in Language« carries studies in both language for general purposes and language for special purposes (LSP). Its scope covers grammar, semantics, pragmatics, text linguistics and translation, from a theoretical as well as an applied perspective.

It is the editors' policy to bring out thematic volumes. The present volume on computer assisted language learning is the latest in a series of volumes following this policy. While based at the Copenhagen Business School, the series is open to contributions from linguists in other institutions, in Denmark and abroad.

STUDIES FROM THE FACULTY OF LANGUAGE, COMMUNICATION AND CULTURAL STUDIES AT THE COPENHAGEN BUSINESS SCHOOL

ISBN 87-593-1176-2

9 788759 311769

DISTRIBUTED BY
SAMFUNDSLITTERATUR
ROSENØRNS ALLÉ 9
DK-1970 FREDERIKSBERG C

GENERAL EDITOR:
GYDE HANSEN
THIS VOLUME IS EDITED BY:
PETER JUEL HENRICHSEN