ISSN 0349-1021

GOTHENBURG PAPERS IN THEORETICAL LINGUISTICS

86.

TRANSLITERATION BETWEEN SPOKEN LANGUAGE CORPORA: MOVING BETWEEN DANISH BYSOC AND SWEDISH GSLC

Jens Allwood, Peter Juel Henrichsen, Leif Grönqvist,
Elisabeth Ahlsén and Magnus Gunnarsson

OCTOBER 2002

ABSTRACT

The paper discusses problems that arise in trying to transfer a spoken language corpus transcribed and formatted according to one standard into the standard and format of another corpus. Some of the problems that arise are related to the differences that exist between the standards and formats of different corpora. Other problems are related to human errors and lack of reliability in creating the transcriptions.

Although the discussion is based on transfer and transliteration between two specific corpora (the Swedish GSLC (Göteborg Spoken Language Corpus) and the Danish BySoc (By Sociolingvistik Corpus), we believe the discussion in the article documents and highlights problems of a general kind which have to be faced whenever spoken language corpora of different formats are to be compared.

Table of Contents

Transliteration between spoken language corpora
Moving between Danish BySoc and Swedish GSLC

Jens Allwood, Peter Juel Henrichsen, Leif Grönqvist,
Elisabeth Ahlsén and Magnus Gunnarsson

1. Introduction and purpose

The advent of corpus linguistics has meant that an increasing number of spoken language corpora are being established. These corpora are often created according to different standards. Since it is becoming increasingly desirable to be able to compare data from different corpora, the methodological problem of how to overcome differences in standards and formats needs to be solved. This report presents some of the problems and possible solutions.

The report contains a comparison of two major contemporary spoken language corpora of Scandinavian languages, the Danish BySoc (BySociolingvistik) corpus and the Swedish GSLC (Göteborg Spoken Language Corpus), each containing 1.3 million words of transcribed spoken interaction.

The purposes of the report are (i) to compare the transcription standards and formats of the two corpora, (ii) to document "translation" or rather "transliteration" programs for transferring transcriptions which have been made according to the DS - Dansk (Danish) Standard (the standard used in BySoc) to GTS (Göteborg Transcription Standard) the standard used in GSLC and from transcriptions which have been made according to GTS to DS, (iii) to generally discuss problems, choices and solutions for corpus transcription and transference between different formats for spoken language corpora.

The report, thus, discusses some of the general questions that have to be addressed in transcription and in doing transliteration between corpora transcribed according to different standards. Such questions include, for example, questions relating to lack of compatibility of standards and questions relating to actual translation from existing transcriptions, which have errors which may not be sanctioned by the standards but rather be caused by difficulties in carrying out what the standard demands. In particular, examples of transliteration originating from the use of two tools for doing automatic transfer, ds2gts (Dansk Standard to Göteborg Transcription Standard) (applied to transfer from BySoc to GSLC) and gts2ds (Göteborg Transcription Standard to Dansk Standard applied to transliteration from GSLC to BySoc) will be considered. Since the discussion is fairly specific, it should also be possible to use the report as a manual for making comparisons and transfers between GSLC and BySoc.

2. Similarities

Before we go into the differences between the two corpora, we want to point to the fairly extensive similarities between them. Both corpora consist mainly of spoken, in most cases fairly informal, spoken language interaction between two or more speakers. They have roughly the same size and the main parts were collected during the same period of time. They represent two Scandinavian languages with considerable similarities.

Both corpora are done according to standards which are a compromises between the three purposes of (i) representing spoken language with as much ecological validity as possible, (ii) creating a standard which supports transcription and is both rapid and reliable and (iii) making possible the use of computerized tools for analysis. This means that both corpora are transcribed into basically orthographic word representation, but that the transcription standards are specially designed for *spoken* language.

Neither of the two transcription standards uses any form of written punctuation.

3. Differences between the two corpora

3.1 Activities and speakers

The two corpora were collected for somewhat different purposes and this is reflected in the types of activities and speakers which are included.

The BySoc corpus was originally recorded and transcribed in 1986-1990 in the project BySoc (The Copenhagen Study in Urban Sociolinguistics). It consists of so called Labovian sociolinguistic interviews or conversations with about 80 citizens of Copenhagen, representing different ages, genders and social classes. They are informal conversations. The transcriptions were made in score format. They have been converted into text files and homogenized/standardized into the present BySoc corpus (Henrichsen 1997, 1998a, 1998b).
[?]

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

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

3.2 Transcription formats

The general format of the files included in the two corpora, the information included in the headers, the choice of what is transcribed, the types of comments included and the adaptation of standard orthography to spoken language all differ in some respects. BySoc is transcribed with Dansk Standard (DS) (Gregersen et al. 1991, Juel Henrichsen 1998), GSLC is transcribed with the Göteborg Transcription Standard (GTS) (Nivre 1999b) , which gives language universal traits of transcription (GTS general), in combination with Modified Standard Orthography 6 (MSO6) (Nivre 1999a), which gives the traits particular to Swedish. An overview of the differences, which have to be considered in "translating" between the corpora and in making comparisons, is given in tables 1-5 below. [REF]

Table 1. Comparison of transcription standards GSLC (GTS) – BySoc (DS)

	GSLC (GTS)	**BySoc (DS)**
Basic file organization of transcription	One file for transcription, but new line for each new utterance	Score format, separate files for each speaker and a separate file for all headings
Header containing information about transcription	First part of transcription file	In separate file
Sections	§ name of subsection	No subsections
Tokenization	Words separated by space	Words separated by space
Utterance delimiter	New line	2 or more spaces
Indication of new speaker	$I: (I = capital initial letter)	A>, B> … (for interviewers) 1> , 2> … (for informants)
Names	No special indication	Indicated with capital letters
Time line	# Hr. min. sec. 00.30.15 from start of recording. Total time can be given at end.	Not included
Anonymized names	Yes	Yes (in public version)

Table 1 presents differences concerning some general features of GSLC and BySoc transcriptions. DS uses score transcription as the basic format. Here every speaker is assigned a speech line which lasts throughout the transcription. The talk of each speaker is stored in a separate file. In GTS transcriptions are utterance based, so that every utterance gets a new line. In GTS, headers are the first part of a transcription. In DS, they are placed in a separate file. GTS transcriptions are also generally divided into subsections, which are given names on section lines, starting with a § sign. BySoc transcriptions are not divided into subsections. A similarity between the two corpora is that both are tokenized using words as the basic unit. In the transcriptions, words are separated by spaces. Because of the difference in basic format, the two standards are different in how utterances are separated. In GTS every utterance is given a new line (note that a line in the computer stored transcription does not necessarily correspond to a line in the printed output which depends on page and font size) while in DS utterances are only separated by spaces included in the line of a particular speaker, cf. table 2 below.

GTS allows for time lines, e.g. # 00.30.15 means 30 minutes, 15 seconds into the recording after start. A time line at the end can be used to give the total duration of the transcribed recording.

In Both GSLC and the public version of BySoc all names are anonymized.

Table 2. Illustration of GTS utterance format and DS score format (see also Appendix 1 and 2).

GSLC	BySoc
$ A: xxxx	A>xxxx xxx
$ B: zz	1> zz zzz
$ A: xxx	
$ B: zzz	

Table 2 illustrates a difference in how new speakers are indicated, in GTS this is done by $ A:, i.e. $ for speaker, capital initial for name and : to signal that what will follow is a speech line. In DS, there is a constant participant role, i.e. that of interviewer A, followed by interviewees given by digits (1, 2, 3 …).

