The Spoken Language Corpus at the Department of Linguistics, Göteborg University

Jens Allwood, Maria Björnberg, Leif Grönqvist,
Elisabeth Ahlsén and Cajsa Ottesjö

Abstract: This paper describes work on spoken language at the Department of Linguistics Göteborg University. First, we describe what recordings are contained in the Spoken Language Corpus of Swedish at Göteborg University. Secondly, we describe the standard of transcription (MSO) which is used in the transcriptions. Thirdly, we describe computer tools that have been developed to support transcription, coding and analysis. Fourthly, we describe some types of quantitative and qualitative analysis that have been done. Fifthly, we briefly mention some of the results which have been obtained.

Key words: Spoken language, corpus, Göteborg, transcription standard, MSO, coding, computer tool, qualitative analysis, quantitative analysis.

1. Introduction

The paper contains a description of the Spoken Language Corpus of Swedish at Göteborg University and a summary of the various types of analysis that have been done on this corpus. The corpus is an incrementally growing corpus of spoken language which presently consists of 1,26 million words from about 25 different social activities. It is based on the fact that spoken language varies considerably in different social activities with regard to pronunciation, vocabulary, grammar and communicative functions. The goal of the corpus is to include spoken language from as many social activities as possible to get a more complete understanding of the role of language and communication in human social life.[1]

The corpus is based on audio (50%) or video/audio (50%) recordings of naturalistically occurring interactions. The recordings have been transcribed according to the transcription standard Modified Standard Orthography MSO, 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 pronunciation variants of a word. When there are several variants, these are kept apart graphically. Although the goal is to keep transcription simple, MSO includes features of spoken language such as contrastive stress, overlaps and pauses. MSO also includes procedures for anonymizing transcriptions and for introducing comments on part of the transcription. [2]

Besides describing our corpus and standard of transcription, we will also describe several tools we have developed for using the corpus, e.g.: [3]

1. The corpus browser, which is a tool used for searching in the spoken language corpus according to different criteria like gender, age, activity type etc.

2. Tools for coding(several different coding schemas are available) and for calculating frequencies. [4]

3. A tool for synchronizing video and audio recordings, codings, transcriptions and acoustic analysis called MULTITOOL. This tool is presently being developed. [5]

The corpus has been used for various kinds of quantitative and qualitative analysis which will be briefly reported. A book of frequencies of Swedish spoken language has been produced. The book contains word frequencies both for the words in MSO format and in standard orthographic format. It also contains comparisons between word frequencies in spoken and written language. There is statistics on the parts of speech represented in the corpus, based on an automatic probabilistic tagging, yielding a 97% correct classification. The corpus has also been the basis for work using various kinds of manual coding, e.g. communication management (including hesitations, changes, feedback and turntaking), speech acts, obligations, misunderstandings, etc. Finally, we point to ways of using the corpus for other types of qualitative analysis, e.g. for CA-related sequential analysis. [6]

The corpus is continuously being digitized using CD:s with Mpeg compression. Each CD contains both transcriptions and recordings. [7]

2. Corpus Description

The Göteborg Spoken Language Corpus consists of spoken language samples from several languages. So far the following languages are included. Table 1. Over and above this we also work with other spoken language corpora collected by other teams. [8]

Table 1. Spoken language corpora at Göteborg University. Some of the corpora are Multimodal

- Göteborg Spoken Language Corpus (Kernel Corpus - adult 1st language Swedish), 1.2 million words
- Adult language learners of Swedish
- Aphasic speakers
- Child language corpus (Swedish and Scandinavian), 0.5 million words including the adults
- Non-Swedish adult spoken language corpus
  * Chinese (70 000 words)
  * Bulgarian (25 000 words)
  * Arabic
  * English (10 000 words), BNC
  * Finnish
  * Italian (3000 words)
  * Norwegian (140 000 words)
  * Spanish
- Wizard-Of-Oz Corpus, Bionic
- Intercultural communication corpus [7]

As can be seen the largest corpus is the kernel corpus of Adult 1st language Swedish speakers. This is the corpus we will focus on in this article. In Table 2, below, we present basic data on this corpus. As will be seen the corpus is organized on the basis of social activities rather than for example on the basis of dialects or categorizations of speakers such as social class or gender. However, regroupings of, or selections from, the corpus according to criteria such as these are also possible. The limitations which exist in our ability to create subcorpora are dependent on the fact that we do not always have the relevant information about individual speakers. [9]

