*SweDanes for CALL - A corpus and computer-based student's aid for comparison of Swedish and Danish spoken language"; in NorFA CALL Net - Demo CDROM 2005 [10 Nordic applications for computer assisted language learning; ed. by C. Hansen, H. Selsøe Sørensen and PJH; sponsored by the Nordic Research Academy, 2005*

Jens Allwood
Peter Juel Henrichsen

# SWEDANES CALL

## - A corpus and computer-based aid for comparison of Swedish and Danish

This subproject is based on a corpus for spoken Danish BySoc (BySociolingvistik) and a corpus for spoken Swedish GSLC (Göteborg Spoken Language Corpus), each containing 1.3 million words of transcribed spoken interaction. Both corpora are available for browsing on line, BySoc2 at http://www.id.cbs.dk/~pjuel/cgi-bin/BySoc_ID/index.cgi and GSLC at http://www.ling.gu.se/projekt/tal/index.cgi? PAGE=3.

Both corpora consist mainly of fairly informal, spoken language interaction between two or more speakers. They have the same size and the main parts were collected during the same period of time. They represent two Scandinavian languages with considerable similarities. Both corpora are transcribed according to standards that are compromises between the three purposes of (i) representing spoken language collected in naturalistic circumstances with as little interference from a researcher as possible, (ii) creating a standard which supports transcription and is both rapid and reliable and (iii) making possible the use of computerized tools for analysis. This means that both corpora are transcribed into basically orthographic word representation with spaces between words, but that the transcription standards are specially designed for *spoken* language (cf. Allwood 1998).
Neither of the two transcription standards uses any form of written punctuation.

The two corpora were collected for somewhat different purposes and this is reflected in the types of activities and speakers which are included. The BySoc corpus was originally recorded and transcribed in 1986-1990 in the project BySoc (The Copenhagen Study in Urban Sociolinguistics). It consists of so called Labovian sociolinguistic interviews or conversations with about 80 citizens of Copenhagen, representing different ages, genders and social classes. They are informal conversations. The transcriptions were made in score format, i.e. with parallel running lines for the different participants. They have been converted into text files and homogenized/standardized into the present BySoc corpus by Henrichsen (1997, 1998a, 1998b).

The GSLC (the Göteborg Spoken Language Corpus) was mainly recorded in the period 1978-2000 as part of many different projects, with the main purpose of representing many different social activities. (It does, however, also include a few recordings from the 1960:s.) The corpus contains around 20 different social activity types (for an overview of activity types, see appendix 3). It is described in Allwood (1999, 2001, 2002) and in Allwood et al (2000).

This difference in purposes, i.e. that the GSLC puts priority on representing different social activity, whereas BySoc puts priority on representing different individuals in related circumstances, means that BySoc contains a systematic variation of age, gender and social class of the interviewed speakers, while the activity type is mainly the same, i.e., sociolinguistic interview or informal conversation. In most cases this means fairly long interactions between two persons. The GSLC, on the other hand, is systematically varied with respect to social activity, the number of speakers is much larger and the characteristics of participants are not primary criteria for selection but are rather a consequence of the choice of activities, i.e. they are varied and less controlled than in BySoc. The transcriptions are also more varied in length. (For some purposes of comparison, it is therefore suitable to use a subcorpus of the GSLC, containing informal interviews and conversations more similar to BySoc.)

Different corpora exhibit differences in format, based on the tradition in which they were created and the purpose of the original transcriptions. The two corpora were both standardized prior to their comparison and are as a result of this standardization written in "two modified standard orthographies", one for Swedish and one for Danish. They share many features, but there are still some notable differences that have to be considered in doing comparative analysis of them.

In addition to the two corpora and the browsers for the corpora, there is a program for making comparisons of word frequencies for related words in Swedish and Danish.

Besides corpora, browsers and comparison program, we provide help to users by making available a list of words which are historically related (so-called cognates) and/or semantically similar in both languages, i.e., words like Swedish *jag* and Danish *jeg* or Swedish *säkert* and Danish *sikkert*.