3.3 Background information given about the recording and transcription

InDS, backgrund information is given in a separate file which is produced as a header for a given transcription. In GTS, it is mostly included in a header section at the beginning of each transcription. Over and above this information, there is also in GTS a separate file with more detailed information on some transcriptions.

Table 3 compares the headers of GTS and DS transcriptions. As can be seen, DS provides richer information about participants than GTS. GTS instead normally provides more information about the activity which is recorded. However, GTS does have standard fields for social status and several other properties of speakers and activity, but these fields are mostly empty due to lack of information. Cf Appendix 1 and 2 for examples of GTS and DS headers.

Table 3. Information given in the header of GTS and DS

	GTS	DS
Participant data		
Age of participants	Possibly year of birth(not in most)	Age always included
Gender of participants	Included	Included
Social status	Not included (can be written in header, not included now)	Included
Other participant information	Id Pseudonym Other details in separate file	ID Number Role (interviewer, interviewee) Name Class Social and geographical origin
Data on recording		
Duration	Hr. min.sec	Min.
Unique ID exists for every recorded activity ID	Yes	Yes
Recorded activity title	Hierarchy of activity types 25 activity types on top level	2 activity types: Person interview, Group conversation
Data on transcription		
Versions	Double transcriptions are removed from the core corpus (GSLC) and stored separately.	Double transcriptions are included. Main transcriptions = subcorpus "a", secondary transcriptions = "b" etc.
Name of transcriber Name of controller	Yes Yes	Yes No controller
Transcribed (the segment transcribed in the recording/activity)	Transcribed segments of recording marked	Total or Excerpt marked No excerpt identification
Transcription standard	GTS + MSO	Dansk Standard
Automatically generated statistics	No of utterances, tokens, overlaps etc.	Not provided
Additional free comments allowed	Yes	Three types: comment concerning participants, interview situation and transcription

3.4 What is transcribed?

What is transcribed can be divided into three parts:

(i) General features of what is transcribed
(ii) Comments on what is transcribed
(iii) Specific features of the systems of written representation used for Swedish and Danish

Table 4 presents the general features included in the transcriptions.

Table 4. What is transcribed in GSLC and BySoc

	GSLC (GTS + MSO)	BySoc (DS)
What vocal information is included	Everything said that is conventional, includes hesitation, feedback - standardized by MSO	Only what can be represented in standard orthography, supplemented by a list of reserved special words (e.g. ik', hva')
Hesitation	OCM-morpheme, like äh, eh etc. (OCM = Own Communication management)	_~_
Specification of Feedback (FB) expressions	Many variants, like ja, jaa, ja:, a, a: -standardized by MSO	Only ja, nej, jo, næ, næh, mm, nå and a few more
Rendering of numbers	Letters: två	Letters: to
Lengthening of vowel	spo: ö:l bi:len	spo~ øl~ bilen~
Rising intonation	Not standardly indicated, but can be represented by standard comment	? (sparsely used)
Pause with exhalation	Not indicated, but can be represented by non-standard comment, like </> @<sigh>	#
Contrastive stress	Capitals	Not indicated
Overlap	Start and end marked (only complete words) A: xxx [2 xxx]2 xx B: [2 zzzzz]2	Start but not end marked A> xxx xxx xx 1> zzzzz
Pause + time	3 degrees / // /// (short, normal, long)	3 degrees £ ££ £££ (unmarked pause, long, very long)
Interrupted word	spo+	spo-
Incomprehensible	(…)	(uf)
Uncertain transcription	(XYZ)	[XYZ]

Table 4 shows us that GTS includes more specific spoken language material, such as hesitation and feedback words. The basic format is the utterance, where also non-turns can be utterances, e.g. a totally overlapper yes or m. We can also see that vowel lengthening is done in two different ways in GTS (colon (:) directly after vowel) and DS (tilde (~), defined as "hesitation", before or after the word closest to the lengthened vowel). Rising intonation and pause with exhalation are regularly marked in DS in principle, but not in GTS, where it can however be included as a comment, cf. below. Contrastive stress is marked in GTS but not in DS (capital letters are used to indicate names in DS). When it comes to overlaps, beginning and end are marked in GTS but only beginning in DS, In GTS, overlaps are indicated with square numbered matching brackets in DS and by alignment on the score speaking line. Pause lengths are marked both in GTS and DS. However, the lengths are not the same. GTS has short, normal and long pause, while DS has pause, long pause and extraordinarily long pause (see further below, section 4). Another difference is that GTS allows time indicators after the pause symbol, either in clock time or in subjective time (counting one-one-thousand, two-one-thousand etc) to harmonize with speaker's speed. Interrupted words are marked in both corpora in two different ways (GTS uses + and DS uses -).

3.5 Comments

In table 5, we give an overview of the comments used in GSLC and BySoc.

Table 5. Comments in GTS and DS.

Types of comments	GTS	DS
Comments	< > in text to mark scope, @ <XYZ> on comment line below text line	(XYZ) in the text General comments on line above, marked K
Standardized comments	See listing in Transcription manual	(uf) (ler) (latter) also uncontrolled
Quotes of other speaker/own speech	Indicated as a regular comment.	”XYZ”
Deviating genre	Not standardly indicated. Can be indicated as subactivity or comment	{XYZ} English, reading test

The table shows that GTS has one format for comments, angular brackets @ <xyz>, on the line following the utterance containing what is commented on, while DS has two, (xyz) in text line and K> xyz for comments above speaker line (“K” represented as a pseudo speaker). GTS has a manual of standardized comments (Nivre 1999b), but also allows nonstandardized comments. In DS, there are three standardized comments included in speech lines, (uf) incomprehensible, (ler) laughs and (latter) laughter. In addition, non-standardized comments are allowed both in speech lines and above speech lines. Quotes are marked by quotation signs “ “ in DS. In GTS, quotes have no special status, but can be indicated by the angular brackets for comments described above. In DS, there is a special sign for indicating deviating genre { }. In GTS this would have to be indicated as a comment or possibly using a section line to indicate a specific subsection.

3.6 Level of standardization and phonetic specificity of the transcriptions

Another issue in comparing GTS and DS concerns the level of phonetic specificity employed in the transcriptions. In GTS, MSO (Modified Standard Orthography), a standard allowing for three levels of specification is used. It includes the following three levels allowing for disambiguation from IDT to the level of ambiguity in written language.

GTS

IDT: Non-disambiguated speech transciption (Icke Disambiguerat Tal)
Written “as it sounds” if conventionalized variants exist in speech, otherwise with standard orthography, e.g. spoken “ja” (can mean I or yes), while in writing “ja”(yes) is diffentiated from “jag” (I).

DT: Disambiguated transcription (Disambiguerat Tal)

The basic format for transcription in GTS, which can be used for transfer to IDT and to SSM (see below), but not back again, since DT contains more information than either IDT or SSM. DT represents IDT forms with additions allowing correspondence with standard written language words by curly brackets or numerical indices, e.g. ja => ja{g} (I), och -> å0 (and).
SSM: Written language correspondent (SkriftSpråksMotsvarighet)

DT represents the way it would be represented in standard written language, , e.g. ja{g} => jag (I).