Table 2. The corpus of adult list language Swedish at Göteborg University

Göteborg Spoken Language Corpus

Activity	Record-ings	Speakers	Sec-tions	Tokens	Audible	Duration
Auction	2	6.0 *	111	26 776	26 459	3:14:11
Bus driver/ passenger	1	33.0 *	20	1 360	1 345	0:13:33
Consultation	15	2.3 *	198	24 916	24 450	1:47:25
Court	6	5.0 *	79	33 401	33 261	3:58:33
Dinner	5	8.0 *	30	30 738	30 001	2:49:54
Discussion	33	5.9 *	255	240 426	237 583	17:02:54
Factory conversation	5	7.4 *	48	29 024	28 860	2:19:47
Formal meeting	12	9.8 *	153	206 564	202 923	14:14:39
Hotel	9	19.2 *	183	18 950	18 137	6:47:50
Informal conversation	22	4.4 *	152	94 490	93 436	7:48:41
Information Service (phone)	32	2.1 *	40	14 700	14 614	0:13:40
Interview	56	2.7 *	1 021	388 959	386 444	30:20:27
Lecture	2	3.5 *	3	14 682	14 667	1:38:00
Market	4	24.2 *	38	12 581	12 175	2:18:37
Religious Service	2	3.5 *	10	10 273	10 234	1:10:45
Retelling of article	7	2.0 *	7	5 331	5 290	0:42:00
Role play	2	2.5 *	7	5 702	5 652	0:39:16
Shop	48	7.5 *	137	32 339	30 970	6:09:27
Task-oriented dialogue	26	2.3 *	46	15 475	15 347	2:05:20
Therapy	2	7.0 *	8	13 841	13 529	2:04:07
Trade fair	16	2.1 *	16	14 353	14 116	1:12:46
Travel agency	40	2.7 *	112	40 370	40 129	5:53:57
Total	**347**	**4.9 ***	**2 674**	**1 275 251**	**1 259 622**	**114:45:49**

Table 2 contains 7 columns. The first column gives information about what type of social activity the recordings concern. The second column describes how many recordings of each activity type exist. The third column describes how many speakers a particular activity type has on average (indicated by asterisks (*)). The fourth column gives information about the number of sections in each activity instance. A section is a longer phase of an activity with a distinct subordinate purpose. The bus driver/passenger recording, for example, has 20 sections, where each section involves talk with a new passenger. Column five gives information about word tokens as well as about pauses and comments, while column 6 only includes words actually uttered in the recording. Finally, column 7 gives the temporal duration of each activity. Due

to lack of resources, the duration has, in most cases, been estimated on the basis of the number of word tokens. The estimate is conservative and probably under-represents actual duration by about 30 hours. [10]

3. Storage

Around 50% of our 1.2 million tokens corpus is stored on audio tapes and the rest on video tapes (Umatic, VHS or BetaCAM). In order to preserve the recordings, tapes are being copied to newer tapes, while simultaneously being digitized. There are several possible formats for storage: [11]

(i) Analog video: BetaCAM is probably the best analog video format but VHS is almost the only one used nowadays. One problem with analog formats is that the quality gets worse for every copy which is not the case with digital formats.[12]

(ii) DV (digital video): One mini DV-tape takes 60 minutes or a DVCam 180 minutes. This format requires a fast computer. [13]

(iii) Mpeg: We have tried to use a constant data rate of around 200 kb per second. This will give a fair quality and the format may be used on almost any PC/Mac. For phonetic analysis the sound should not be compressed with mpeg but with some non-destructive method. An Mpeg card capable of creating Mpeg 1 or 2 with a variable data rate and a speed of 200 kb/sec should be enough for very good video quality. The sound could probably be stored as CD-quality raw-data, compressed separately without loss. The mpeg audio/video + raw audio could be recorded on CD:s with up to 60 minutes per disc compared to 3 minutes in the DV-format. [14]

4. Description of MSO (Modified Standard Orthography) - the corpus transcription standard

The transcription standard we have used is called MSO (Modified Standard Orthography). It can perhaps most rapidly be explained through exemplification. Consider the example below: [15]

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: [1 {j}a]1
$D: [2 nä]2

$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: [2 no]2

The example shows the following properties of the transcription standard: [16]

(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. [17]

5. Tools which have been developed

The following tools have been developed to aid work related to the corpora. [18]

5.1 TransTool (http://www.ling.gu.se/~sylvana/SLSA/TransTool.html)

TransTool is a computer tool for transcribing spoken language in accordance with the transcription standard (Nivre 1999). It will help the user to transcribe correctly and make it much easier to keep track of indices for overlaps and comments (cf. Nivre et al. 1998). [19]

5.2 The Corpus Browser

The Corpus Browser is a tool that makes it possible to search for words, word combinations and phrases (as regular expressions) in the Gothenburg Spoken Language Corpus. The results can be presented as concordances or lists of utterances with as much context as you wish and with direct links to the transcription.[20]

5.3 TRACTOR (http://www.ling.gu.se/~sl/tractor.html)

Tractor is a coding tool which makes it possible to create new coding schemas and annotate transcriptions. Coded segments can be discontinuous and it is also possible to code relations. A coding schema can be represented as a tree with strings on all nodes and leaves, and a coding value is a path through the tree. That model is similar to the file and folder structure on a computer harddisk. This framework makes it easy to

analyze the codings in a Prolog system, but it is not possible to order the codings or code a coding, because a coding only consists of two discontinuous intervals and a coded value. [21]

