Cecilie Carlsen and Eli Moe (eds.)

A Human Touch to Language Testing

A collection of essays in honour of
Reidun Oanæs Andersen
on the occasion of her retirement
June 2007

Kari Tenfjord

"ASK and you will find what you seek"

One of Reidun Oanæs Andersen's visions was to establish an electronic corpus of texts and personal data based on the archives at Norwegian Language Test (NLT= the institution which Reidun has been in charge of). Her idea was that the written texts produced by candidates taking the official tests in Norwegian as a second language would be a rich source for research in the field of second language acquisition (SLA) and language testing. It is therefore a pleasure to present ASK[1], a language learner corpus of Norwegian as a second language in this book. The ASK project was initiated by Reidun, and thanks to her mild pressure, SLA researchers at the University of Bergen started to build this corpus in 2002. In this article I will give a short presentation of the design and interface of the corpus, some theoretical challenges in building a language learner corpus, as well as its potential for research.

1. Introduction

ASK is an electronically searchable corpus of Norwegian as a second language linking linguistic and personal data, which can serve as a research tool and database for SLA and language test research. Moreover, the corpus has potential qualities as a computer aided language learning (CALL) instrument.

The main aim of building this corpus was to facilitate research on second language acquisition. The corpus gives us the potential to test specific hypotheses generated from earlier studies in Norwegian as a second language as well as more general hypotheses of SLA. The corpus may be a source for generating new hypotheses of lexical, grammatical and textual features of written SLA, and for explorative and descriptive studies of the proficiency levels that are represented in the corpus. There is also a novel possibility to conduct statistical analyses of correlation between linguistic features and personal variables. This makes the corpus a rich source for

exploring individual and external factors influencing the acquisition process and language use in a test situation.

There are three stakeholders involved in the ASK project. Firstly, there is *Norwegian Language Test* (Norsk språktest), which develops and validates the official language tests for migrants in Norway. The written responses of the test have been collected together with personal data about the test takers. Secondly, the *Department of Culture, Language and Information Technology* (Aksis) has the language resource competence that is of vital importance for establishing an electronic corpus. And thirdly, researchers at the *Department of Scandinavian Language and Literature* hold the second language research competence. This kind of interdisciplinarity is of great value in corpus building and, as I see it, this is in accordance with Granger's (2002:28) recommendations for corpus design and research.

2. The design and interface of ASK[2]

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 at Norwegian Language Test represents a rich source from which we could 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 contrastive studies in general, and in the question of L1 influence on the acquisition process in particular. Yet, the corpus annotation itself is theory-neutral. We have chosen to use the term 'error code' in our annotation. This term is a technical one; it refers only to differences between the learner language and the standard norm of written Norwegian (bokmål). It must not be interpreted as a theoretical stance regarding what the inherent properties of the learner languages are.

2.1. The database

ASK contains essays collected from language tests at two different proficiency levels (compared to the level description given in the *Common European Framework of Reference for Languages* it is much the same as level B1, Threshold level, and level B2, Vantage level). In addition to the texts the corpus contains personal data from the test takers.

As already mentioned the basic criterion for selecting texts for the corpus was learners' L1s, and the language groups chosen were German, Dutch, English, Spanish, Russian, Polish, Bosnian-Croatian-Serbian, Albanian, Vietnamese and Somali.

Among the personal data included are L1, country of origin, 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 makes sure that the learners' identity may not be deduced from the texts or personal data. Therefore, names, places and dates (among others) had to be anonymised.

In addition to the texts and personal data from language learners of Norwegian, we have collected personal data and comparable essays from native speakers of Norwegian. This control group is stratified to match the groups of learners as far as possible. The native speakers have been selected from groups (such as choirs) where we expected a variation in age, sex, educational background and occupation, for instance.

2.2. 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*[3] in the texts 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 avoiding that the coding involved 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' (see below).

