In Proceedings Fonetik 99:The Swedish Phonetics Conf. June 1999: Göteborg University, Sweden

The Swedish Spoken Language Corpus at Göteborg University

Jens Allwood, Department of Linguistics, Göteborg University

1. The Corpus

The Swedish Spoken Language Corpus at Göteborg University is an incrementally growing corpus of spoken language from different social activities. It is a part of the Göteborg spoken language corpora. Besides these 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. For an overview of the activities transcribed so far, see table I below. You can also view a complete listing in pdf format or download the postscript format. Some of the transcriptions are also available for download at: http://www.ling.gu.se/SLSA/transcriptions/.

Table 1: Overview of Göteborg Spoken Language Corpus

Type of activity	**Number of activities**	**Amount of words**
Shop	16	46 599
Occupational therapy	1	7 907
Auction	2	26 459
Discussion	22	134 722
Court	6	33 261
Formal meeting	11	176 887
Quarrel	1	760
Hotel	9	18 137
Informal conversation	18	74 610
Interview	54	395 395
Classroom interaction	2	14 667
Consultation	15	24 450
Dinner	5	30 001
Trade fair	16	14 116
Sermon	2	10 234
Radio:Spanama	2	14 012
Role play	3	8 055
Conversation in factory	5	28 860
Seminar discussion	4	39 652
Physical therapy	1	5 622
Phone	32	14 614
Travel Agency Dialogue	35	30 195
Market	4	12 175
TV	5	47 291
Task-oriented dialogue	20	15 347
Retelling of article	7	5 290
Bus driver/pass.convers.	1	1 345
TOTAL	299	1 230 663

2. Transcription

The spoken language material has been transcribed according to the transcription standard Modified Standard Orthography MSO (Modifierad Standardortografi), Nivre 1999 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. The transcription standards are described in detail in Nivre (1999) and compared to other transcription standards used in Sweden in Allwood, Abelin and Grönqvist (1998). They are available for download in ps and pdf format at
http:/www.ling.gu.se/SLSA/Postscripts/AppMSO6.ps. MSO.

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.

Through this practice, sometimes words which are pronounced the same way, but kept apart in standard orthography, will coincide. This, for example, happens to "jag" (I) pronounced as "ja" and "ja" (yes). When this happens, the words have been disambiguated by brackets or numerical indexes. In this case, 'ja{g}" (jag) and "ja" (yes). If the spoken form is produced by just removing letters from the standard form, then brackets are used to indicate the corresponding standard form. If the spoken forms can't be disambiguated by brackets, then numerical indexes are used. For example, the spoken form "å' can mean "och" ("and") or "att" ("to" - infinitive marker), so the transcribed form is "å0" for "och" and "ål" for "att". Thus, MSO maintains the same degree of disambiguation as standard written orthography but adds to this the disambiguations which are actually added by spoken language, e.g. between Swedish standard orthography "att" (that, to) which can be pronounced as "å" ("to" –infinitive marker) or "att" ("that" conjunction). However, no attempt is made to separate homonyms which are separated neither in written or spoken language. This means that one can not know from a word form like "springa" (run, chink) whether it is a verb or a noun. The conventions of MSO are exemplified and briefly explained below:

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

§ 1. Small talk	
$D: säger du de0 ä0 de0) ä0 de0) så0 besvärlit då	$D: oh I see is it it is so troublesome then
$P:jaøja0	$P: yes yes
$D: m0// ha0 / de0 kan ju bli0 så0 se1 du	$D: m f/ yes / it can be that way you see
$P: <jaha>	$P<yes>
@ <ingressive>	<ingressive>
$D: du tal den på morronen	$D: you take it in the morning
$P: nej inte på MORRONEN kan jal ju tar allti en0 promenad på förmiddan [l å0]1 vill ja0 inte ha0 [2 den]2 medicinen å0 sen1 nä1 ja0 kommer hem möjligtvis	$P: no not in the MORNING I always take a walk before lunch fl and] I then I1 då don't want [2that]2 medicine and then when I get home possibly
$D: [1 a0]l	$D: [1 yes] l
$D: [2 nä0]2	$D: [2 no]2

As we can see the transcription has the following properties

(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 (II) to indicate pauses.
(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 (de0 corresponds to written language det).
(vii) Overlaps are indicated using square brackets ([1) with indices which allowdisambiguation 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.

Talspråksklubben: http://www.ling.gu.se/SLSA/talklubben. html is an arena for contact with interested persons in the general public who would like to help us make recordings and transcriptions of spoken language in different social activities.