5.4 Visualization of codings with FrameMaker (http://www.ling.gu.se/SDS/doc)

This document describes a toolbox that makes it possible to visualize coding schemas and coding values with colors. bold, italics, etc. directly in the transcription as a FrameMaker document. Different parts of the transcription may also be marked (or removed!) to get a legible view of it without all details you are not interested in. [22]

5.5 TraSA (http:// www.ling.gu.se/SDS/doc)

If you have a corpus transcribed according to the Göteborg Transcription Standard, TraSA it is very easy to calculate some 30 statistical measurements for different sections and/or speakers. You will be able to count things like number of tokens, types, utterances, theoretical vocabulary. No other tool makes it possible to partition a corpus and calculate all these measurements without programming and statistical skills. [23]

5.6 SyncTool

SyncTool is developed (as a prototype for MultiTool) for synchronizing transcriptions with digitized audio/video recordings. It is also meant to be a viewing tool allowing the researcher to view the transcription and play the recording without having to manually locate the specific passage in the recording. [24]

5.7 Work on a synchronizing tool – MULTITOOL (http://www.ling.gu.se/SDS/multitool)

Work has been done on a tool for synchronizing dialog transcriptions with audio and/or video files for the same dialogs (cf. Nivre et al., 1998). [25]

MultiTool is an attempt to build a general tool for linguistic annotation and transcribing dialogues, browsing, searching and counting. The system can handle any number of participants, overlapped speech, hierarchical coding schemes, discontinuous coding intervals, relations and synchronization between codings and the media file. [26]

The internal state: The fundamental idea is to collect all information in an internal state containing only codings and synchronizations. Even the transcription is made by codings. The internal state can be visualized with a number of different views. [27]

The views: The Standard View shows one utterance on each line, overlaps and other details that the user wants are marked. [28]

The Partiture View has one line for each participant and the codings are viewed in chronological order along the x-axis. This will give a clear view of the dialogue structure and the overlapping sections. [29]

The Coding View shows the tree structure of all coded values so far, and their frequencies. Each value can be expanded to the next level in a similar way as Windows Explorer. [30]

The Media Player will play audio and video. The user can navigate through the media file to find interesting sections. [31]

The Time Scale shows the codings in linear time and the sound waveform which is very useful when aligning coding points and media. [32]

Why views? One important detail is that the views can be synchronized to show the same sequence when the user scrolls in one of them. The internal state contains all information so it is possible to have many views of the same kind, showing different parts of the dialogue. Changes made in one view will immediately change in the internal state and the other views. [33]
Codings: A coding consists of two discontinuous intervals (lists of starting and ending coding points), one list of speakers, and a coded value. It should be interpreted as a relation between the two intervals. Transcribed words is a special case where the first interval is continuous and the second an empty list. A synchronization indicates that a specific coding point corresponds to a specific time. [34]
Implementation: MultiTool is written in JAVA+JMF which makes it platform independent, and the interpreters are rapidly getting more efficient so the performance will probably be good enough on the major platforms very soon. A second prototype is now finished and in use. The architecture makes it easy to expand the system with new type of views. [35]
Download: Multitool as well as examples and The MultiTool User's Manual may be downloaded from http://www.ling.gu.se/SDS/multitool. [36]

6. Types of quantitative analysis

Using the information provided by the MSO compliant transcriptions, we have defined a set of automatically derivable properties which include the following: [37]

(i) **Volume**: Volume comprises measures of the number of words, pauses, stresses, overlaps, utterances, turns relative to speaker, activity and subactivity. [38]

(ii) **Ratios:** Various ratios can then be calculated based on the volume measures.

For example:			
	MLU	=	words/utterances
	% pauses	=	100*pauses/(words+pauses)
	% stress	=	100*stressed words/words
	% overlap	=	100*overlapped words/words [39]

Alternatively, pause, stress and overlap can be given per utterance. All of these measures can then be relativized to speaker, activity or subactivity. [40]

(iii) **Special measures**: One example of a special type of measure is "vocabulary richness" as measured through type/token, Guiraud, Über, Herdan or "theoretical vocabulary", cf. van Hout & Rietveld (1993). Another measure we have constructed is "stereotypicality" which looks at how often words and phrases are repeated in an activity. [41]

(iv) **Lemma:** We also implemented a simple stemming algorithm which enables us to collect regularly inflected forms together with their stem. [42]

(xii) **Parts of speech**: Parts of speech are assigned using a probability based statistical (Viterbi - trigram) parts of speech tagger which 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). Words subdivided according to parts of speech can then be assigned to speaker, activity or subactivity. [43]

(vi) **Collocations**: All speakers, activities and subactivities can be characterized in terms of their most frequent collocations. [44]

(vii) **Sequences of parts of speech**: Utterances of different length can be characterized as to sequence of parts of speech. This allows a first analysis of grammatical differences between speakers, activities and subactivities. [45]

(viii) **Similarities:** Similarities between activities are captured by looking at the extent to which words and collocations are shared between activities. [46]

