In Nordisk Sprogteknologi, Nordic Language Technology: 2002

Sweden Göteborg

Corpus Based, Spoken Language and Computational Research

by

Jens Allwood,
Department of Linguistics, Göteborg University

The Department of Linguistics at Göteborg is active both in the empirical study of dialog and the study of computational dialog systems. It has a spoken dialog corpus of 1.3 M words transcribed and annotated. The department contains researchers in linguistics, language and speech technology with a common interest in work on spoken language and dialog and of using spoken language corpora as a basis for research. Some European projects in which the department has participated are: e.g. PLUS (A Pragmatics Based Language Understanding System), TREE . Scientific coordinator for EC LE project TRINDI (Task Oriented Instructional Dialogue).Contributed to SDS Swedish Dialogue Systems (NUTEK & HSFR), part of SIRIDIUS, involving speech recognition. The department is host of the Swedish "Forskarskola" for Language Technology.

Work on spoken language structures in the Scandinavian languages needs to be strengthened. We feel that the there is need for cooperation between the Nordic countries, especially for the typologically similar Scandinavian languages, in establishing spoken language corpora and tools for employing these corpora. Researchers from linguistics/language and speech technology can all gain by cooperating in solving problems of spoken language processing that have direct relevance for applications of language and speech technology.

Corpus based spoken language grammar is our main interest in the NORDTALK network. So far, only a few persons in the Nordic countries have worked seriously in this area and we feel that it is important to bring them together. We also feel that cooperation with speech technology groups can lead to interesting combinations of approaches in spoken language analysis.

Finally, but not least, the area of language and speech technology needs educational programs producing future researchers with a broad background for working with speech based and multimodal systems and we believe that the network can provide a basis for providing such education.

GSLC - Göteborg Spoken Language Corpus

The Swedish Spoken Language Corpus at the Department of linguistics, Göteborg University is an incrementally growing corpus of spoken language from about 25 different social activities. It is part of the Göteborg spoken language corpora. Besides the spoken language corpora there are also several written corpora. Based on the fact that spoken language varies considerably in different social activities with regard to pronunciation, vocabulary and grammar, the goal of the corpus is to include spoken language from as many social activities as possible. The activites transcribed so far are the following.

Activities	**Word tokens**	**Activities**	**Word tokens**
Auction	26 459	Interview	389 416
Bus driver/passenger	1 345	Lecture	14 667
Church	10 234	Market	12 175
Consultation	34 285	Phone	14 614
Court	33 722	Retelling of article	5 290
Dinner	30 001	Role play	8 055
Discussion	243 090	Shop	50 492
Factory conversation	28 860	Task-oriented dialogue	
Formal meeting	215 582	Therapy	13 529
Hotel	18 137	Trade fair	14 116
Informal conversation	74 502	Travel agency	40 111
		Total	**1 294 029**

For an overview of the activities transcribed so far, see also http://www.ling.gu.se/projekt/SLSA/medium_overview.html. You can also view a complete listing at http://www.ling.gu.se/projekt/SLSA/big_overview.html.

The spoken language material has been transcribed according to the transcription standard Modified Standard Orthography MSO, Modifierad Standardortografi, which is a standard for transcription which is more faithful to spoken language than Swedish standard orthography but less detailed than a phonetic or phonematic transcription would be.

In MSO, standard orthography is used unless there are several spoken language pronunciations of a word. When there are several variants, these are kept apart graphically. According to this principle, the Swedish word "jag" (I), which is mostly pronounced "ja" but occasionally as "jag" is written in both these ways, depending on which form is actually used. What variants can be distinguished is, however, to some extent arbitrary and has, therefore, in some cases been decided on a stipulative basis. Thus, we have not, in general, distinguished words on the basis of vowel length. For an example of a transcription example, see below.

Example 1. Transcription according to the MSO standard with translation.