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

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

List of error categories used in ASK	
1. 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 **2. Morphological codes:** F deviant selection of morphosyntactic category INFL deviant paradigm selection, but interpreted to be in accordance with the morphosyntactical category in Norwegian **3. Syntactical codes:** M word or phrase missing R word or phrase redundant O deviant word or phrase order *The category O has the following subcategories*: INV non-application of subject verb inversion OINV application of subject verb inversion in inappropriate contexts MCA incorrect position for main clause adverbial SCA incorrect position for subsidiary clause adverbial	**4. Punctuation codes:** PUNC wrong selection of punctuation mark PUNCM punctuation mark missing PUNCR punctuation mark redundant **5. 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

Table 1. List of error categories used in ASK

The manual error coding is not a straightforward procedure. There may be different ways to reconstruct an interlanguage structure deviating from the target language. What was important to have in mind when conducting the error coding procedure was that all the texts had already been assessed to be at or above certain reasonably well-defined levels of language proficiency. For both tests, the central criterion of assessment is communicative functionality. That means that the candidates have been able to communicate the contents of their intentions sufficiently according to the descriptors relevant at each level. When coding the texts we presuppose that they express a reasonably clear, identifiable and coherent content, and that this content is intelligible and processable directly 'on line' by any native speaker with Norwegian

as their only linguistic resource. A reader of any text will, whether the text is produced by a native speaker or a language learner, interpret a text not merely upon the basis of its literal content, but also upon the expectations motivated by its contexts. Systematic analysis of learner language is therefore not required for interpretation of these texts. There have been two important principles guiding the error coding procedure: 1) select the most probable interpretation from a pragmatic point of view and, 2) select the reconstruction which deviates least from the original text. Anyway, the error codes chosen in ASK may be reconsidered by the researcher using ASK as a research tool.

The ability to view parallel sentences is of special interest both for those doing the error coding and for researchers using the corpus for text analysis, since it displays a synopsis of original and reconstructed text in a user friendly way.

3. The System Architecture

The ASK corpus system is designed as a client-server application with a web-based user interface. As the underlying corpus query system we are using Corpus Workbench (CWB), a corpus engine developed at IMS (University of Stuttgart), whereas the remaining parts of the system were developed at Aksis, University of Bergen.

When a text (as an 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 are added as additional XML elements; a corrected version of the text; and finally a grammatically tagged corrected version. 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). The original and the corrected texts are searchable as parallel corpora.

Among the attributes indexed are 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.

It is of course problematic to use a tagger written for standard Norwegian on learners' texts with their high frequency of orthographic, morphological and syntactic deviations from the target language norms. However, the tagger we used (the Oslo-Bergen tagger) is based on the Constraint Grammar (CG) formalism and as

such it is 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, however, implemented a couple of strategies to improve the quality of the grammatical tagging and to make its shortcomings less severe.

Among the errors categorized in ASK the most problematic ones, from the tagger's point of view, are orthographic errors (which are generally 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 neighbouring words.