7. Types of qualitative analysis

7.1 Overview

In order to increase reliability, qualitative analysis in Göteborg has often resulted in the development of coding schemas. [47]

The following provides an overview of the Göteborg coding schemas: [48]

1. Social activity and Communicative act related coding
   1.1 Social activity
   1.2 Communicative acts
   1.3 Expressive and Evocative functions
   1.4 Obligations
2. Communication management related coding
   2.1 Feedback
   2.2 Turn and sequence management
   2.3 Own Communication Management
3. Grammatical coding
   3.4 Parts of speech (automatic, probabilistic)
   3.5 Maximal grammatical units
4. Semantic coding. [49]

7.2 Contributions, utterances and turns

Following Grice (1975), Allwood, Nivre and Ahlsén (1990) and Allwood (1995), the basic units of dialog are gestural or vocal *contributions* from the participants. The term *contribution* is used instead of *utterance* in order to cover also gestural and written input to communication. Verbal contributions can consist of single morphemes or be several sentences long. The term *turn* is used to refer to the right to contribute, rather than to the contribution produced during that turn. One may make a contribution without having a turn and one may have the turn without using it for an

active contribution, as demonstrated in the example below, in which B's first contribution involves giving positive feedback without having the turn (square brackets indicate overlap) and his second contribution involves being silent and doing nothing while having the turn. [50]

A: look ice cream [would] you like an ice cream
B1: [yeah]
B2: (silence and no action) [51]

Contributions, utterances and turns are not coded since they are obtainable directly from the Göteborg transcription standard – MSO.6 (Modified Standard Orthography, version 6). [52]

7.3 Coding related to Social activity and Communicative acts

7.3.1 Social activity

Each transcription is linked to a database entry and a header containing information on. [53]

(i) The purpose, function and procedures of the activity
(ii) The roles of the activity
(iii) The artefacts, i.e. objects. furniture, instruments and media of the activity
(iv) The social and physical environment
(v) Anonymous categorical data on the participants, such as age, gender, dialect and ethnicity. [54]

In addition, the major subactivities of each activity are given. [55]

7.3.2 Communicative Acts

Each contribution can be coded with respect to one or more communicative acts which can occur sequentially or simultaneously as in the following example from a travelling agency dialog, where the customers utterance *ja typ den: ä:{h} tredje fjärde <7 <8 april >7 / [3 nån]3 gång där > 8 <9 / >9 så billi{g}t [4 som möjli{g}t]4* has been coded with several communicative act labels both sequentially and simultaneously. [56]

$P6.1: / <5 <6 >5 >6 ja:	Hesitation +
$P6.2: typ den:	Initiated (answer(J4)/statement/ specification(J4))
$P6.3: ä:{h}	Hesitation
$P6.4: tredje fjärde <7 <8 april >7 / [3 nån]3 gång där > 8 <9 / >9	Continued(answer(J4)/Statement/ specification(J4))
$P6.5: så billi{g}t [4 som möjli{g}t]4	Statement/specification of price range/ Request for low price ticket

The communicative acts make up an extendible list, where often used types have been provided with definitions and operationalizations. Some often used types are the following. [57]

Request
Statement
Hesitation
Question
Answer
Specification
Confirmation
Ending interaction
Interruption
Affirmation
Conclusion
Offer [58]

7.3.3 Expressive and evocative functions

In accordance with Allwood (1976, 1978), each contribution is viewed as having both an *expressive* and an *evocative* function. These functions make explicit some of the features implied by the communicative act coding. The *expressive* function lets the sender express beliefs and other cognitive attitudes and emotions. What is "expressed" is made up of a combination of reactions to the preceding contribution(s) and novel initiatives. The *evocative* function is the reaction the sender intends to call forth in the hearer. Thus, the evocative function of a statement normally is to evoke a belief in the hearer, the evocative function of a question is to evoke an answer, and the evocative function of a request to evoke a desired action. For a discussion of the relations between these functions and Bühler's (1934) symptom, symbol and signal function as well as Austin's (1962) locutionary, illocutionary and perlocutionary functions see Allwood (1976, 1977 and 1978). [59]

Each contribution to a dialog is associated with the following default evocative functions, cf Allwood (1987). A contribution is intended to make the receiver: [60]

(i) have contact/continue (C),
(ii) perceive (P),
(iii) understand (U),
(ii) react in accordance with main evocative function (R). [61]

These four default "evocative functions" are connected with four default "expressive functions" which are default consequences of normal cooperative communication. Thus, each contribution, except possibly the first, is associated with the following 4 expressive functions which express an evaluation and reaction to the evocative functions of the preceding utterance. [62]

(i) ability and wish to continue (C)
(ii) ability and wish to perceive (P)
(iii) ability and wish to understand (U)
(iv) ability and wish to react in accordance with main evocative function (R) [63]

These functions can be expressed explicitly or implicitly. Implicitly, they are expressed by carrying out desired actions or by carrying out actions which presuppose a positive evaluation of the ability and wish to carry out the main evocative function (and usually the three CPU functions as well). [64]

