FQS FORUM: QUALITATIVE SOCIAL RESEARCH SOZIALFORSCHUNG

Volume 1, No. 3, Art. 9
December 2000

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

Jens Allwood, Maria Björnberg, Leif Grönqvist, Elisabeth Ahlsen & Cajsa Ottesjö

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

Abstract: This paper summarizes work on spoken language at the Department of Linguistics Göteborg University. In addition to describing the recordings contained in the Spoken Language Corpus of Swedish at Göteborg University, we discuss the standard of transcription (MSO) which is used in creating the transcriptions, as well as some types of quantitative and qualitative analysis that have been done. Finally, we describe the computer tools that have been developed to support transcription, coding and analysis and briefly mention some of the results which have been obtained.

Table of Contents

1. Corpus Description

The Spoken Language Corpus of Swedish at Göteborg University is an incrementally growing corpus of spoken language samples from several languages which presently consists of 1.26 million words from about 25 different social activities. Because 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 in order to facilitate research that will provide a more complete understanding of the role of language and communication in human social life. [1]

The corpus consists of about 50% audio and 50% video/audio (Umatic, VHS or BetaCAM) recordings of naturalistically occurring interactions from as long ago as the early 1980s. These recordings have generally been made on an as-needed basis for various student and faculty projects and as student course assignments. Students are still actively encouraged to look for new types of activities to be recorded. Additional recordings have been added to the corpus as funding has become available. There are several possible formats for storage, including analog video, digital video and MPEG. In order to preserve the recordings, tapes are being copied to newer tapes while simultaneously being digitized using CDs with MPEG compression. Each CD contains both transcriptions and recordings[1]. Over and above this we also work with other spoken language corpora sometimes collected by other teams (see table 1).

[1] Storage formats:

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.

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

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 CDs with up to 60 minutes per disc compared to three minutes in the DV format.

- Göteborg Spoken Language Corpus (Kernel Corpus—adult first-language Swedish), 1.2 million words
- Adult language learners of Swedish
- Speakers with aphasia
- Child language corpus (Swedish and Scandinavian), 0.5 million words including those of adult interlocutors
- Non-Swedish adult spoken language corpus

  Chinese (70 000 words)
  Bulgarian (25 000 words)
  Arabic
  English (10 000 words), British National Corpus
  Finnish
  Italian (3000 words)
  Norwegian (140 000 words)
  Spanish
- Wizard-Of-Oz Corpus, Bionic (types of human–computer interaction)
- Intercultural communication corpus

Table 1: Spoken language corpora at Göteborg University (Some of the corpora are recorded in more than one medium. Word counts are not currently available for all languages) [2]

As can be seen, the largest corpus is the *Kernel Corpus* of adult first-language Swedish speakers. This is the corpus we will focus on in this article. 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 possible. The limitations which exist in our ability to create subcorpora are dependent on the fact that we do not always have the pertinent information about individual speakers. [3]

In Table 2, basic data on this corpus is presented. The first column labels the type of social activity recorded. The second column lists how many separate recordings of each activity type exist in the corpus. The number of recordings usually corresponds to the number of instances of the activity type recorded. The third column gives the number of speakers the activity type has on average. The fourth column tells 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. The discrepancy between the number of speakers and the number of sections in this example is due to several passengers talking to the driver in one section. Column five gives information about word tokens as well as about pauses and comments, while column six only includes words actually uttered in the recording. Finally, column seven 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.

Activity Type	**Recordings**	**Speakers (average)**	**Sections**	**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
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
Medical Consultation	15	2.3	198	24 916	24 450	1:47:25
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: The corpus of adult first-language Swedish at Göteborg University [4]

2. Description of Modified Standard Orthography (MSO)—The Corpus Transcription Standard

The recordings have been transcribed according to the Modified Standard Orthography (MSO) transcription standard. This standard was developed internally by the Department of Linguistics at Göteborg University and standardized for the first time in 1983 (for the version valid at present, cf. NIVRE 1999). It 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 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. It can perhaps most rapidly be explained through exemplification. Consider the following example:

§1. Small talk	
$D: säger du de{t} ä{r} de{t} ä{r} de{t} så besvärlit då	$D: oh I see is it it is so troublesome then
$P: ja ja	$P: yes yes
$D: m // ha / de{t} kan ju bli så se{r} du	$D: m // yes / it can be that way you see
$P: < jaha >	$P < yes >
@ <ingressive>	@ <ingressive >
$D: du ta{r} den på morronen	$D: you take it in the morning
$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	$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 {j}a]1	$D: [1 yes]1
$D: [2 nä]2	$D: [2 no]2

Table 3: Transcription according to the MSO standard with translation [5]

This example contains the most important properties of the transcription standard:

a. Section boundaries paragraph sign (§). These divide a longer activity up into subactivities. A doctor-patient interview can, for example have the following subactivities. (a) greetings and introduction, (b) reason for visit, (c) investigation and (d) prescribing treatment.
b. Words and space between words.
c. Dollar sign ($) followed by a capital letter, followed by a colon (:) to indicate a new speaker and a new utterance.

d. 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.
e. Double slash (//) to indicate pauses. Slashes /, // or /// are used to indicate pauses of different length.
f. 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.
g. Capital letters to indicate contrastive stress.
h. Overlaps are indicated using square brackets ([ ]) with indices which allow disambiguation if several speakers overlap simultaneously. [6]

Following GRICE (1975), ALLWOOD, NIVRE and AHLSÉN (1990) and ALLWOOD (2000a), the basic units of dialogue 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 choice of contribution involves giving positive feedback without having the turn (square brackets indicate overlap) and his second choice of contribution involves being silent and doing nothing while having the turn.

A: look ice cream [would] you like an ice cream

B1: [yeah]

B2: (silence and no action) [7]

Contributions, utterances and turns are not coded after the transcription has been made (see section 3.1) since they are decided on in the process of transcription using the Göteborg transcription standard MSO.6 (Modified Standard Orthography, version 6). [8]

In ALLWOOD, ABELIN and GRÖNQVIST (1998), MSO is compared with transcription formats used by the Department of Linguistics and Phonetics at Lund University and by 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 by the Department of Computer and Information Science at Linköping University. In the report, we describe a computer-based translation of all three formats to MSO. [9]

In addition to this report, we have compared MSO to the standard of transcription used in Conversation Analysis (CA) as it is available through the journal *Research on Language and Social Interaction*. As we did for the Lund, Linköping and Telia formats, we have also been able to provide an automatic way of converting MSO-based transcriptions to CA-based transcriptions and vice versa. [10]

3. Qualitative and Quantitative Analyses

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. 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. Several published articles have been written on the basis of the corpus (e.g. ALLWOOD 1999). 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. The book contains word frequencies both for the words in MSO format and for those in standard orthographic format. There are statistics on the parts of speech represented in the corpus, based on an automatic probabilistic tagging, yielding a 97% correct classification. 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 & HAGMAN 1994). [11]

3.1 Overview of qualitative analyses with Göteborg Coding Schemas

Because the corpus has been the basis for work using various kinds of manual coding, qualitative analysis in Göteborg has often resulted in the development of new coding schemas. This coding is done after the data has been transcribed and the type of coding performed depends on the researcher's needs. The following provides an overview of the Göteborg coding schemas:

1. Social activity and Communicative act-related coding
   - Social activity
   - Communicative acts
   - Expressive and Evocative functions
   - Obligations
2. Communication management-related coding
   - Feedback
   - Turn and sequence management
   - Own Communication Management

3. Grammatical coding
   - Parts of speech (automatic, probabilistic)
   - Maximal grammatical units
4. Semantic coding [12]

3.1.1 Coding related to social activity and communicative acts

3.1.1.1 Social activity coding

Each transcription is linked to a database entry and a header containing information on

a. the purpose, function and procedures of the activity,
b. the roles of the activity,
c. the artifacts, i.e. objects, furniture, instruments and media of the activity,
d. the social and physical environment and
e. anonymous categorical data on the participants, such as age, gender, dialect and ethnicity. [13]

In addition, the major subactivities of each activity are given. [14]

3.1.1.2 Communicative acts coding

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 dialogue. The customer's 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.

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

Table 4: Example of sequential and simultaneous (P6.5) coding of communicative acts of 1 utterance (The codings are based on a division of utterance p. 6. into smaller parts) [15]

The communicative acts make up an extensible list, where often-used types (listed below) have been provided with definitions and operationalizations (cf. ALLWOOD et al. 2000).