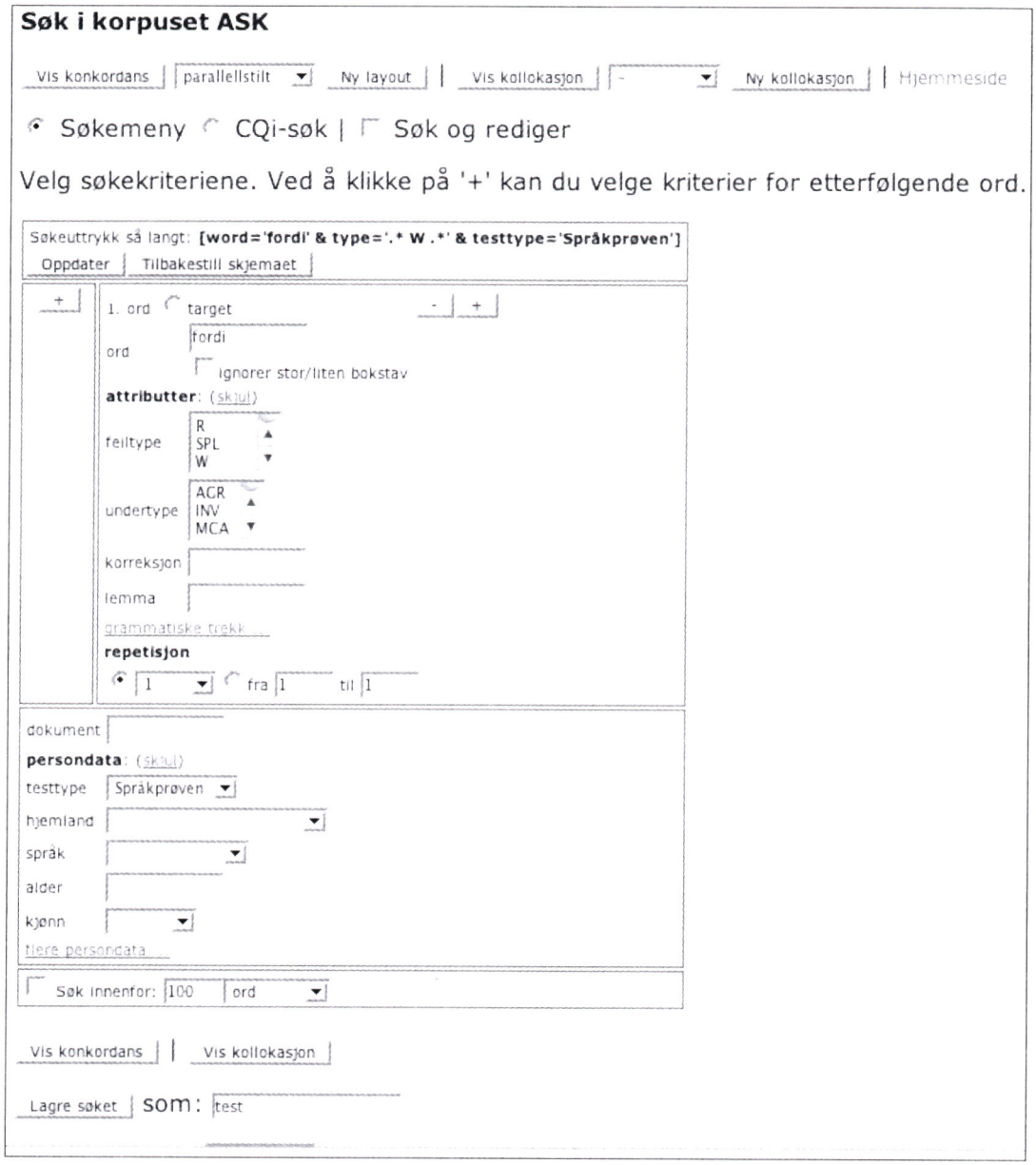
3.1. Querying and Result Display

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. Picture 1 shows the interface, and a search for the word 'fordi', when the word represents an error of the type W.

Search results can be displayed either as traditional KWIC-concordances (Picture 2), as pairs of matching sentences from the original and the corrected corpus (Picture 3), 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 (Picture 4), although the possibilities are still rather limited. There are also possibilities for generating different kinds of word lists (Picture 5).

4. Research potential

As already mentioned, the error coding is theory independent, and as a consequence ASK may serve as a research tool for researchers of different theoretical positions.



Picture 1. The interface of ASK showing a search for the word 'fordi' when 'fordi' represents the errror type W (wrong word).

Golden, Kulbrandstad and Tenfjord (2007) have just conducted a study of the history of the field of Norwegian as a second language which shows that in the core field of SLA – the study of learner language systems – the prototypical master thesis is a syntactic analysis, based on adult learners' written texts, the data are collected by the researchers themselves and the numbers of informants are between 20 to 50 people (Golden et al. 2007:19). Only by pointing to these facts it is obvious that ASK may improve and facilitate the research possibilities. Not only will the researchers save

Korpus: ASK, **Søk:** [word='fordi' & type='.* W .*' & testtype='Språkprøven']

Treff 1 - 14 av 14. | Kun ett treff per sic | KWIC | bredde: 200px | Last ned | Nytt søk | Hjemmeside