In addition to the 4 default evocative and expressive functions attached to contributions/utterances there are then other default functions attached to moods as well as a list of non-default functions which often occur. [65]

Since perception and understanding mostly function as a means for the sharing of the expressive and evocative functions of each contribution, a cooperative response usually consists in one of the following responses, used separately or in combination: [66]

(i) overtly signaling the result of the listener's evaluation through the use of an explicit positive or negative feedback expression, such as a head nod, a head shake or a verbal expression like *m*, *what*, *yes*, *no* or *OK*, after a statement or request
(ii) direct verbal action, as when a question is answered
(iii) direct nonverbal action, as when a window is closed after a request to do so
(iv) implicitly accepting an evocative intention by contributing a response that implies acceptance, as when you accept a stated belief by exploring one of its consequences. [67]

Since the main thrust of a dialog revolves around evocative intentions which are aimed at achieving more than mere perception and understanding, a cooperative response that signals only perception and understanding usually occurs only in the following circumstances: [68]

(i) when a message can be perceived and understood but no commitment is made to its evocative function or a message cannot be perceived or understood. In the first case often low key feedback expressions like *m* or *well* are used and in the second we find instead negative feedback expressions such as *pardon* or *what*. [69]

For an illustration of how Expressive and Evocative Functions are coded, see Table 1 below. [70]

7.3.4 Obligations

If the dialog and communication is to be cooperatively pursued, whether it be in the service of some activity or not, they impose certain obligations on both sender and receiver. With regard to both expressive and evocative functions, the sender should take the receiver's perceptual, cognitive and behavioral ability into consideration and should not mislead, hurt or unnecessarily restrict the freedom of the receiver. The receiver should reciprocate with an evaluation of whether he/she can hear, understand and carry out the sender's evocative intentions and signal this to the interlocutor. [71]

The sender's and receiver's obligations can be summarized as follows (see also Allwood 1994): [72]

Sender:
1. *Sincerity*:
The sender should, unless she/he indicates otherwise, have the attitude normally associated with a particular type of communicative act, e.g. statement - belief, request - desire (cf. Allwood 1976). [73]
2. *Motivation*:
Normally, communicative action, like other action should be motivated. [74]
3. *Consideration*:
If communicative action is to be cooperative and ethical it must take the other person into cognitive and ethical consideration. [75]

Receiver:
1. *Evaluation*:
The receiver should evaluate the preceding utterance with regard to whether he/she can continue the interaction, perceive and understand and accept its main evocative intention. [76]
2. *Report*:
After having evaluated, the receiver should report the result verbally or nonverbally. [77]
3. *Action:*
In some activities and roles, a positive evaluation of the ability to carry out the main evocative intention also obligates the listener to carry out the action associated with this intention. [78]

7.4 Communication management related coding

7.4.1 Introduction

The term "communicative management" refers to means whereby speakers can regulate interaction or their own communication. There are 3 coding schemas related to communication management (cf Nivre, Allwood & Ahlsén 1999). [79]
1) Feedback coding
2) Turn and sequence management coding and
3) Own Communication Management (OCM) coding[80]

7.4.2 Feedback coding schema

A feedback unit can be described as "a maximal continuous stretch of utterance (occurring on its own or as part of a larger utterance), the primary function of which is to give and/or elicit feedback concerning contact, perception, understanding and acceptance of evocative function" (Allwood, 1988). All feedback units are coded with respect to "Structure", "Position/Status" and "Function". Coding structure means coding grammatical category (part of speech, phrase or sentence) and also "structural operations". "Structural operations" is subdivided into "phonological", "morphological" and "contextual" operations, each of which have different values as below: [81]

Tags	Values
phon_op	lengthening
	cont_redupl(fricative)
	cont_redupl(stop)
	vowel_addition
	truncation(pure)
	ingressive
	prosody
morph_op	reduplication
	derivation
	compounding
	reduction
context_op	repetition
	reformulation

When coding Position/Status one is coding the position of the feedback unit in the utterance. This could be coded as "single" (the unit constitutes an entire utterance by itself), "initial", "medial" or "final" in the utterance. [82]

"Function" coding is divided into coding of "function type" and "attitudes". "Function type" indicates whether the feedback unit is giving or eliciting feedback or both giving and eliciting feedback. [83]

Coding of "CPU attitudes" and "acceptance of evocative function" at the present stage overlaps with coding Communicative Acts, Expressive and Evocative function and Obligations. Work on eliminating this is in progress. [84]

7.4.3 Turn and sequence management coding

Turn and sequence management coding encompasses the following phenomena: [85]

(A) Overlap and interruption: Overlap is coded in the transcriptions and can be extracted automatically. Interruption is a code for those overlaps which aim/at or succeed in changing the topic or taking away the floor from another speaker. [86]

(B) Intended recipient: This type of coding has 4 self explanatory values

(i) particular participant
(ii) particular group of participants
(iii) all participants
(iv) no participant (talking to oneself). [87]