- Request
- Statement
- Hesitation
- Question
- Answer
- Specification
- Confirmation
- Ending interaction
- Interruption
- Affirmation
- Conclusion
- Offer [16]

<u>3.1.1.3 Expressive and evocative functions coding</u>

In accordance with ALLWOOD (1976, 1978,1987) each contribution is viewed as having both an *expressive* and an *evocative* function. These functions make some of the features implied by the communicative act coding explicit. 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 elicit from the recipient. 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 is 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). [17]

Each contribution to a dialogue is associated with the following default (i.e. they are assumed unless explicitly denied) evocative functions (cf. ALLWOOD 2000b). A contribution is intended to make the receiver:

a. have contact/continue (C),
b. perceive (P),
c. understand (U) and
d. react in accordance with main evocative function (R). [18]

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 four expressive functions which express an evaluation and reaction to the evocative functions of the preceding utterance:

a. ability and wish to continue (C),
b. ability and wish to perceive (P),
c. ability and wish to understand (U) and
d. ability and wish to react in accordance with main evocative function (R). [19]

These functions can be expressed explicitly or implicitly. They are expressed implicitly by carrying out desired actions or by carrying out actions which presuppose a positive evaluation of both the ability and the wish to carry out the main evocative function (and usually the three CPU functions as well). [20]

In addition to the four default evocative and expressive functions attached to contributions/utterances, there are other default functions attached to moods, as well as a list of non-default functions which often occur. [21]

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

a. overtly signaling the result of the listener's evaluation through the use of an explicit positive or negative feedback expression, such as a head node, a head shake, or a verbal expression like m, what, yes, no, or OK, after a statement or request,
b. direct verbal action, as when a question is answered,
c. direct nonverbal action, as when a window is closed after a request to do so, or
d. 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. [22]

Since the main thrust of a dialogue 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: when a message can be perceived and understood but no commitment is made to its evocative function, or when a message cannot be perceived or understood. In the first case, low-key feedback expressions like *m* or *well* are often used and in the second we find instead negative feedback expressions such as *pardon* or *what*. [23]

3.1.1.4 Obligations coding

If the dialogue and communication are 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. [24]

The sender's and receiver's obligations can be summarized as follows (see also ALLWOOD 1994, 2000b):

> **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).
>
> *2. Motivation*
>
> Normally, communicative action, like other action, should be motivated.
>
> *3. Consideration*
>
> If communicative action is to be cooperative and ethical, it must take the other person into cognitive and ethical consideration. [25]
>
> **Receiver**
>
> *1. Evaluation*
>
> The receiver should evaluate the preceding utterance with regard to whether he/she can continue the interaction and perceive, understand and accept its main evocative intention.
>
> *2. Report*
>
> After having evaluated the utterance, the receiver should report the result verbally or nonverbally.
>
> *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. [26]

3.1.2 Coding related to communication management

3.1.2.1 Introduction

The term "communicative management" refers to means whereby speakers can regulate interaction or their own communication. There are three coding schemas related to communication management (cf. NIVRE, ALLWOOD & AHLSÉN 1999).

a. Feedback coding
b. Turn and sequence management coding
c. Own Communication Management (OCM) coding [27]

3.1.2.2 Feedback coding

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" (phon_op), "morphological" (morph_op) and "contextual" (context_op) operations, each of which have different values. Examples of these values include vowel lengthening under phonological operations, reduplication (e.g. saying "Yes, yes") under morphological operations and repetition and reformulation (i.e. referring back to your own or the other speaker's utterance) under contextual operations.

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

Table 5: Values for phonological, morphological and contextual operations used in coding feedback [28]

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. [29]

"Function" coding is divided into coding of "function type" and "attitudes." "Function type" indicates whether the feedback unit is either giving or eliciting feedback or both giving and eliciting feedback. [30]

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. [31]

3.1.2.3 Turn and sequence management coding

Turn and sequence management coding encompasses the following phenomena (cf. ALLWOOD & BJÖRNBERG 2000):

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.
b. Intended recipient: This type of coding has four self-explanatory values.
    - particular participant
    - particular group of participants
    - all participants
    - no participant (talking to oneself)
c. Marking of the opening and closing of subactivities and/or the interaction as a whole. [32]

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. 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. [33]