§1. Small talk
$D: säger du de{t} ä{r} de{t} ä{r} de{t} så besvärlit då
$P: ja ja
$D: m // ha / de{t} kan ju bli så se{r} du
$P: < jaha >
@ <ingressive>
$D: du ta{r} den på morronen
$P: nej inte på MORRONEN kan ja{g} ju tar allti en promenad på förmiddan [1 å0]1 då vill ja{g} inte ha [2 den]2 medicinen å0 sen nä ja{g} kommer hem möjligtvis
$D: oh I see is it it is so troublesome then
$P: yes yes
$D: m // yes / it can be that way you see
$P < yes >
@ <ingressive >
$D: you take it in the morning
$P: no not in the MORNING I always take a walk before lunch [1 and]1 then I don't want [2 that]2 medicine and then when I get home possibly
$D: [1 yes]1

$D: [1 {j}a]1 $D: [2 no]2
$D: [2 nä]2

The example shows the following properties of the transcription standard:

(i) Section boundaries paragraph sign (§). These divide a longer activity up into subactivities. A doctor-patient interview can, for example have the following subactivities. (i) greetings and introduction, (ii) reason for visit, (iii) Investigation, (iv) prescribing treatment.
(ii) Words and space between words.
(iii) Dollar sign ($) followed by capital letter, followed by colon (:) to indicate a new speaker and a new utterance.
(iv) Double slash (//) to indicate pauses. Slashes /, // or /// are used to indicate pauses of different length.
(v) Capital letters to indicate contrastive stress.
(vi) Word indexes to indicate which written language word corresponds to the spoken form given in the transcription (å0 corresponds to written language *och*). In the cases where spoken language variants can be viewed as abbreviated forms of written language, we use curly brackets {} to indicate what the standard orthographic form would be, e.g. de {t} = *det*.
(vii) Overlaps are indicated using square brackets ([ ]) with indices which allow disambiguation if several speakers overlap simultaneously.
(viii) Comments can be inserted using angular brackets (< > to mark the scope of the comment and @< > for inserting the actual comment). These comments are about events which are important for the interaction or about such things as voice quality and gestures.

Regarding analysis of the corpus we have produced a first book of frequencies of Swedish spoken language. The book contains word frequencies both for the words in MSO format and in standard format. It also contains comparisons between word frequencies in spoken and written language. These lists are given in alphabetical and frequency order. There are list of frequencies for collocations in MSO, standard orthography and written language. Connected with the word frequencies, there are lists of words which are unique to or very much more common in spoken MSO spoken language rendered in standard orthography of written language. Finally, there is statistics on the parts of speech represented in the corpus, based on an automatic probabilistic tagging, yielding a 97% correct classification. Below we present the most frequent Swedish words in non-disambiguated speech, in speech disambiguated to the same level of disambiguation as in writing and in writing.

Rank number	**Non-disambiguated speech**		**Disambiguated speech**		**Writing**	
Swedish	**English**	**Swedish**	**English Swedish**	**English**		
1	de	(it/they/the/that/there)	det	(it/that/there)	och	(and)
2	ja	(yes/I)	är	(is)	i	(in)
3	e	(is/eh)	och	(and)	att	(that)
4	å (	and/that)	ja	(yes)	det	(it/that/there)

5	så	(so)	att	(that	en	(a)
6	att	(that)	jag	(I)	som	(which)
7	va	(what/was)	som	(which)	på	(on)
8	som	(which)	inte	(not)	är	(is)
9	vi	(we)	har	(has/have)	med	(with)
10	inte	(not)	vi	(we)	för	(for)

There are three columns in the table. In the first column there is spoken language fairly closely to real speech. In the second column we have written the spoken language words, the way they would be written in standard orthography and in the third column we have frequencies from a written language corpus. As we can see, the differences are considerable. In fact, they are even bigger if punctuation marks are included. Period (.) and comma (,) are actually the two most common meaningful units of the written language corpus. In the following table, we present some more data from our analysis – the most common words in the different parts of speech. The words are presented in standard written language orthography. Parts of speech – Five most frequent words (disambiguated speech to the level of writing)