(C) Marking of the opening and closing of subactivities and/or the interaction as a whole. [88]

Some turn and sequence related functions can be derived from other parts of the coding schema. For example, Turn acceptance - rejection is derived from communicative acts and expressive function. [89]

Many sequences of communicative acts are derived from the exchange types generated by the communicative acts coding and from the list of subactivities given by the initial activity description. [90]

7.4.4 OCM coding schema

OCM means "Own Communication Management" and stands for processes that speakers use to regulate their own contributions to communicative interaction. OCM function coding concerns classifying whether the OCM unit is: [91]

- choice related - helps the speaker to gain time for processes concerning continuing choice of content and types of structural expressions, or: [92]
- change related - helps the speaker to change already produced content, structure or expression. [93]

OCM units are also coded with respect to structure of the OCM related expression. This structure can be divided into "basic OCM features", "basic OCM operations" and "complex OCM operations". Pauses, simple OCM expressions such as hesitation sounds etc and explicit OCM phrases count as basic OCM features. Basic OCM operations are: "lengthening of continuants", "self interruption" and "self repetition". The category "Complex OCM operations" stands for different ways to modify the linguistic structure. These operations always involve self interruption, often together with a number of other basic OCM structures. The OCM coding schema is described in Allwood, Ahlsén, Nivre & Larsson (1997). [94]

7.5 Grammatical coding

There are also ways of coding grammatical structure. One of these is an automatic coding of parts of speech. Another is a coding of "maximal grammatical units". For a description see below. [95]

7.5.1 Maximal Grammatical Units coding schema

The Maximal Grammatical Units coding schema is described in Allwood, Björnberg & Weilenmann (1999). When coding Maximal Grammatical Units, one should primarily try to find as large units as possible, the largest unit being complete sentences. Sentences are subclassified by using the schema "**sentences**". In spoken language, there are many utterances that are not sentences, so secondarily, one should try to find complete phrases, which should be coded in the schema "**phrases**". If it isn't possible to find either complete sentences or complete phrases, single words should be coded by parts of speech in the schema "Parts of speech". Each one of the three mentioned schemes contains different categories as seen below. [96]

7.5.2 Sentences

The coding schema "sentences" consists of the following categories:

Tag	Type of sentence
declarative_s	Declarative sentence
exclamative_s	Exclamative sentence
imperative_s	Imperative sentence
disj_question	Disjunct question
wh_question	Wh-question
yes/no_question	Yes/no-question

All complete sentences are coded in this scheme. If the sentence contains pauses, hesitation sounds, repeats etc, these should not be coded in this scheme (but in the OCM scheme) and the sentence should still be coded as a complete sentence. Indirect speech is also considered as part of the sentence. [97]

7.5.3 Phrases

The coding schema "phrases" contains the following categories:

Tag	Type of phrase
adjp	Adjective phrase
advp	Adverb phrase or adverbial clause
conj	Conjunction phrase
fbp	Feedback phrase (see below)
np	Nominal phrase
nump	Numerical phrase
pp	Prepositional phrase
subordinate_clause	Subordinate clause
vp	Verb phrase

For something to be considered and coded as a feedback phrase, the phrase must contain a primary feedback word. These are: "ja", "jo", "nej", "nä", "nja", "m", "okej" and "va". [98]

7.5.4 Parts of Speech

The coding scheme "Parts of speech" contains the following categories:

Tag	Parts of speech
adj	adjective
adv	adverb
art	article
conj	conjunction
fb	feed-back-word (see above)
inf	infinitive marker
interj	interjection
n	noun

num	numeral
ocm	OCM word (see below)
part	particles
pron	pronoun
v	verb

OCM word*s* are certain words that always or often have OCM function, for example hesitation sounds like "eh" and "m". [99]

8. Relation to other types of transcription and coding

We have compared the transcription standard and types of coding used in Göteborg to some other types of transcription and coding. [100]

8.1 Transcription

In Allwood, Abelin and Grönqvist (1998), MSO is compared with transcriptions formats used at the department of phonetics, Lund University and at Telia - the former Swedish national telephone company. Both Lund and Telia use a word based, time coded format with some extra annotations. The report also compares MSO with the standard orthography format used at the department of computer science of Linköping University. In the report, we describe a computer based translation of all three formats to MSO. [101]

In addition to this report, we have compared MSO to the standard of transcription used in CA (Conversation Analysis) as it is available through the journal Discourse Analysis. Also in this case, we have been able to provide an automatic way of converting MSO-based transcriptions to CA-based transcriptions and vice versa. [102]