3.1.2.4 Own Communication Management (OCM) coding

OCM means "Own Communication Management" and stands for processes that speakers use to regulate their own contributions to communicative interaction (ALLWOOD, AHLSÉN, NIVRE & LARSSON 1997). OCM function coding concerns classifying whether the OCM unit is:

- choice related-helps the speaker to gain time for processes concerning continuing choice of content and types of structural expressions, or
- change related-helps the speaker to change already produced content, structure or expression. [34]

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. [35]

3.1.3 Grammatical coding

There are also ways of coding grammatical structure. One of these is an automatic coding of parts of speech. Another is a manual coding of "maximal grammatical units." [36]

3.1.3.1 Parts of speech coding

One of the ways of coding grammatical structure is an automatic, probabilistic coding of *parts of speech* (see also Parts of speech in section 3.2). This coding scheme contains the following categories:

Tag	**Part of Speech**
adj	Adjective
adv	Adverb
art	Article
conj	Conjunction
fb	feedback word
interj	Interjection
n	Noun
num	Numeral
ocm	OCM word
part	Particles
pron	Pronoun
v	Verb

Table 6: The parts of speech used in the automatic tagging of the Göteborg corpus (For additional information see feedback word[2] and OCM word[3] [37]

[2] The part of speech "feedback words" (and also the type of phrase "feedback phrase", see Note 5 below) includes primary feedback words like: "ja", "jo", "nej", "nä", "nja", "m", "okej", and "va".

[3] OCM (Own Communication Management) words are certain words that always or often have OCM function, for example hesitation sounds like "eh" and "m" (see section 3.1.4 on OCM).

3.1.3.2 Maximal Grammatical Units coding

The Maximal Grammatical Units coding schema is described in ALLWOOD, BJÖRNBERG & WEILENMANN (1999). When coding Maximal Grammatical Units, one should primarily try to find units as large as possible, the largest unit being complete sentences. Sentences are subclassified by using the schema "sentences".[4] 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".[5] If it is not possible to find either complete sentences or complete phrases, single words should be coded by parts of speech in the schema "parts of speech" (see section 3.4.1). Each one of the three mentioned schemes contains different categories. [38]

3.1.4 Some comparisons

We have also compared the coding schemas used in Göteborg with two related coding schemas, ELIN and LINLIN, developed at the Department of Computer and Information Science, Linköping University (DAHLBÄCK & JÖNSSON 1998), but used both by the Department of Linguistics and Phonetics at Lund University (ELIN) and by the Department of Computer and Information Science at Linköping (ELIN and LINLIN) (ABELIN & ALLWOOD 1998, ALLWOOD & BJÖRNBERG 1999). In the report we find that the three coding schemas code partly different aspects of dialogue but that there is a large overlap. The schemas produced in

[4] 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.

[5] 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 Note 2)
np	Nominal phrase
nump	Numerical phrase
pp	Prepositional phrase
subordinate_clause	Subordinate clause
vp	Verb phrase

Göteborg are the most detailed, while the LINLIN Schemas are the least detailed with ELIN being in the middle. In some cases differences are almost exclusively terminological, as when "discourse opening, continuation and ending" are used in LINLIN and "dialogue 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 the 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. Some differences are more subtle, as when ELIN "repairs" seems to cover both "own repairs" and "other-repairs", while these functions are separated in Göteborg. [39]

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. [40]

A possible unification of all three 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. [41]

3. 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:

a. *Volume*: Volume comprises measures of the number of words, pauses, stresses, overlaps, utterances, turns relative to speaker, activity and subactivity.

b. *Ratios*: Various ratios can then be calculated based on the volume measures. For example:

   MLU[6] = words/utterances

   % pauses = pauses/(words+pauses)*100

   % stress = stressed words/words*100

   % overlap = overlapping words/words*100

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

c. *Special measures*: One example of a special type of measure is "vocabulary richness" as measured through type/token or through "theoretical vocabulary" (cf. VAN HOUT & RIETVELD 1993). Another measure we have constructed is

[6] Mean Length per Utterance

"stereotypicality" which looks at how often words and phrases are repeated in an activity.

d. *Lemma*: We also implemented a simple stemming algorithm which enables us to collect regularly inflected forms together with their stem.

e. *Parts of speech*: Parts of speech are assigned using a computer program that tags parts of speech based on statistical probability (developed by Viterbi) 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.