This word list can be used together with the comparison program or the browsers to make comparisons between Swedish and Danish. Some of the comparisons which are possible are the following:

1. Use of graphemes in the two languages
2. Use of morphemes in the two languages
3. Use of words in the two languages
4. Use of collocations in the two languages
5. Use of discontinuous constituents in the two languages

Furthermore, both corpora make it possible to create subcorpora of men or women or of recordings of a certain age. Additionally, GSLC makes it possible to study the language of specific activities such as auction, shop, interaction, etc.

Below we will now give examples of exercises comparing Swedish with Danish that can be performed by interested users.

## Example 1. Comparing graphemes/phonemes

Compare Swedish and Danish words ending in *a* and *e*. This exercise will give data on the relative frequency of the two graphemes and indirectly on final vowels in Swedish and Danish. After having obtained the frequencies, users are encouraged to look at some of the examples of differences in frequency and to attempt to give an explanation of why the differences exists. Below are some examples of questions that can be asked.

### Word-final vowels

Many Swedish words end in a vowel (e.g. 'hemma', 'tvŒ', 'bli'). In table 1 below, all words in GSLC ending in a vowel are counted.

*Table 1* **Swedish words ending in a vowel**

| Final vowel | Number of words |
|---|---|
| ...a | 40546 |
| ...e | 33161 |
| ...i | 1047 |
| ...o | 1992 |
| ...u | 474 |
| ...y | 215 |
| ...å | 5347 |
| ...ä | 166 |
| ...ö | 132 |

*NOTE: very short words (L<3) are not counted*

Also many Danish words end in a vowel (e.g. 'hjemme', 'to', 'blive'). Table 2 shows the number of words in BySoc ending in a vowel.

*Table 2* **Danish words ending in a vowel**

| Final vowel | Number of words |
|---|---|
| ...a | 1984 |
| ...e | 73245 |
| ...i | 2887 |
| ...o | 355 |
| ...u | 870 |
| ...y | 386 |
| ...æ | 18 |
| ...¿ | 59 |
| ...å | 14086 |

*NOTE: very short words (L<3) are not counted*

### Exercises

Sort the end-vowels in table 1 by frequency ( *a>e>Œ>* ... ). Then do the same with vowels in table 2.

Compare the two sorted lists. What are the most significant differences?

Find 20 examples of Danish and Swedish words with the same (or equivalent) stem,

but different end-vowel (e.g. 'koka' vs. 'koge').
Describe some typical differences.
Explain why 'a' is the most frequent end-vowel in Swedish while much less frequent in Danish.
Words ending in 'o' are far more frequent in Swedish than in Danish. Why?
There are surprisingly many words ending in 'Œ' in (spoken) Danish. What words in this category are most frequent?

**Example 2. Comparing morphemes**

Compare Swedish and Danish for plural morphemes.

Compare Swedish *-or* (e.g. *yxor*), *-ar* (e.g. *bilar*) and *-er* (e.g. *böcker*) with their Danish counterparts and note the correspondences between Swedish and Danish plural.

**Example 3. Comparing morphemes**

Compare gender, find 5 words that have different gender in Danish and Swedish.

**Example 4. Comparing words**

Find two words from the list of cognates which have very different frequencies in Danish and Swedish and try to see if you can explain the difference in frequency.

**Example 5. Comparing Single word utterances**

Many utterances in spoken Swedish and spoken Danish consist of a single word only. The most common examples for GSLC and BySoc are listed in table 3 and 4 below (sorted by frequency).