8.2 Coding

We have also compared the coding schemas used in Göteborg with two related coding schemas, ELIN and LINLIN. developed at the department of computer science, Linköping university (Dahlbäck & Jönsson 1998) but used both at the department of phonetics, Lund university (ELIN) and at the department of computer science, Linköping (ELIN and LINLIN). The comparison in reported in Abelin & Allwood, (1998), and Björnberg (1999). In the report we find that the three coding schemas code partly different aspects of dialog but there is a large overlap. The schemas produced in Göteborg are the most detailed, while the LINLIN Schemas are the least detailed with ELIN in the middle. In some cases differences are almost exclusively terminological, as when "discourse opening, continuation and ending" are used in LINLIN and "dialog opening, continuation and encoding" are used in ELIN. Some substantial differences are that LINLIN has no detailed speech act coding while ELIN has a more detailed schema and the Göteborg schema is even more detailed, in addition to being open for addition of new speech act labels. The obligation aspect of communicative acts is coded in Göteborg but not in ELIN or LINLIN. ELIN and LINLIN have however, done more work on coding related to topic and knowledge sources. [103]

Finally, some differences are more subtle as when ELIN "repairs" seem to cover both "own repairs" and "other-repairs", while these functions are separated in Göteborg. [104]

Another complicated difference concerns what in Göteborg are called the "expressive" and "evocative" functions of an utterance and in LINLIN "initiatives" and "responses". In Göteborg, these are seen as aspects of every utterance while they in LINLIN are seen as utterance types, in accordance with the main functions of specific utterances. [105]

A possible unification of all 3 schemas would probably be possible for more than 80% of the codes but would result in a very large number of coding categories. A solution to this problem would be to subdivide the codes into different types with codes on different levels of abstraction and specificity. [106]

9. Some results

The establishment of the Göteborg spoken language corpus has already resulted in many different kinds of analysis. The analyses have been both of an automatic-quantitative and manual-qualitative kind, sometimes done separately and sometimes done in combination. [107]

The corpus has been used extensively by both undergraduate and graduate students as a resource for their course papers and probably about 40 student papers have been written using the corpus as a basis. Material in the corpus has been the basis for four Ph.D. theses. The corpus has resulted in a frequency dictionary (Allwood 1996 and later editions), where spoken and written language are systematically compared with regard to words, collocations and parts of speech. This is the first dictionary of this type for Swedish and it is still possibly unique also in comparison to other languages. Work on the corpus has also resulted in papers concerned with developing tools, coding schemas, transcription formats and automatic measures, cf. e.g. Allwood and Hagman (1994). Several published articles have been written on the basis of the corpus, for example, Allwood (1999). [108]

10. Conclusions and Future Directions

In this paper we have described work done at the department of linguistics, Göteborg University to collect, transcribe and store spoken language material. We have also described some of the tools we have developed in order to aid work on analyzing the data both automatically and manually. Finally, we have described some or the results obtained so far. [109]

Future work will include incremental expansion of the corpus both to obtain data from new social activities and in order to equalize the size of the material from different activity types. We will also be making increased efforts to make the corpus more multimodal by making the audio and video recordings on which the transcriptions are based more available. [110]

Work on tools for analyzing the corpus will continue. The most immediate goal is to complete MULTITOOL which will hopefully give us a better possibility of working

with multimodal data. Similarly, work on qualitative and quantitative analysis will be continued. An ambitious goal is to work toward a grammatical description of spoken language and toward a systematic description (perhaps not a grammar) of multimodal face-to-face communication. [111]

11. References

Abelin, Åsa & Allwood, Jens (1998). *Jämförelse mellan OCM kodningsstandard och Robert Eklunds disflueringskodningsstandard.* Internal Report. Department of Linguistics,. Göteborg University

Allwood, J. (1976). Linguistic Communication as Action and Cooperation. *Gothenburg Monographs in Linguistics 2.* Göteborg University, Department of Linguistics.

Allwood, Jens (1977). A Critical Look at Speech Act Theory. In Dahl (Ed.). Logic, *Pragmatics and Grammar*, Lund, Studentlitteratur, pp. 53-69.

Allwood, Jens (1978). On the Analysis of Communicative Action. In Brenner (Ed.) *The Structure of Action*, Basil Blackwell, Oxford.

Allwood, Jens (1987). A Semantic Analysis of Understanding. In V. Rosén (Ed.) *Papers from the Tenth Scandinavian Conference of Linguistics,* University of Bergen, Dept of Linguistics and Phonetics.

Allwood, Jens (1988). Feedback in Adult Language Acquisition. Final Report II. *Ecology of Adult Language Acquisition* (ESF).

Allwood, Jens, Nivre, Joakim & Ahlsén, Elisabeth (1990). Speech Management: On the Non-Written Life of Speech. *Nordic Journal of Linguistics*, 13, 3-48.

Allwood, Jens (1994). Obligations and Options in Dialogue, *Think*, Vol 3, May, ITK, Tilburg University.

Allwood, Jens & Hagman, Johan (1994). Enkla mått på samtal. In Gregersen, F. And Allwood, J. (Eds.) Spoken Language, *Proceedings of the XIV Conference of Scandinavian Linguistics.*

Allwood, Jens (1995). An Activity-based Approach to Pragmatics. Gothenburg. *Papers in Theoretical Linguistics 76*. Göteborg University, Department of Linguistics. Forthcoming in Bunt and Black (Eds.) Approaches to Pragmatics.