Example:

IDT:	de	å	å
DT:	de{t}	å0	å1
SSM:	det (it/that)	och (and)	att (that/to)

Dansk standard

The basic format for transcription in DS is Standard orthography, which is most similar to th GSLC format SSM. This means that in transfer between DS and GTS, SSM should always be preferred.

The strictly orthographic style was introduced in the proof reading and restructuring of BySoc in 1996-97. Dansk Standard is not very specific in this respect, allowing transcribers too much freedom to guarantee a homogeneous corpus.

4. Problems in transliteration – conflicts between standards

4.1 Introduction

In general, incompatibilities between standards are related to the fact that transcription standards support different kinds of information. What is captured by one standard is missing from another.

For example, when something is regularly transcribed in one standard that is not transcribed in the other. The following phenomena in DS lack regular equivalents in GTS: some sociobiographical information, score format, names, very long pauses, rising intonation, pause with inhalation, while the following phenomena in GTS lack regular equivalents in DS: information about transcriber, controller, activity, subsections, time indications, anonymization, some OCM and FB morphemes, contrastive stress, end of overlap and conventionalized deviations from standard orthography.

The solutions in general are the following

(i) Leave phenomenon which is not indicated out of second transcription, i.e. loss of information.

(ii) Provide general way of adding information. The comment facility in GTS provides this sort of help. Instead of using ? to mark rising intonation, a comment can be added. Thus A> xxxxx? becomes

A: <xxxxx>
@ <rising intonation>.

(iii) Providing a facility for deriving missing information, cf below discussion of how endings of overlaps which are missing in DS have been derived in the GTS transliteration.

Another example of “loss of information” occurs with regard to the levels of standardization and phonetic specificity used in GSLC and BySoc. Since BySoc only uses standard orthography, the differences between MSO *ja, ja:, a* and *a:* would all disappear in BySoc and be rendered *ja.*

Let us now consider some examples of incompatibilities between standards.

4.2 The problem of underspecified background information

4.2.1 Introduction

The DS and GTS standards both distinguish two kinds of data, here referred to as 'background' and 'transcription'. *Background* data include participants' personal data, information about the recording (id-no., duration, quality, date, etc.), transcribers' personal data, and information about the structure of the transcription (no. of words, anonymization, transcription code, subsectioning, etc.). *Transcription* data include the transcribed words and other communication parts, and also the comments referring directly to the recorded events.

In this section, we study the conditions for transferring *background information* between DS- and GTS-formatted documents (problems concerning transcription data are discussed in later sections).

In both regimes, DS as well as GTS, background information is relocated to a data structure called a *header*. In GTS, headers are included in the respective activity files. In DS, in contrast, all headers are contained in a single background file. Thus in GTS *all* information related to a particular recorded activity is contained in a single file, while this is not the case in DS.

Headers, then, are the loci of background information. The DS-header and the GTS-header both consist of two different kinds of data fields:

- designated fields for conventionalized information (with controlled syntax)
- comment fields where all kinds of information may be inserted (with uncontrolled syntax)

The two regimes, however, do not agree on which particular information types to be conventionalized. For example, *Transcriber's name* is a dedicated field in GTS and DS on a par, while *Transcription date* only in GTS, and *Participant's name* only in DS.

Information types for which both standards have designated data fields, are easy to map, requiring just a formal conversion. Easier still are information types not conventionalized in either regime, as they can be transferred unchanged from one comment field to another. The remaining cases concern background data of types which are only conventionalized in one of the two regimes. Mapping in direction from controlled data fields to unspecific comment fields is fairly simple. Consider an example: a transcription date to be transferred from a GTS-header to a DS-header.

```
...
   Transcription date: 990316
   ...
```

Applying a little syntactic makeup, the data can be copied to a DS-comment line:

```
...
   EVTT: Transcription data is 990316
   ...
```

After transferring all conventionalized data, the target header may however still be incomplete, lacking essential data which are not present at all in the exporting header or present in the uncontrolled form of comments (in which case they cannot be recovered by automatic methods since comment lines have uncontrolled syntax). Consider a case of information transfer from a GTS-header to a DS-header leading to conflict.

```
...
   Participant: A = A1552
   ...
```

Applying a little syntactic makeup, the data can be copied to a DS-comment line:

```
   DELTAGER: A
      ...
     KOEN: ???
      ...
```

KOEN is sex of participant - information not provided in the GTS-header.

In such cases, default strategies (qualified guessing, default values, heuristic methods) have to be applied so that essential data will not be missing in the produced header.

4.2.2 Mapping DS-headers on GTS-headers

Field in DS	**gloss**	**Mapped to GTS-field**
INTERVIEW	activity id	Recorded activity id
BDNR	tape id	Tape
ITLE	duration	Duration
ADEL	no. of participants	(implicit)
ATRS	no. of transcriptions	(implicit)
BSTY	type of interview ("personal" or "group")	Activity type
EVTI, EVTD, EVTT	comments (interview/participant/transcription level)	Comment
DELTAGER	speaker index	Participant
BSGR	sociolinguistic category	no
NAVN, ALDR, KOEN, KLAS, TILH	name/age/sex/soc.class/origin of participant	no
TRANSSKRIPTION	transcription index	Transcription name

TRDK	transcription coverage	Transcribed segments
ITTR	dur. of transcribed segment	Duration
TRAN	transcriber id	Transcription name

All DS-fields except EVT*x* have controlled syntax.

4.2.3 Mapping GTS-headers on DS-headers

An actual DS-header is seen in the appendix.

Field in GTS	**Type of value**	**Mapped to DS-field**
Activity type	[type]	BSTY
Audible tokens	[no.]	no
Checker	[name]	no
Checking date	[date]	no
Comment	[free text]	EVTI, EVTD, EVTT
Duration	[time figure]	ITLE
Participant	[index]	DELTAGER
Recorded activity date	[date]	no
Recorded activity id	[id]	INTERVIEW
Recorded activity title	[free text]	no
Tape	[id]	BDNR
Transcriber	[name]	TRAN
Transcription date	[date]	no
Transcription name	[id]	TRANSCRIPTION
Transcription system	[id]	(implicit)

GTS-headers also include a range of statistical information that is derived from the transcription.

All GTS-fields except Comment have controlled syntax.
Examples of a DS-header and a GTS-header are found in Appendix 4.

4.3 Transliteration of pauses

Another type of problem arises when the two formats are almost similar but not quite. As an example of this, we will discuss the transliteration of pauses + time from GTS to DS.

The GTS format and DS format each provide a set of three pause symbols, viz. {/,//,///} and {£,££,£££} respectively. In addition, the GTS format includes the extended notation //*t*, where *t* is a time code (e.g. "//3.50" for pause in three and a half second). The formal similarity between the two notations suggests a straight forward translation scheme:

Pause translation scheme 1:
==========

```
  GTS <=> DS
  ----------

  /       <=> £
  //      <=> ££
  ///     <=> £££
  //t1     => £     for t1<1″
  //t2     => ££    for 1″<t2<2″
  //t3     => £££   for t3>2″
```

The relation between GTS and DS looks simple and information preserving (except for the time indicators). However, it hides a conflict in the intended meaning of the pause symbols. In GTS, the three pause symbols are glossed 'short pause', 'normal pause', and 'long pause', while the corresponding DS glosses are 'pause', 'long pause', and 'extraordinarily long pause', suggesting two semantically motivated alternatives, described in translation schemes 2 and 3 below.