0833	:r på i midten av vinteren i Norge, **fordi** at vi gjøre sånn i Argentina, for el	W	|selv|
-0392	e slags våpner i begynnelsen <sic> **fordi** </sic> den etter hvert måtte gi seç	W PUNCM	|men|,|
0065	person som liker å skrive brev, og **fordi** er det lett for meg å svarer snart	W	|derfor|
0270	g liker veldig god Norsk natyr <sic> **fordi** </sic> har frisk luft, masse forskjel	W	|på grunn av|
-0061	ıar mørketida ca. 2 måneder <sic> **fordi** </sic> hjemme brukes mye lyss. </	W PUNCM	|derfor|,|
-0425	; kansje at det er litt rart. <s> <sic> **fordi** </sic> jeg bare er bare ANTALL år	W	|siden|
-0358	eiser så mye i våre dager er, <sic> **fordi** </sic> vi har mer penger, mer mul	W	|at|
-0391	bli. <s> Vi må gjøre det riktig <sic> **fordi** </sic> vi får ingen sjanse å gjøre d	W PUNCM	|for|,|
0816	Årsaken at vi mener sonn er <sic> **fordi** </sic> vi kan representere oss best	W	|at|
0615	ra Tyskland. <s> Det var bare <sic> **fordi** </sic> vår venner. </s> Vi har kont.	W	|på grunn av|

Picture 2. The search result of the word 'fordi' when 'fordi' represents the errror type W (wrong word) displayed as a KWIC-concordance.

Korpus: ASK, **Søk:** [word='fordi' & type='.* W .*' & testtype='Språkprøven']

Treff 1 - 14 av 14. | Kun ett treff per sic | parallellstilt | Nytt søk | Hjemmeside

Jeg er en person som liker å skrive brev, og **fordi** er det lett for meg å svarer snart til brevene.
Jeg er en person som liker å skrive brev, og derfor er det lett for meg å svare raskt på brevene.
Men komunisme partiet tok det feil, **fordi** befolkningen ville gjerne ha en regjering forandring.
Men kommunistpartiet tok feil, for befolkningen ville gjerne ha en regjeringsendring.
Vi må gjøre det riktig **fordi** vi får ingen sjanse å gjøre det igjen.
Vi må gjøre det riktig, for vi får ingen sjanse til å gjøre det igjen.
fordi jeg bare er bare ANTALL år gammel (ikke pensjonist).
siden jeg bare er bare ANTALL år gammel (ikke pensjonist).
Vi har et varmt temperament og vi er glad **fordi**.
Vi har et varmt temperament, og vi er glad for det.
Det var bare **fordi** vår venner.
Det var bare på grunn av våre venner.
Det er sikkert og jeg er glad **fordi**.
Det er sikkert, og derfor er jeg glad.
Årsaken at vi mener sonn er **fordi** vi kan representere oss best vi kan.
Årsaken til at vi mener det, er at vi vil framstå som best vi kan.
Pentagonen nektet å ha brukt dette slags våpner i begynnelsen **fordi** den etter hvert måtte gi seg.
Pentagon nektet for å ha brukt denne slags våpen i begynnelsen, men etter hvert måtte de gi seg.

Picture 3. The search result of the word 'fordi' when 'fordi' represents the error type W (wrong word) displayed as pairs of matching sentences from the original corpus and the corrected corpus which is generated from the error coding (sic) and the correction (corr).

Korpus: ASK, **Søk:** [word='fordi' & type='.* W .*' & testtype='Språkprøven'] ;

Ny kollokasjon | Nytt søk | Vis konkordans | Last ned kollokasjon

Søket ga 14 treff. Det ble funnet 7 forskjellige kollokasjoner.

	match lang	absolutt frekvens	relativ frekvens	mutual information
☐	nederlandsk	4	0.28571	-14.24329
☐	serbokroatisk	4	0.28571	-14.10074
☐	spansk	2	0.14286	-15.22219
☐	albansk	1	0.07143	-15.08257
☐	polsk	1	0.07143	-16.38382
☐	somali	1	0.07143	-14.67358
☐	tysk	1	0.07143	-16.31720

Picture 4. The search result of the collocation L1 and the error type W when the wrong word is 'fordi'.