1. Pronouns	**2. Verbs 3.**	**Adverbs**
det (it, that, there)	är (is)	så (so)
jag (I)	har (have/has)	inte (not)
vi (we)	kan (can	då (then)‹
man (one	ska (will)	ju (you know)
du (you)	var (was/were)	här (here
4. Nouns	**5. Conjunctions**	**6. Prepositions**
naturen (the nature)	och (and)	på (on)
sätt (manner)	att (that)	i (in)
kronor (crowns)	men (but)	för (for)
fall (case)	om (if)	med (with)
exempel (example)	eller (or)	till (to)
7. Feedback words words	**8. Adjectives**	**9. Own communication**
ja (yes)	sådant (such)	eh (eh)
m (m)	bra (good)	äh (eh)
nä (no)	sådan (such)	öh (eh)
va (what	sådana (such)	-hm (-hm)
jo (yes	olika (different)	-s+ (-s+)
10. Numerals	**11. Interjections**	
två (two)	tack (thanks)	
tre (three)	hej (hi)	
femtio (fifty)	oj (oh)	
fem (five)	fan (devil)	
första (first)	varsågood (please)	

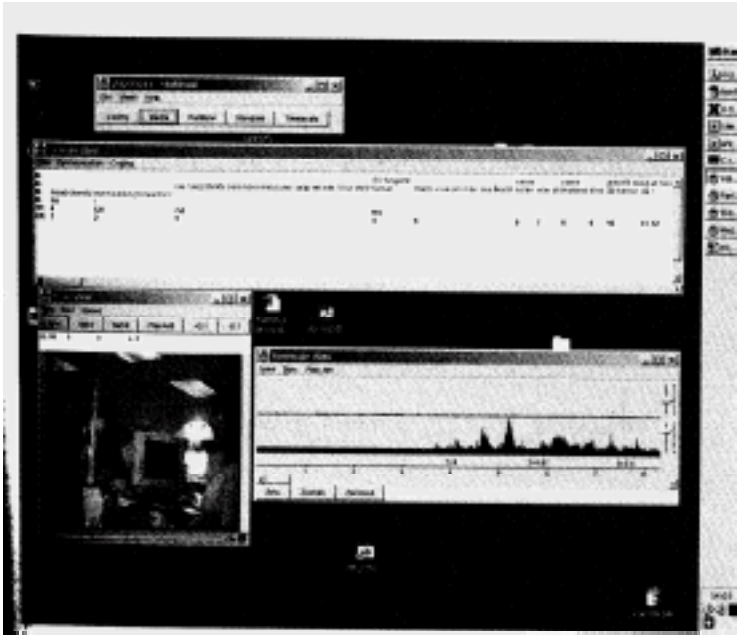
The parts of speech are given in descending rank order. Thus, we can see that pronouns are the most common part of speech in spoken language.

Tagging and Coding

A statistical (Viterbi - trigram) parts of speech tagger has been adapted to spoken language. Using this, a parts of speech coding has been done for the whole Göteborg

Spoken Language Corpus, roughly 1.2 million transcribed words. The correctness of the coding is about 97% (cf Nivre & Grönqvist, 1999).

Besides parts of speech, work has also been done on coding the following linguistic features and coding schemas are available for download at http://www.ling.gu.se/projekt/SLSA/coding.html.



- Addressee, Turn & Sequence Management (Tsm)
- Speech Acts
- Expressive & Evocative Functions and Obligations
- Linguistic feedback
- Own communication management (OCM)
- Maximal grammatical categories
- Subactivities

Tools

1) MULTITOOL a synchronizing tool for video, acoustic analysis, transcription and coding •
2) TRANSTOOL: a prototype tool for supporting transcriptions TRACTOR: a prototype tool for supporting coding
3) TraSA: a prototype tool for selecting subcorpora and automatic analysis of our spoken language corpus.
4) Coding Visualizer: a tool for visualizing codings in framemaker
5) Corpus Browser: A tool that makes it possible to search for words, word combinations and phrases (as regular expressions) in the corpus and present them as concordances or lists of utterances.

In Multitool, a transcription of a spoken language file is synchronized with an audio or video file. You can hear and see the people producing the transcribed words. You can also see an acoustic analysis ans a coding of what has been said. For contact with persons in the general public who are interested in spoken language, we have started Talspråksklubben, which is an arena for contact with persons who would like to help us make recordings and transcriptions of spoken language in different social activities, see http://www.ling.gu.se/projekt/SLSA/talklubben.html.