Pause translation scheme 2:
==========

```
GTS <=> DS
  ----------

  /        => £
  //      <=> £
  ///     <=> ££
  //t      => £, ££, or £££   (depending on t)
```

Pause translation scheme 3:
==========

```
  GTS <=> DS
  ----------

  /        => (nothing)
  //      <=> £
  ///     <=> £££
  //t      => £, ££, or £££   (depending on t)
```

However, both scheme 2 and 3 introduce formal problems in the translation from DS to GTS: The scheme 3 translation of '££' insists on including a time figure (which is not provided in the DS transcriptions), while scheme 2 has a similar problem concerning "£££". In short: Scheme 1 is the only feasible alternative.

The remaining question is: How bad is this?

Table 6. Distribution of pause symbols. Pauses are given in absolute numbers and share of total number of pauses in each corpus.

Pause	**1st degree** **'/' and '£'**	**2nd degree** **'//' and '££'**	**3rd degree** **'///' and '£££'**
GTS	65 701 (67.4%)	27 981 (28.7%)	3 728 (3.8%)
DS	88 026 (77.6%)	22 790 (20.1%)	2 627 (2.3%)

As seen, '//' is relatively more frequent than '££'. This is expected, since a 'normal pause', arguably, is the unmarked case, while a 'long pause' is special. What is more surprising is that '//' is only *slightly* more frequent than '££', and certainly less frequent than '/' (making '/' the de facto *normal* pause). Given the fairly equal distribution of pause degrees over the two corpora, we suspect that the average lengths of the '//'- and '££'-marked pauses are not all that different (and similarly for 1st and 3rd degree pauses). If so, translation scheme 1 may be justified after all, even on semantic grounds.

But of course, a conclusive answer cannot be given without consulting the sound recordings.

4.4 Overlap

4.4.1 Different types of overlap

In GTS, overlaps are marked both at start and end. This will give four different types of overlapped segments:

- Initial: $A: [this] is an utterance
- Final: $A: this is [an utterance]
- Medial: $A: this [is an] utterance
- Complete: $A: [this is an utterance]

In the normal case an overlap consists of two segments from different speakers. In some cases there are more speakers, but with two involved speakers we will get 16 combinations. Below, some of these are given with possible interpretations:

Final (A) + Initial (B)	The most likely interpretation of this is that B interrupts A
Complete (A) + Medial (B)	A could, for example, give feedback to B

Some cases are not as intuitive, less clear to analyze, and also less common:

Complete (A) + Complete(B)	Both speakers start and stop at the same time
Complete (A) + Initial (B)	Both start at the same time but B keeps the turn
Complete (A) + Final (B)	A breaks in but they end at the same time

Some cases are impossible:

Initial+Initial, Final+Final, Medial+Medial, Medial+Initial, Medial+Final

The distinctions between the cases above are impossible to make in the BySoc corpus, but are still possible in the files created by gts2ds, because of the addition of underscores marking end of overlapped segments.

The following is a short example showing one of the possible cases of overlap position combination in GTS but not in BySoc.

$A: {j}a nä de{t} e0 ju skillna{d} på // kulturen i rom ol{i}ka samhällena / me{n} ja{g} tycke{r} inte att {d}e{t} behöv+ finnas nå{gon} motsättning [1 ändå mella{n} natur å0 kultur i vårt s+]1
$B: [1 ne:j jo: det]1 tro{r} ja{g} visst att det måste göra

In this example we have two segments overlapping each other. The segment in A's utterance is final and the segment in B's utterance is initial. Therefore, based on the overlap structure, we conclude that B probably interrupts A. In DS after a transfer with gts2ds, the example would look like this:

```
A> {j}a nä de{t} e0 ju skillna{d} på // kulturen i rom
-------------------------------------------------------------
A> ol{i}ka samhällena / me{n} ja{g} tycke{r} inte att
-------------------------------------------------------------
A> {d}e{t} behöv+ finnas nå{gon} motsättning
-------------------------------------------------------------
A> ändå mella{n} natur å0 kultur i vårt s+
B> ne:j jo: det                          tro{r} ja{g}
-------------------------------------------------------------
B> visst att det måste göra
```

Without listening to the tape it is difficult to see that B starts and utterance that interrupts A. From this representation, it looks more like two utterances. A transfer back to GTS with ds2gts would now look like this:

$A: {j}a nä de{t} e0 ju skillna{d} på // kulturen i rom ol{i}ka samhällena / me{n} ja{g} tycke{r} inte att {d}e{t} behöv+ finnas nå{gon} motsättning [1 ändå mella{n}]1 natur å0 kultur i vårt s+
$B: [1 ne:j jo: det]1
$B: tro{r} ja{g} visst att det måste göra

Now, the first part of B's original utterance looks like a totally overlapped utterance, and the rest of it like another utterance that follows after A has finished his utterance. However, as mentioned before, the underscores added by the gts2ds program will preserve all information about the overlap positions and the problem above would not arise.

Another example of the differences in transcribing overlap between GTS and DS can be illustrated by the following made up example of missing information in DS:

```
A> hello one and two ££ how are you
1>       hello a what do you say
2>               hello
```

In this case it is impossible to know if 2's "hello" starts at the same time as A's uttering of the word "two" or 1's uttering of the word "what". It looks as if all the three words start at the same time but, since there is a correspondence between A and 1only at the initial point of overlap, this is impossible to know. In GTS, on the other hand, an overlapped utterance like

2's would force the transcriber to state the position where the utterance starts both in relation to A's and 1's utterance.

4.4.2 Complex overlapping

The following example of overlap, even if unrealistic, is possible to describe in DS.

```
A>one two three four five_________ six twenty plus
B>seven_____________
C>              eight and nine_________________
D>              ten eleven_____________
E>        twelve thirteen fourteen
F>                   fifteen__________
G>                        seventeen__________
```

However, as the example suggests, such complex encodings are extremely demanding on the transcriber.

This could not be transcribed in GTS, (and is actually not allowed). It has to be simplified, since overlap symbols may not be placed inside words.

If the highly improbable section above really were to be recorded it would be impossible to transcribe that accurately in GTS. One would have to transcribe a simplified version and lose some information. A simplified but correct (according to the standard) transcribed version would be:

$A: [1 one two three four]1 [2 five six]2 [5 twenty]5 plus
$B: [1 se:ve:n]1
$C: [2 eight and]2 [5 nine]5
$D: [2 ten]2 [5 eleven]5
$E: [2 twelve thirteen]2 [5 fourteen]5
$F: [2 fifteen]2
$G: [5 seventeen]5

If this simplified version were to be transliterated back to DS, it would look as follows.

```
A>one two three four five six_______ twenty___ plus
B>seven_____________
C>                  eight and______ nine_____
D>                  ten____________ eleven___
E>                  twelve thirteen fourteen_
F>                  fifteen________
G>                                 Seventeen
```

5. Transfer tools – problems and solutions

Two tools for doing automatic transfer between the two corpora were designed. Transfer from BySoc to GTS was done with the tool ds2gts, which takes Dansk Standard (DS) into Göteborg Transcription Standard (GTS) and transfer from GSLC to DS was done with the tool gts2ds, which takes GTS into DS.

Below we will discuss some actual problems and solutions we have found in doing transfer from BySoc to GTS and from GSLC to DS.