Allwood, Jens (Ed) (1996 and later editions). *Talspråksfrekvenser*, Ny och utvidgad upplaga. Gothenburg Papers in Theoretical Linguistics S21. Göteborg University: Department of Linguistics.

Allwood, Jens, Ahlsén, Elisabeth, Nivre, Joakim & Larsson, Staffan (1997). *Own communication management.* Göteborg University, Department of Linguistics.

Allwood, Jens, Abelin, Åsa & Grönqvist, Leif (1998). Kort beskrivning och jämförelse av transkriptionssystem från Lund, Telia, Linköping och Göteborg http://www.ling.gu.se/~leifg/doc/jfresle transkriptionssystem.pdf (pdf).

Allwood, Jens (1999). Some Frequency based Differences between Spoken and Written Swedish. In *Proceedings of the 16th Scandinavian Conference of Linguistics*, Turku University, Department of Linguistics.

Allwood, Jens & Björnberg, Maria (1999). *Coding Schemas within the SDS Project – A Comparison* http://www.ling.gu.se/SLSA/Documents/coding_schema_comparison.ps

Allwood, Jens, Björnberg, Maria & Weilenmann, Alexandra (1999). *Kodning av maximala grammatiska enheter- Manual,* Göteborg University, Department of Linguistics.

Austin, John Langshaw (1962). *How to Do Things with Words*. Harvard University Press.

Bühler, Karl (1934). *Sprachtheorie*. Jena: Fischer.

Dahlbäck, Nils & Jönsson, Arne (1998). *A coding manual for the Linköping dialogue model*. Department of Computer and Information Science, Linköping University.

Grice, H. Paul (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.) *Syntax and Semantics* Vol. 3: Speech Acts. New York: Seminar Press, 41-58.

Nivre, Joakim (1999). Transcription Standard. Version 6. Göteborg University. Department of Linguistics.

Nivre, Joakim, Tullgren, Kristina, Allwood, Jens, Ahlsén, Elisabeth, Holm, Jenny, Grönqvist, Leif, Lopez-Kästen, Dario & Sofkova, Sylvana. Towards multimodal spoken language corpora: TransTool and SyncTool. *Proceedings of ACL-COLING 1998*, June 1998. http://www.ling.gu.se/~leifg/doc/COLING98.pdf

Nivre, Joakim & Grönqvist, Leif (1999) Tagging a corpus of Spoken Swedish. Forthcoming in *International Journal of Corpus Linguistics*. http://www.ling.gu.se/SLSA/Documents/sot.ps

Nivre, Joakim, Allwood, Jens & Ahlsén, Elisabeth (1999). *Interactive communication management - Coding manual V1.0.* Göteborg University, Department of Linguistics.

Van Hout, Roeland & Rietveld, Toni (1993) *Statistical Techniques for the Study of Language and Language Behaviour*. Berlin & New York: Mouton de Gruyter.

Author Descriptions

Jens Allwood is since 1986 professor of linguistics at the Department of Linguistics at Göteborg University. He is also director of the interdisciplinary cognitive science center SSKKII at the same university. His research primarily includes work in semantics and pragmatics. He is also investigating spoken language interaction from several perspectives, eg corpus linguistics, computer modelling of dialog, sociolinguistics and psycholinguistics as well as intercultural communication. Presently he is heading projects concerned with the semantics of spoken language phenomena, multimodal communication, cultural variation in communication and the influence of social activity on spoken language.
phone: +46-31 773 1876
e-mail: jens@ling.gu.se

Leif Grönqvist, M.Sc in computing science, and two years of studies in mathematics and physics, has been working at the Department of Linguistics, Göteborg University in various research projects since 1994. The latest project, "A Platform for Multimodal Spoken Language Corpora" included development of a tool for multi-modal corpus studies. He has also been working at the Informatics Department in "The Internet Project" and in a small consulting company with information retrieval in large free text medical databases.
phone: +46- 31 773 1177
e-mail: leifg@ling.gu.se

Maria Björnberg is a senior student of the master's program in computational linguistics, Göteborg University. She has worked as a part-time assistant. She has been involved in administrating the spoken language corpus, transcriptions and coding work as well as in writing manuals for coding.
phone: +46-73 773 4446
e-mail: mariab@ling.gu.se

Elisabeth Ahlsén is professor of neurolinguistics and has worked for many years in research involving disability and multimodal communication, both with respect to face-to-face-communication and in relation to alternative and augmentative communication using computer support. Some of her relevant research projects deal with "intermodal translation", "gestures and meaning", and "communication involving non-speaking children".
phone: +46- 31 773 1923
e-mail: eliza@ling.gu.se

Cajsa Ottesjö
PhD student at the Department of Linguistics, Göteborg University. She is interested in spoken language interaction, especially the use of particles in interaction.
phone: +46-73 773 4446
e-mail: cajsao@ling.gu.se

Authors' mailing address:
Dept of Linguistics
Göteborg University
Box 200
405 30 Göteborg, Sweden.