*Table 3* **Swedish single-word utterances**

| Single word utterances | N |
|---|---|
| ja,jaa,ha | 11975 |
| m,mm,hm | 10783 |
| nej,nä | 3217 |
| jaha | 972 |
| okej | 379 |
| hej | 358 |
| men | 340 |
| va | 318 |
| jo | 312 |
| och | 309 |
| så | 289 |
| aha | 268 |

*Table 4* **Danish single-word utterances**

| Single word utterances | N |
|---|---|
| ja | 27685 |
| mm | 13575 |
| nej,næ | 5633 |
| nå | 5130 |
| jo | 1082 |
| så | 1080 |
| og | 1028 |
| men | 848 |
| altså | 754 |
| ik' | 579 |
| det | 568 |
| aha | 323 |

*NOTE: Single-word utterances are generally more frequent in BySoc than in GSLC. This has to do with differences in activity types and utterance definitions for the two corpora, rather than differences in the languages as such.*

GSLC Some of the most significant differences are:
Danish "nå" (rank #4 in table 4)
Danish "ik'" (rank #10 in table 4)
Neither of these have Swedish counterparts. Likewise,
Swedish "va" (rank #8 in table 3)
has no equivalent in Danish.

Search for these three particles ("interjections") in GSLC and BySoc.
Form an opinion of how the Swedish "va" translate to Danish
Likewise, suggest Swedish translations for Danish "nŒ" (as feedback particle) and "ik'".
You may need to include a preceding utterance in your translations in order to exemplify the typical use of these feedback particles.

**Example 6. Comparing collocations**

Departing from the list of cognates pick two words and some collocations in which they usually appear then use the two available browsers to determine their frequencies. Finally, try to explain possible differences. For example, you could compare Danish *sådan noget* with Swedish *sådant något, något sådant* eller *nåt sånt.*

**Example 7. Comparing utterance initial copula constructions**

Many Danish utterances begin with "der er" (2588 occurrences in BySoc). Here are some examples:
1. der er ingen der har mærket noget
2. der er et rimeligt antal af dem
3. der er ikke noget at tage fejl af
Swedish utterances beginning with "där är" are much less frequent (only 144 in GSLC). In addition, "där" rarely functions as an expletive (unlike English "there" and

Danish "der"). Some examples of utterances beginning with "där är":
4. där är lika bra vägar
5. där är jag född ja
6. där är vad dom har sagt till mig i alla fall KOLLA SÅ ATT DETTA ÄR RÄTT (DVS ”DÄR” I BÖRJAN, VAD ÄR KONTEXTEN? DET LÅTER OSVENSKT UTOM I MYCKET SPECIFIK KONTEXT, T EX OM MAN PEKAR PÅ EN KARTA
What is the Swedish equivalent to the Danish expletive "der"? Search GSLC and BySoc, and find 10 examples of Swedish utterances grammatically analogous to the Danish examples 1-3.

### Example 8. Comparing discontinuous collocations

Compare a discontinuous collocation in Danish with a corresponding collocation in Swedish, e.g. try Danish *ta den ut* and check Swedish collocations *ta X ut*.

### Example 9. Comparing the language of men and women in Denmark and Sweden

Using the browsers, create subcorpora then compare one or more expressions in the four corpora, e.g. compare Danish *jo* with Swedish *ju*.

# SweDanes
# spoken Swedish and spoken Danish compared

========================================

Introduction

In recent years, large amounts of spoken Swedish and spoken Danish has been recorded and collected for use in linguistic research. Two of the most largest and most significant collections are known as **GSLC** (Göteborg Spoken Language Corpus) and **BySoc** (Corpus BySociolingvistik, corpus of Danish spoken language).

Spoken language put down in writing (so called 'transcribed speech') does not look like ordinary written text:

men jag har en liten känsla av att intejuvar man ungdom av idag
det gäller inte bara om pensionärer utan om alla saker
eh dom inte mycket och säga egentligen
jag menar det jag tycker jag vet inte jag tycker det
jag tittade igår och det var någon jag satte på eller om det var i förrgår
det var ung+ tonåringar
ja det var öh ingenting var roligt egentligen
var det radio eller teve
ja att dom ska göra uttalanden och så vidare
jag tycker det är skäligen tunt det dom har
det roligaste var det var att gå och gå och dansa och lyssna på musik
jaha gå på bio ja
nä men jag menar vad är det skolan var tråkig hur gå
gå och köpa kläder handla kläder och höra på musik
det var ro+ men skolan var botten