5.1 Errors in the original transcription – Examples from translating GSLC to Dansk Standard using the gts2ds tool

A third type of problem occurs when the transcription which is to be transferred contains errors. The errors of course make consistent transference very difficult. As an example of this type of problem we will discuss some difficulties that arise because GSLC, in spite of having been checked, is not free of transcription errors.

Generally speaking, transcription excerpts not conforming to the standard are identified and rejected by the program. All such conflicts are reported by the program with error messages such as:

```
BAD overlap '[126 ]126' in line 553    pseudo overlap

BAD left context in 'Z' c21431 at [127]overlapping cannot be resolved

BAD body top (can't find '§ Start' or '§ Introduction')
                                          no explicit 'BEGIN'

BAD overlap index [126]: singleton      only one instance of [126]??
```

There are however certain types of ambiguities and minor coding errors that can be safely corrected on-the-fly. A few examples are discussed below.

5.1.1 Superfluous pauses

By definition, '/', '//' and '///' denote *pauses*. Intuitively, the term 'pause' is ambiguous between two readings: (i) 'any silence produced by a participant', or (ii) 'a (turn holding) participant is silent'. Of course, the choice of definition has implications for the transcription produced, as illustrated by the translation fragment from GTS into DS below.

Pause definition (i): a pause only arises as an internal part of a turn:

```
A>                  ä{r} de{t} berjstett där
X>TACK ann kristin                          näe // de{t}
-----------------------------------------------------------
A>              ursäkta mej              gu{d} va{d} de{t}
X> ligger (...)             ursäkta mej
-----------------------------------------------------------
A> e0 kallt / ja{g} kommer ihåg när vi (...)
X>                                          ja visst
```

Pause definition (ii): *any silence produced by a participant is a pause*

```
A>                  ä{r} de{t} berjstett där ///
X>TACK ann kristin //                        näe // de{t}
-----------------------------------------------------------
A>              ursäkta mej //          gu{d} va{d} de{t}
X> ligger (...) //          ursäkta mej ///
-----------------------------------------------------------
A> e0 kallt / ja{g} kommer ihåg när vi (...) //
X>                                            ja visst
```

Definition (ii) is clearly unreasonable leading to transcriptions with loads of redundant pause tags - merely denoting 'turn shift' - and so definition (i) is adopted by all transcribers (even without being stipulated in the coding manuals for GTS and DS). Because of this unclarity, redundant pauses have sometimes been inserted, such as in the second line of the following example from GSLC.

$PG: hej [10 //]10 ja{g} vi{ll} tanka på / [11 gå{r} de{t} bra]11
$C: [10 tack hej]10 //
$C: [11 de{t} sk+]11 de{t} ska vi höppes att de{t} gör

The conflicts are hardly visible in this transcription format. In transliteration to the DS score format, however, they jump to the eye:

```
C >      tack hej                                  de{t} sk+_______
de{t}
PG> hej //       ja{g} vi{ll} tanka på / gå{r} de{t} bra
-------------------------------------------------------------------
C > ska vi höppes att de{t} gör
```

(The underscore '_' is not part of DS, but here used to indicate the utterance endpoints in order to facilitate translation from GSLC.)

As seen, '//' above conforms to definition (ii), and '/' to (i). Such inconsistency is quite disturbing, since it corrupts the timing information of the transcription. What good is knowing that GSCL contains exactly 97,410 pauses, if you don't know how many of each kind?

In consequence, all pauses not conforming to definition (ii) are deleted by the gts2ds tool.

5.1.2 Transcribing complex overlapping

Many instances of complex overlapping structures occurring in GSLC are clearly unintentional. So in designing a transliteration algorithm, a precautious policy should be adopted. Instances of unusual overlapping can be considered as 'suspicious by default' and rejected by the program (even when they are not logically impossible).

There are however a few exceptions to the rule of rejecting by default. In cases of more than two segments with the same overlap index, the *two first* instances are considered valid and are mapped onto the score, creating a genuine overlap (if logically possible). All subsequent instances are left uninterpreted in the score.

The second exception to the rule concerns crossing overlaps of this simple type:

$A: [1 [2 actually not]1 crossing scopes]2 at all

In cases such as this, where crossing scopes can be avoided by merely swapping two adjacent indices, the program does so without further notice.

As mentioned, crossing scopes are hard to administer and often lead the transcriber to errors of great complexity. This quote is from A8211011.MS6 - notice the entangled scopes of [205], [206], and [392].

$S: ja men ä{r} de{t} bara / om du ä{r} intresserad av djur så ä{r} de{t}oftast så att du ä{r} intresserad av en viss ras å0 mena{r} / där ja{g}[202 pratar om ä{r}]202 all{t}så / om du ä{r} intresserad av djur de{t} e0 al{d}ri{g} så att du ä{r} intresserad av typ djur som helhet [203 å0]203 därför av maskar / fiskar / ormar [204 /]204 kor / ja{g} mena{r} verkligen [205 kör in dej på exakt alltihopa // och / ja{g} menar]205
$J: [202 vadå en viss ras]202
$V: [203 jo: då]203
$V: [204 jo: då]204
$J: [205 nä ja{g} e0 ju ja{g} e:{h} nä [206 nä de{t} ä{r} ju: vissa]205 / de{t} ju de{t} att [392 ///]206]392 nä ja{g} vill inte ha // utan
$K: [206 pappa /// pappa [392 du få+]206]392
$V: [392 ja{g} kan ta den]392
$V: karin ja{g} kan ta
$C: 1 ja{g} kan ta den å0 så ge{r} du viking å0 pia ja{g} sa{de} se{r} hu{r} myck+
(comment lines omitted)

For a sample transcription transliteration, see Appendix 4.

6. Conclusions

The main conclusion from this comparison is perhaps that corpora can be compared in spite of being fairly different in many ways. GSLC and BySoc have been created for different purposes, resulting in slightly different material being collected. In GSLC we have a rich variation of speech from many activities, while BySoc provides more representative data from one or two activities. There are two ways of handling this kind of sampling difference.

(i) Neglect. The difference can be ignored in some cases since all properties of spoken language are perhaps not equally sensitive to activity variation (Allwood 19XX).
(ii) Comparison of subcorpora. For properties which are activity sensitive, a subcorpus of GSLC, consisting of “interviews” and “conversations”, can be used to compare with BySoc (Allwood 19XX).

We have also seen how a systematic working through of the differences between the formats and standards used in the two corpora can be used to pinpoint where the differences lie and to suggest remedies that are good enough to allow programs for automatic transference to be

constructed. Above we have given a fairly complete survey and transliteration of such differences connecting them with

(i) Standard
(ii) Header
(iii) What is transcribed
(iv) Allowable comments
(v) Level of standardization and phonetic specificity

We then discussed three types of problems and solutions that can arise in attempting to automatically transfer from one type of transcription to another considering both problems that arise because of incompatibilities between standards and problems that arise because of difficulties in implementing the standards.

Concerning incompatibilities between standards, the problem we are faced with is considering what is not so essential in a transcription. We also have to consider if transcriptions should be subdivided into an obligatory part and an optional part which can always in principle be expanded to accommodate new information from another transcription format.

In general, differences between standards can be brought out by increasing the validity and reliability of transcriptions via the use of operational definitions. If such definitions are present, it will in the end always be possible to fairly specifically determine the nature of the differences.

Finally, the discussion of difficulties caused by errors in the original transcription points to the necessity of having simple and reliable transcription formats and standards. It also points to the advantage of transcribing in a format which is homomorphic with speech. When it comes to overlaps, such ease of transcription seems to be more true of the score format than of the utterance format.