f. *Collocations*: All speakers, activities and subactivities can be characterized in terms of their most frequent collocations.

g. *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.

h. *Similarities*: Similarities between activities are captured by looking at the extent to which words and collocations are shared between activities. [42]

4. Tools Which Have Been Developed

Several tools have been developed internally for using the corpus, including a browser for searching according to various criteria, tools for coding the transcriptions and for calculating frequencies and a tool for synchronizing a recording with different analyses of it for concurrent utilization. [43]

4.1 TransTool[7]

TransTool is a computer tool for transcribing spoken language in accordance with the MSO transcription standard (NIVRE 1999). It is meant to facilitate and partially automate the task of a human transcriber by automatically adding certain standard elements, such as the header and by giving prompts when other elements are missing, for example closing brackets. This makes it much easier to keep track of indices for overlaps and comments (cf. NIVRE et al. 1998). [44]

4.2 The Corpus Browser

The Corpus Browser makes it possible to search for words, word combinations and phrases 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. [45]

[7] For additional information, see http://www.ling.gu.se/~sylvana/SLSA/TransTool.html [Broken link, FQS, December 2004].

4.3 TRACTOR[8]

TRACTOR is a coding tool which makes it possible to create new coding schemas and to 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 hard disk drive. This framework makes it easy to analyze the codings in a Prolog system[9]. [46]

4.4 Visualization of codings with FrameMaker[10]

This document describes a toolbox that makes it possible to visualize coding schemas and coding values by using colors, boldface, 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 the details that are not of interest. [47]

4.5 TraSA (Transcription Statistics with Automation)[11]

If you have a corpus transcribed according to the Göteborg Transcription Standard, using TraSA makes 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 and utterances and also theoretical vocabulary. No other tool makes it possible to partition a corpus and calculate all these measurements without requiring additional programming and statistical skills on the part of the researcher. [48]

4.6 SyncTool

SyncTool was developed for aligning transcriptions with their respective 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. This application later became the prototype for MuliTool (cf. NIVRE et al. 1998). [49]

[8] For additional information, see http://www.ling.gu.se/~sl/tractor.html.

[9] Prolog is a logic programming language. For additional information, see http://www.sics.se/isl/sicstus.html.

[10] For additional information, see http://www.ling.gu.se/~leifg/doc/kodvisualisering.pdf. Some corpus data on this site are password protected, interested readers must contact one of the authors to arrange access.

[11] For additional information, see http://www.ling.gu.se/~leifg/doc/trasa08e.pdf.

4.7 MultiTool[12]

MultiTool is an attempt to build a general tool for linguistic annotation and transcribing of dialogues, as well as browsing, searching and counting. The system can handle any number of participants, overlapping speech, hierarchical coding schemes, discontinuous coding intervals, relations and synchronization between codings and the media file. [50]

The fundamental idea is to collect all information in an internal state containing only codings and synchronizations[13]. These are the basic types of information the computer program requires. (For purposes of computer programming, transcriptions are considered as coding. For researchers using the audio/video recordings of the corpus, the transcriptions themselves are merely a coding of the fundamental data, also known as the recordings.) One important detail is that views pertaining to the same point in time can be synchronized to show the same sequence from different points of view whenever the user scrolls only in one of them. The internal state contains all the information so it is possible to have many different views of the same sequence of the dialogue.[14] Changes made in one view will immediately change the internal state and as a consequence the other views. [51]

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 types of views. [52]

[12] MultiTool as well as examples and The MultiTool User's Manual may be downloaded from http://www.ling.gu.se/projekt/multitool/ [Broken link, FQS, December 2004]. Some corpus data on this site are password protected, interested readers must contact one of the authors to arrange access.

[13] 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.

[14] The views in MultiTool:

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

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.

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.

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

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

5. Conclusions and Future Directions

Work on the Göteborg Spoken Language Corpus has resulted in new developments concerning ways to collect, transcribe, analyze and store spoken language material. It has also resulted in new tools used to manipulate the data in the corpus and in coding schemas used in the analysis based on the corpus. The work has been reported on in several papers, doctoral dissertations and a new comparative frequency dictionary of spoken and written Swedish. [53]

Future work will include incremental expansion of the corpus in order to both obtain data from new social activities and 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. [54]

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. As the tools and technology we use improves, so will our ability to extract more meaning from the corpus in order to increase our understanding of the role of language and communication in human social life. [55]

References

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

Allwood, Jens (1976). Linguistic communication as action and cooperation. Gothenburg Monographs in Linguistics 2. Göteborg: Göteborg University, Department of Linguistics.