Korpus: ASK, **Søk:** [((features='.* verb .*')

Søket ga 50014 treff. Det ble funnet 1019 forskjellige kollokasjoner.

	match lemma	absolutt frekvens	relativ frekvens	mutual information
☐	være	10476	0.20946	-1.28097
☐	ha	4420	0.08838	-1.18995
☐	kunne	2701	0.05400	-1.35915
☐	måtte	1369	0.02737	-1.34804
☐	bli	1335	0.02669	-1.63244
☐	skulle	1129	0.02257	-1.11065
☐	gå	975	0.01949	-0.92950
☐	bo	928	0.01855	-0.80647
☐	få	900	0.01799	-1.72052
☐	gjøre	863	0.01726	-1.32627

Picture 5. The 10 most frequent verbs in the texts collected from 'Språkprøven'.

time by not having to collect all the data themselves, they will also have access to a much higher number of informants than have been possible in earlier studies. In addition they will have the possibility of linking text variables with internal and external factors, and ASK may thus function as a database not only for language system studies, but also for studies of factors affecting the acquisition process. Golden et al. (2007) show that to a certain extent researchers and students have been collecting personal data from the informants, but these data have seldom been used in the studies conducted. This is probably connected with the fact that the number of informants has been too low, and it is a clear tendency in the discussions of research results in Norwegian language studies, that it is not possible to draw general conclusions due to lack of statistically significant results. The corpus is not only a source for statistical analysis, it is also a rich source for explorative and descriptive studies.

Studies have already been conducted based on preliminary versions of ASK[4], but only the future will show if the ASK corpus will improve the research situation and research possibilities in the scientific field of Norwegian as a second language.

Notes

1 ASK is acronymic for the three constituent morphemes of Norwegian AndreSpråksKorpus (SecondLanguageCorpus).
2 The presentation is partly based on Tenfjord, Meurer & Hofland 2006.
3 The question whether the practice of error recording and error coding in itself is theoretically misguided by virtue of the so-called 'comparative fallacy' argument (Bley-Vroman 1983) is discussed in Tenfjord, Hagen og Johansen 2006.
4 Master thesis Hagland (2005), Busterud (2006), Johansen (2007), doctoral lectures Golden (2005), and there are a number of works in progress, both master thesis and Ph.D. thesis.

References

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

Golden, Anne, Lise I. Kulbrandstad & Kari Tenfjord. 2007. Norsk andrespråksforskning – utviklingslinjer fra 1980 til 2005. In *Nordand* 1 2007, 5–37.

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.

Tenfjord, Kari, Paul Meurer og Knut Hofland. 2006. The ASK Corpus: a Language Learner corpus of Norwegian as a Second Language. *In Procedings from the 5th International Conference on Language Resources and Evaluation* (LREC 2006) European Language Resources Assosiation: Genova

Tenfjord, Kari, Jon Erik Hagen og Hilde Johansen. 2006. The Hows and Whys of Coding Categories in a Learner Corpus (or "How and Why an Error Tagged Learner Corpus is not IPSO FACTO one big Comparative Fallacy"). In *Rivista di Psicolinguistica Applicata (RiPLA)* VI.3.2006. Special issue on 'Interlanguage: current thought and practices'

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

[Oslo-Bergen Tagger]. http://omilia.uio.no/obt/

Reidun Oanæs Andersen has retired from the position of director of Norsk språktest (the University of Bergen/Folkeuniversitetet), a post which she has held for the last two decades. Through her work, Reidun has made an enormous contribution to the professionalisation of the field of language testing in Norway. Reidun's professionalism, as well as her openheartedness, and amazing energy, care and concern for everyone around her, has made her a treasured colleague and friend to many, both within and outside Norway.

This volume is entitled A Human Touch to Language Testing, and the contributions address language assessment issues, primarily in a European or Nordic context. The focus of the articles varies, ranging from broad, general perspectives on language assessment to specific case studies examining test validation and second language acquisition.

ISBN 978-82-7099-452-6

NOVUS PRESS

Cecilie Carlsen and Eli Moe (eds.)

A Human Touch to Language Testing