References

Allwood, J., Björnberg, M., Grönqvist, L., Ahlsén, E. & Ottesjö, C. 2000. The spoken language corpus at the department of linguistics, Göteborg University. *FQS – Forum Qualitative Social Research*, Volume 1, No. 3 – December 2000.

Allwood, J., Grönqvist, L., Ahlsén, E. & Gunnarsson, M. 2002. *Göteborgskorpusen för talspråk (The Gothenburg Spoken Language Corpus)*. Nydanske Studier & Almen Kommunikationsteori, 30. Köpenhamn: Akademisk.

Allwood, J., Grönqvist, L. Ahlsén, E. & Gunnarsson, M. 2002. Annotations and tools for an activity based spoken language corpus. Forthcoming in van Kuppevelt, J. (ed.) *Current and New Directions in Discourse and Dialogue*" (Proceedings from SIGDial workshop Aalborg Aug. 2002). Kluwer Academic Publishers.

Gregersen, F. et al. 1991. *The Copenhagen Study in Urban Sociolinguistics;* 1 + 2. København: Reitzel.

Henrichsen, P. J. 1997. Talespog med ansigtsløftning. Utilisering af et stort dansk talesprogskorpus. *Instrumentalis* 10/1997, IAAS, Københavns Universitet.

Henrichsen, P. 1998a. Talesprog med netstrømper, Internet-adgang til et stort dansk talesprogskorpus. Instrumentalis 12/1998, IAAS, Københavns Universitet.

Henrichsen, P. 1998b. Peeking into the Danish living room. In *Proceedings from NODALIDA* 1998.

Nivre J. 1999a. *Modifierad Standardortografi Version*. Göteborg University, Department of Linguistics. http://www.ling.gu.se/projekt/SLSA/Publications2.html

Nivre J. 1999b. *Transcription Standard Version 6.2*. Göteborg University, Department of Linguistics. http://www.ling.gu.se/projekt/SLSA/Publications2.html

Appendix 1: GSLC-transcription V8203011.MS6 (in toto)

@ Activity type, level 1: Travel agency
@ Activity type, level 2: Face to face
@ Activity type, level 3: Göteborg 5
@ Anonymized: Yes
@ Audible tokens: 271
@ Checker: Anna Maria Szczepanska
@ Checking date: 991016
@ Comment: Fiona is talking with a foreign accent
@ Duration: 00:02:16
@ For external use: ???
@ KERNEL: yes
@ Participant: F = F1552 (Fiona)
@ Participant: R = F1540 (Rita)
@ Participant: T = F1553 (Tintin)
@ Recorded activity date: 981126
@ Recorded activity id: V820301
@ Recorded activity title: Travel Agency, Face to Face, Göteborg, dialog 5
@ Section: Start
@ Section: End
@ Short name: TravelAgencyFaceGbg5
@ Stat.Contributions: 38
@ Stat.Overlapped tokens: 7
@ Stat.Overlaps: 4
@ Stat.Participants: 3
@ Stat.Pauses: 42
@ Stat.Sections: 1
@ Stat.Stressed tokens: 0
@ Stat.Turns: 37
@ Tape: V8203,KV8203
@ Time coding: Yes
@ Tokens: 275
@ Transcribed segments: All
@ Transcriber: Helen Tak
@ Transcription date: 990316
@ Transcription name: V8203011
@ Transcription system: MSO6
§ Start
# 00:00:00
$R: < m / då ska vi se om ja{g} kan hjälpa dej > / < hej >
@ < event: R is looking through some papers >
@ < mood: cheerful >
$F: hej (...) ja{g} vill väldi{g}t gärna resa på lörda{g} [0 å0]0 sen komma på sönda{g} / e0 de{t} möjli{g}t att resa så
$R: [0 m]0
$R: < å0 komma hem på sönda{g} >
@ < mood: asking >
$F: ja
$R: <1 <2 vart vill du åka då / >1 >2
@ <1 smiling >1
@ <2 gesture: R lutar huvudet >2

$F: < london >
@ < name of city >
$R: < london > / < ja'a har vi bara platser så / / >
@ < name of city >
@ < event: R is writing on her computer >
$F: < e{h} men hur mycke{t} kostar de{t} / >
@ < event continued: R is writing on her computer >
$R: < bara flyg du vill ha >
@ < event continued: R is writing on her computer >
$F: < ja / bara / / > < >
@ < event continued: R is writing on her computer >
@ < sigh >
$T: < ja{g} har bara (...) kvar >
@ < comment: T is a person talking somewhere in the background > , < quiet >
$R: < > < > e:1 billi{g}aste flyget e0 me{d} < british airways > / vi skall se om vi har nå{g}ra platser ledi{ga} på lörda{g} / / < / / >
@ < gesture: shaking her head >
@ < click >
@ < name of company >
@ < event: conversation in the background between T and a client >
$F: ibland ni hade om < sista minut / >
@ < gesture: R is shaking her head >
$R: < ja men de{t} e0 bara > < chartern > då och då måste du va{ra} borta en hel vecka /
@ < gesture: R is turning her head back and forth >
@ < loan English: charter >
$F: < jaha man måste vara borta en hel vecka >
@ < quiet >
$R: ja'a /
$T: < men de{t} va{r} ju skönt > /
@ < event: T is talking to her client in the background >
$F: heter dom sista minut / / va{d} heter < dom >
@ < ingressive: R >
$R: sista minuten ja de{t} e0 me{d} < charter > ja / ja'a
@ < loan English: charter >
$F: ja
$R: men om du skall åka på lörda{g} å0 hem på sönda{g} då får du ju åka me{d} regulejär flyg å0 / då e0 < british airways > billi{g}ast
@ < name of company >
$F: hur micke e0 de{t}
$R: de{t} e0 tvåtusennittifem plus flygskatt < / / >
@ < event continued: T is talking to a client in the background >
$F: m'm / de{t} e0 micke för en dag
$R: < ja'a men > du kan ju stanna i en månad / de{t} har ingen betydelse på / dagen där /
@ < gesture: R is showing her palms >
$F: < m / men hade ni plats / ni hade plats / de{t} finns plats >
@ < event continued: T is talking to a client in the background >
$R: < de{t} finns plats ut ja > / elva å0 tie
@ < gesture: nods >
$F: du säger tvåtusenniohundra
$R: tvåtusennittifem plus flygskatt tvåhundra så [1 ungefär två å0 tre]1
$F: [1 (...)]1 me{d} pengar då kanske skall betala me{d} pengar /
$R: < ungefär / e{h} cirka tvåtusen+ / +trehundra / inklusiv{e} flyg+ / +skatt >
@ < event: R is writing it down on a paper >
$F: de{t} e0 tvåhundra pound e0 de{t} så < > / ja{g} kan räkna ungefär
@ < event: R is ripping a paper >

$R: ja / < ungefär >
@ < gesture: grimaserar >
$F: < e{h} ja{g} får ta två eller tre (...) me{d} sej >
@ < mumbling >
$R: < ja >
@ < gesture: scratches her nose >
$F: tack så mycke{t}
$R: < ha / tack själv >
@ < smiling >
# 00:02:16
§ End

Appendix 2: BySoc-transcription 600000620a (excerpt)

Transcription files sliced and shown in score format:

A> ... *(interviewer)*
1> ... *(1st informant)*
2> ... *(2nd informant)*
3> ... *(3rd informant)*
K> ... *(transcriber's comments and observations)*

(to be provided)

Fragment of extralin.txt (representing interview 600000620):

```
(...)
INTERVIEW: 60000620
 BDNR: 6032-4-61, 6032-4-62
 BS96: /Gruppe_IIa/id62/tekst.txt
 ITLE: 102
 ADEL: 4
 ATRS: 1
 BSTY: pers
 EVTI:
 DELTAGER: A
  BSID: 997
  BSGR:
  ROLL: itv
  NAVN: Jens Andersen
  INIT: JA
  ALDR: 33
  KOEN: M
  KLAS:
  TILH: ikke Nyboder; fra Nørrebro
  EVTD:
 DELTAGER: 1
  BSID: 62
  BSGR: IIa
  ROLL: inf
  NAVN: Pernille Ferner
  INIT:
  ALDR: 32
  KOEN: F
  KLAS: MK
  TILH: Nyboder
  EVTD:
 DELTAGER: 2
  BSID:
  BSGR:
  ROLL: inf
```

```
 NAVN: Malene
 INIT:
 ALDR:
 KOEN: F
 KLAS:
 TILH:
 EVTD: Pernille Ferner's datter
DELTAGER: 3
 BSID:
 BSGR:
 ROLL: inf
 NAVN: Mogens
 INIT:
 ALDR:
 KOEN: M
 KLAS:
 TILH:
 EVTD: Pernille Ferner's søn
TRANSSKRIPTION: a
 BS97: /60000620/60000620a
 TRDK: T
 ITTR: 102
 TRAN: JA
 EVTT:


...
```

Excerpt from interview 60000620

The score is slightly edited. Person names are changed/masked (e.g. `K%%%%%%`, preserving only the initial letter and the word length).

```
-------------------------------------------------------------
1>                                                        mm
2>
3>der er også en der hedder B%%%%% £ i~ vores kamp ik'  £ men
A>
K>
-------------------------------------------------------------
3>ved du hvad han £ gjorde han skød hele tiden sådan nogle £
-------------------------------------------------------------
1>                                        mm
3> høje £ høje~ højdere £ med bolden ik'  £ så han er blevet
-------------------------------------------------------------
3>udvist hele tiden ££ (ler) så jeg tror nok vi skal spille
-------------------------------------------------------------
1>                    nej ej det tror jeg ikke det er alt
2>                        nej det tror
3>udendørs i dag eller i morgen
-------------------------------------------------------------
1>                                                          for
2>jeg ikke £
3>          hvorfor skal jeg ikke det ?
A>                                     det er for vådt
-------------------------------------------------------------
1>vådt mand £           (uf)           hvor er dine
3>           hvad £ det er godt nok ££
-------------------------------------------------------------
1>overtræksksbukser er det dem fra I%%% ?
2>                          du kan sgu da ikke spille ude £ i
-------------------------------------------------------------
1>              I sp-~ skal ikke spille ude før til foråret
2> fodboldshorts                       (uf)
-------------------------------------------------------------
1>££                 vel ?
2>           det skal vi da heller ikke
3>                                    ~ £ hvad hedder det nu
A>   (hoster)
-------------------------------------------------------------
3>£££ han sagde at vi skulle han £ han troede nok at vi skal
-------------------------------------------------------------
1>               ja~          nej men det er altså heller
3>spille ude £ ~    i~ £ (uf) vanter
-------------------------------------------------------------
1> ikke til dig det er til M%%%%% ££ så lad dem bare være
-------------------------------------------------------------
1> £££   £ har du ikke noget du kan sidde og lave ?
3>                                                  nej (surt)
A>    mm
-------------------------------------------------------------
1>££ nå (sukkende) men det varer lidt inden~ K%%%%% £ kommer
3>
-------------------------------------------------------------
1>hjem ££          det varer en time
3>        (laver lyde)
A>                                   er det legekammeraten ?
-------------------------------------------------------------
1>det er legekammeraten ja                    P%%%%
2>                              han er snart ikke
3>                         (larmer)
A>
-------------------------------------------------------------
1>råbende til hunden)            åh de slås jo bare som alle
2>legekammerat med M%%%%% mere ££
-------------------------------------------------------------
1>andre~ (uf) £ det £ er ikke særlig alvorligt
2>ja~                          ja det hørte jeg
3>                                  det er bare fordi han
-------------------------------------------------------------
1>                åh      han er en halv gang større end dig
2>                   (uf)
3>ikke er så stærk mand
-------------------------------------------------------------
1> ££       han er en halv gang større end dig ik'
3>    hva' ?                                        det kan
-------------------------------------------------------------
1>                  (ler)
2>                  K%%%%% han K%%%%% han er ikke højere
```

```
3>være lige meget £              ( råbende uf ) da ikke bange
-------------------------------------------------------------
1>    nå                     nå~ er er du det ?
2>end mig   det tror jeg nok     jeg er       jeg er hundrede
3> for (uf)
-------------------------------------------------------------
2> ~ tre højere tror jeg £ det er ikke særlig meget vel' £
3>ja
-------------------------------------------------------------
1>                                lad være med det det er
2>det £ men ha- ££       men hva- av M%%%%%
K>      (hunden nyser)
-------------------------------------------------------------
1>da ulækkert med den der £ det er en mus ££    ik' P%%%% ££
3>                                        (uf)
-------------------------------------------------------------
1>kunne man lige have gået til dyrlægen med dig hvis du
2>                    hvorfor tager du ikke dit kødben og
-------------------------------------------------------------
1>havde nået at æde af de der kyllingeskrog
2>(if)                                            (hvisker uf)
3>                                          (ler)
-------------------------------------------------------------
1>mm £ så lad nu være ££
3>                      (voldsom larm på bordet løber ud med
A>                                                    (let
-------------------------------------------------------------
1>                                          ja £ jeg keder mig
3>hunden)
A>leende) ja du har vældigt med liv i huset
-------------------------------------------------------------
1> ikke ££ der er fuld fart på altid ik' £
2>                                          mm ££ (uf)
A>                                                  (højlydt
-------------------------------------------------------------
1>       (uf) hvad med~ hvad med lektier til i morgen ? ££
A>latter)
-------------------------------------------------------------
2>der er                 (uf)         vi skal læse
3>      (kommer ind) mor (råber) tror du godt jeg bruge det
-------------------------------------------------------------
1>           hvordan (uf)
2>                        ££ (uf)
3> sværd til Z%%%%               til Z%%%%%%%%% eller de
-------------------------------------------------------------
1>                           du får sgu ikke andre end det der
3>skal £ have det rigtige £   det sorte
-------------------------------------------------------------
1> det kan jeg da godt fortælle dig det er da rigeligt du har
-------------------------------------------------------------
1> fået det ££    nej det (uf)
3>            (uf)             nå men så tager vi bare andre
-------------------------------------------------------------
1>     jamen hvorfor skal du være sådan noget åndssvagt noget
3>penge
-------------------------------------------------------------
1> £ kunne du ikke være noget så- der var lidt morsomt ?
2>                                                       jeg
-------------------------------------------------------------
1>                          ja det det~ så jeg på den seddel
2>skal også klædes ud mor
A>                                              som hvad
-------------------------------------------------------------
1>der                                           jamen du må
2>      det ved jeg ikke endnu £ (uf) fastelavn
3>                               jeg troede hun skulle
A>   ja
-------------------------------------------------------------
1>godt finde ud af det £ i god tid £ ellers kan jeg ikke nå
2>                                   ja (uf)
-------------------------------------------------------------
1>at lave noget ££ nej vel'                                nej
2>                           jeg ved ikke hvad jeg vil være
-------------------------------------------------------------
1> £ så sæt hjernecellerne i sving £
2>    ~ min
A>                                   plejer du at sy kostumer
-------------------------------------------------------------
1>                                      M%%%%% (uf) M%%%%%
2>           mor har (uf)  gjort altid (uf)
3>                    (     banker       )
A> til dem ?
-------------------------------------------------------------
1>(irettesættende)
```

```
2>                                                  £ jeg var
3>                ja
A>                   hvad~ hvad var I sidste år ?
-------------------------------------------------------------
1>                                            ja ££
2>kat £ tror jeg nok £ ik'
3>                         jeg var en hund £££     ovre i
-------------------------------------------------------------
1>                                                          nej
3>legepladsen £ men £ herhjemme der var jeg brandvæsen ££
-------------------------------------------------------------
1>men (uf) heller ikke noget herhjemme ££
3>                                        jamen £ jo G%%%% og
-------------------------------------------------------------
1>                               (ler) ~ nej til fastelavn det
3>I% var her £ til fastelavn her
-------------------------------------------------------------
1>var nytårsaften (leende) £ der~ havde vi sådan en~ hat på
A>
-------------------------------------------------------------
1>                                             (ler) der er man
3>                                  ja (råber)
A>nå men det er også i og for sig
-------------------------------------------------------------
1>også lidt klædt ud ik' ££   man har i hvert fald hat på £
3>                           ja                              mm
-------------------------------------------------------------
1>mm                                          ved du hvad du
3>   ligesom fastelavn (karikerende udtale)
-------------------------------------------------------------
1>skal ikke gøre det der fordi så~ går det i stykker £ det er
-------------------------------------------------------------
1>ikke særlig solidt ££ og du får ikke andet ££
2>                                               (uf)   stænger
3>                                                  (uf)
-------------------------------------------------------------
1>    i forvejen er det meget mod mine principper det der £
2> ££
-------------------------------------------------------------
2>                                        jeg tror godt jeg ved
K>(det ringer på døren børnene løber ud)
-------------------------------------------------------------
1>             hvem er det ?
2>hvem det er £
K>                            (pause mens døren åbnes og nogen
-------------------------------------------------------------
1>                               nej nej det er en mor (uf)
A>               er det (uf) ?
K>gen lukkes ind)
-------------------------------------------------------------
1>går ud)
2>        skal vi ikke til håndbold hvad er klokken egentlig
-------------------------------------------------------------
2>da ?
3>    (råber) (uf) den er lidt i to
K>                                    (pause mens der larmes
-------------------------------------------------------------
K>ved døren, båndoptageren slukkes)
-------------------------------------
```