Allwood, Jens (1977). A critical look at speech act theory. In Östen Dahl (Ed.), Logic, Pragmatics and Grammar (pp.53-69). Lund: Studentlitteratur.

Allwood, Jens (1978). On the analysis of communicative action. In Michael Brenner (Ed.), The Structure of Action (pp.168-191). Oxford: Basil Blackwell.

Allwood, Jens (1987). A semantic analysis of understanding. In Victoria Rosén (Ed.), Papers from the Tenth Scandinavian Conference of Linguistics (Vol.1, pp.37-51). Bergen: University of Bergen, Department of Linguistics and Phonetics.

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

Allwood, Jens (1994). Obligations and options in dialogue. Think, 3(1), 9-18.

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

Allwood, Jens (1999). Some frequency based differences between spoken and written swedish. In Proceedings of the 16th Scandinavian Conference of Linguistics. Turku: University of Turku, Department of Linguistics.

Allwood, Jens (2000a). An Activity Based Approach to Pragmatics. In Harry Bunt & Bill Black (Eds.), Abduction, Belief and Context in Dialogue; Studies in Computational Pragmatics. Amsterdam: John Benjamins.

Allwood, Jens (2000b). Expressive and Evocative Functions and Obligations. Coding Manual. Version 1.0. Göteborg 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 [Broken link, FQS, December 2004].

Allwood, Jens & Björnberg, Maria (2000). Addressee, Turn and Sequence Management. Coding Manual. Version 2.0. Göteborg University, Department of Linguistics.

Allwood, Jens & Hagman, Johan (1994). Enkla mått på samtal. In Frans Gregersen & Jens Allwood (Eds.), Four special sessions Spoken Language, Proceedings of the XIV Conference of Scandinavian Linguistics (pp.3-22). 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/jfrelse_transkriptionssystem.pdf.

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

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; Ahlsén, Elisabeth; Björnberg, Maria & Nivre, Joakim (2000). Speech Act Coding Manual. Version 1.0. Göteborg University, Department of Linguistics.

Allwood, Jens; Ahlsén, Elisabeth; Nivre, Joakim & Larsson, Staffan (1997). Own communication management. Göteborg: 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). Linköping: Linköping University, Department of Computer and Information Science.

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

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

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 [Broken link, FQS, December 2004].

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

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

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

Authors

Jens ALLWOOD is professor of linguistics at the Department of Linguistics at Göteborg University since 1986. 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, e.g. corpus linguistics, computer modeling of dialogue, 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.

Contact:

Jens Allwood

Department of Linguistics
Göteborg University
Box 200
405 30 Göteborg, Sweden

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 multimodal 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.

Contact:

Leif Grönqvist

Department of Linguistics
Göteborg University
Box 200
405 30 Göteborg, Sweden

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.

Contact:

Maria Björnberg

Department of Linguistics
Göteborg University
Box 200
405 30 Göteborg, Sweden

Phone: +46 - 73 773 4446

E-mail: mariab@ling.gu.se

Elisabeth AHLSEN 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."

Contact:

Elisabeth Ahlsen

Department of Linguistics
Göteborg University
Box 200
405 30 Göteborg, Sweden

Phone: +46 - 31 773 1923

E-mail: eliza@ling.gu.se

Cajsa OTTESJÖ is PhD student at the Department of Linguistics, Göteborg University. She is interested in spoken language interaction, especially the use of particles in interaction.

Contact:

Cajsa Ottesjö

Department of Linguistics
Göteborg University
Box 200
405 30 Göteborg, Sweden

Phone: +46 - 73 773 4446

E-mail: cajsao@ling.gu.se

Citation

Allwood, Jens; Björnberg, Maria; Grönqvist; Leif; Ahlsén, Elisabeth & Ottesjö, Cajsa (2000). The Spoken Language Corpus at the Department of Linguistics, Göteborg University [55 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(3), Art. 9, http://nbn-resolving.de/urn:nbn:de:0114-fqs000391.

Revised 8/2008