(sample from corpus GSLC)

jeg er godt nok også blevet kørt over en gang
men der skete heller ikke noget
det var også dernede der lå i gamle dage før man byggede det nye
dernede der lå der Kongegården
og så lå der en slikker en slikmand og en bager og en købmand på hjørnet
og så skulle man altså over Baronessegade
og så ved jeg ikke hvor jeg har fået nogen penge fra
jeg har nok fået dem af min mor
men jeg skulle i hvert fald over og købe slik
og så drønede jeg simpelthen bare lige over gaden og der kom en bil
jeg kan godt huske jeg faldt og slog min næse sådan skrabede den
der var ikke det der hed kød tilbage på den
men ham der manden han var lige så ulykkelig som jeg var

(sample from corpus BySoc)

Any language student visiting a foreign country will have to face the challenge of taking part in such free-style conversations. For the language student there is thus good reason to prepare by studying real speech.

The material presented in this section of the CALL-demo is compiled with exactly that purpose: Showing how transcriptions of actual speech can be of assistance to language students in achieving a natural conversation style. Since the material covers spoken Swedish and spoken Danish only, it would probably be most relevant to Danish students of Swedish - and vice versa.

In the file **Swedanes_CALL.doc** (link to Word document), you'll find examples of exercises showing how to use the corpora GSLC and BySoc for didactic purposes.

In order to evaluate the exercises, you'll need to consult these three www-based resources:

| | |
|---|---|
| Corpus BySoc homepage | http://www.id.cbs.dk/~pjuel/cgi-bin/BySoc_ID/ |
| Corpus GSLC homepage | http://www.ling.gu.se/projekt/tal/ |
| Comparing BySoc and GSLC | http://www.id.cbs.dk/~pjuel/SweDanesDic |

We also include two so-called "frequency lists" - that is, listings of all the most frequent words occurring in GSLC and in BySoc, sorted by frequency. Notice, for instance, that the most frequent word in spoken Danish as well as Swedish, is pronoun *det*.

Shown here are the beginnings of the two frequency files (including the number of occurrences of each word) - click on links to fetch lists:

| **Freqency_list_BySoc.doc** | | **Freqency_list_GSLC.doc** | |
|---|---|---|---|
| 74159 | det | 84007 | det |
| 47127 | ja | 39965 | är |
| 41538 | og | 37884 | och |
| 39371 | jeg | 34731 | ja |
| 38317 | er | 32912 | att |
| 36193 | så | 30745 | jag |
| 32305 | der | 28079 | så |
| 24869 | ikke | 20417 | som |
| 23467 | var | 20164 | inte |
| 23344 | i | 19880 | vi |

Have fun!

Peter Juel Henrichsen, pjuel@id.cbs.dk
Jens Allwood, jens@ling.gu.se

## The SweDanes project

The research project that laid the grounds for this cd-material is known as *SweDanes*. You can learn more about this project here: Allwood, J.; P.J.Henrichsen, E.Ahlsén, M.Gunnarsson,

L.Grönquist (to appear) *Transliteration between Spoken Language Corpora,* in Nordic Journal of Linguistics

Henrichsen, P.J. (2004) *Siblings and Cousins; Statistical Methods for Spoken Language Analysis*; Acta Linguistica Hafniensia 36, pp.7-33

Allwood, J.; P.J.Henrichsen, E.Ahlsén, M.Gunnarsson, L.Grönquist, K.Voionmaa, H.Vappula, L.Grönquist (2004) *Några Frekvensbaserade Skillnader Mellan Svenskt och Danskt Talspråk;* proceed. of Nordic Symposium on the Comparison of Spoken Languages (ed. P.J.Henrichsen), pp.69-98, Copenhagen Business School Press

·Allwood, J. (ed.) (1999) *Talspråksfrekvenser - Ny og Utvidgad Upplaga,* Gothenburg Papers in Theoretical Linguistics, S21; Gothenburg University Press