Appendix 3
Activity types in GSLC

Activity	Recordings	Speakers	Sections	Tokens	Duration
Auction	2	6.0	113	26 459	3:14:11
Bus driver/passenger	1	33.0	21	1 348	0:13:37
Church	2	3.5	12	10 235	1:47:10?
Consultation	16	3.0	256	34 285	4:09:08?
Court	6	5.2	80	33 722	3:58:33
Dinner	5	8.0	42	30 001	2:49:54
Discussion	35	5.7	293	239 412	27:06:04?
Factory conversation	5	7.4	54	28 883	2:54:47
Formal meeting	14	8.9	210	238 460	28:39:12?
Game playing	1	5.0	2	5 960	0:50:00
Games & play	1	5.0	32	6 220	0:42:00
Hotel	9	19.0	192	18 137	9:49:55
Informal conversation	16	2.2	148	75 238	7:06:23
Interview	57	2.9	1 095	389 416	45:24:07?
Lecture	2	3.5	5	14 667	1:38:00
Market	4	23.8	42	12 175	3:55:07
Party	1	7.0	10	4 356	0:27:01
Phone	32	2.1	73	14 614	2:02:03?
Retelling of article	7	2.0	14	5 290	0:42:00
Role play	3	2.3	19	8 055	0:57:16
Shop	54	7.8	231	50 492	10:34:17?
Task-oriented dialogue	26	2.3	74	15 347	2:05:20
Therapy	2	7.0	10	13 529	2:04:07
Trade fair	16	2.1	32	14 116	1:22:06
Travel agency	40	2.7	118	39 899	6:00:06
Total	**357**	**4.9**	**3 178**	**1 330 316**	**170:32:27?**

Values in the speakers column are average instead of total
Durations marked with '?' are partly estimated according to number of tokens.

Appendix 4

A sample translation

Below is presented a fragment of GSCL-transcription, before and after gts2ds conversion.
(X means unknown speaker)

$D: de{t} kan ja{g} gärna göra
$K: skojar du me{d} mej ///
$D: hm:
$K: e0 de{t} [35 carlos]35
$A: [35 väldi{g}t]35 bra
$K: mycke{t} vällagat
$X: ja den va{r} ju mycke{t} billig //
@
$X: ja men de{t} e0 ju bara början
$C: (kan vi bara) [36 (...)]36
$D: [36 jo å0 sen]36 har [37 ni i den]37
$C: [37 en midda{g} (här igen eller)]37
@
$A: ni verkar allti{d} hm eller när ni träffades alla [38 (...)]38
$X: [38 ann eller sofi]38 (ja{g}) har sånt gott samförstånd
$D: i den finns det fler såna här kårn /
$X: kårn

After conversion into DS by gts2ds:

```
D>de{t} kan ja{g} gärna göra                          hm:
K>                           skojar du me{d} mej ///
-------------------------------------------------------
A>          väldi{g}t bra
D>
K> e0 de{t} carlos___     mycke{t} vällagat
X>                                          ja den va{r}
-------------------------------------------------------
X> ju mycke{t} billig // ja men de{t} e0 ju bara början
-------------------------------------------------------
C> (kan vi bara) (...)____     en midda{g} (här igen
D>               jo å0 sen har ni i den_______________
-------------------------------------------------------
A>       ni verkar allti{d} hm eller när ni träffades
C>eller)
D>______
-------------------------------------------------------
A>alla (...)_________
X>     ann eller sofi (ja{g}) har sånt gott samförstånd
-------------------------------------------------------
D>i den finns det fler såna här kårn /
X>                                       kårn
```

A translation back to GTS (if the underscores are removed) results in:

$D: de{t} kan ja{g} gärna göra
$K: skojar du me{d} mej ///

$D: hm:
$K: e0 de{t} [35 carlos]35
$A: [35 väldi{g}t]35 bra
$K: mycke{t} vällagat
$X: ja den va{r} ju mycke{t} billig //
@
$X: ja men de{t} e0 ju bara början
$C: (kan vi bara) [36 (...)]36
$D: [36 jo å0]36 sen har [37 ni i den]37
$C: [37 en midda{g}]37 (här igen eller)
@
$A: ni verkar allti{d} hm eller när ni träffades alla [38 (...)]38
$X: [38 ann eller]38 sofi (ja{g}) har sånt gott samförstånd
$D: i den finns det fler såna här kårn /
$X: kårn

The only differences are that some overlap ending marks have moved slightly.