3. Analysis and Coding

Regarding analysis of the corpus we have produced a first book of frequencies of Swedish spoken language, Allwood (1998 & 1999). 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 lists 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 96% correct classification.

Further, there has been work on the corpus using various kinds of manual coding for communication management (including hesitations, changes, feedback and turntaking), speech acts, obligations, maximal grammatical units, etc. In Abelin and Allwood (1998), 0CM (Own Communication Management) coding is compared to "Dysfluency" coding. For the coding work we have a link containing a transcription with coding and manuals available, cf `http://www.ling.gu.se/SLSA/dialog.html`. Below is an overview of the kind of codings we are using.

Table 2. Types of coding

- General

1. Activity: Purpose, Roles, Artifacts, Environment
2. Participants: Number, Types of listeners: specific, group, all

- Grammar related

3. Utterances - Maximal grammatical categories
4. Total grammatical classification

- Semantics

5. Several types of Semantic classification

- Functional Coding

6. Holistic Communicative Acts (Implicit, Context relations)
7. Communication Management: 0CM, 1CM (FB, TM, Seq)
8. Directional function (Expressive, Evocative function)
9. Obligations (Speaker & Listener related)
10. Chunking into sequences

Manuals

1. OCM-manual (Own Communication Management):
   http://www.ling.gu.se/SLSA/Postscripts/OCMmanual_vl.0.ps
   http://www.ling.gu.se/SLSA/pdf/OCMmanual_vl.O.pdf

2. 1CM-manual (Interactive Communication Management for coding of feedback and turnand sequence management):
   http://www.ling.gu.se/SLSA/postscripts/fb_manuall.ps
   http://www.ling.gu.se/SLSA/pdf/ICM_manual.pdf

3. MaxGram coding manual:
   http://www.ung.gu.se/SLSA/Postscripts/MaxGram.ps
   http://www.ung.gu.se/SLSA/pdf/NaxGram.pdf

4. Tools for Corpus Work

We have also been working on computer based tools for producing and analyzing spoken language copora, cf Nivre et al. (1998). The following is a list of these tools:

Tools

1. TRANSTOOL
   A prototype tool for supporting transcriptions
   http://www.ung.gu.se/~sylvana/SLSA/TransTool.html

2. TRACTOR
   A prototype tool for supporting coding
   http://www.ung.gu.se/~sl/tractor.html

3. TRASA
   A prototype tool for selecting subcorpora and automatic analysis of our spoken language corpus
   http://www.ling.gu.se/~leifg/doc/TrasA08.ps (postscript)
   http://www.ling.gu.se/~1eifg/doc/trasa08.pdf (pdf)

4. Kodningsvisualisering med FrameMaker
   http://www.ling.gu.se/~leifg/doc/kodvisualisering.pdf
   http://www.ling.gu.se/~leifg/doc/kodvisualisering.ps

5. MultiTool
   A tool for synchronizing transcriptions, codings, audio, video files and acoustic analysing.
   1:st prototype - 2nd under way. JAVA and JMF.

References

Allwood, J. 1998**.** Some observations on Swedish spoken language frequencies. In *Proceedings of the 16th Scandinavian Conference of Linguistics*, Turku University,Department of Linguistics.

Allwood, J. (red.) 1999. Talspråksfrekvenser. Frekvenser för ord och kollokationer i svenskt tal- och skriftspråk, Gothenburg Papers in Theoretical Linguistics, Göteborg University, Department of Linguistics

Abelin, A & Allwood, J. 1998. Jämförelse mellan OCM-kodningsstandard och Robert Eklunds disfluenskodningsstandard. Technical Report, Dept of linguistics, Göteborg University

Allwood, J., Abelin, A., Grönkvist, L. 1998: Kort beskrivning och jämförelse av transkriptionssystem från Lund, Telia, Linköping och Göteborg
http://www.ling.gu.se/~leifg/jfresle_transkriptionssystem.pdf.
(pdf)
http://www.ling.gu.se/~leifg/jfresle_transkriptionssystem.ps.
(ps)

Nivre, J. Tullgren, K., Allwood, J., Ahlsén, E., Holm, J., Grönqvist, L., Lopez-Kästen, D. & Sofkova, S. Towards multimodal spoken language corpora: TransTool and SyncTool. *Proceedings of ACL-COLING 1998*, June 1998.
http://www.ling.gu.se/~leifg/doc/COLING98.pdf (ps)

Nivre, J., 1999. Modifierad Standardortografi, Version 6. Institutionen för lingvistik, Göteborgs universitet