Nor Jnl Ling 28.1, 5–36 © 2005 Cambridge University Press
Printed in the United Kingdom
doi:10.1017/S0332586505001307

Allwood Jens, Peter Juel Henrichsen, Leif Grönqvist, Elisabeth Ahlsén & Magnus Gunnarsson. 2005. Transliteration between spoken language corpora. *Nordic Journal of Linguistics* **28.1**, 5–36.

Transliteration between spoken language corpora

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

Comparison of languages and linguistic data is essential if progress in our understanding of the nature of spoken languages is to be made. We understand phenomena better through comparison and contrast. This 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. The problems that arise are related both to the differences that exist between the standards of the corpora and to human errors leading to lack of reliability in creating the transcriptions. Although the discussion is based on transfer and transliteration between two specific corpora (the Danish BySoc, BySociolingvistisk Korpus, and the Swedish GSLC, Göteborg Spoken Language Corpus), we believe that 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.

Keywords comparison of transcriptions, corpus, corpus comparison, corpus linguistics, language comparison, reliability of transcriptions, spoken language, transcription, transfer of transcriptions, transliteration

Jens Allwood, Leif Grönqvist, Elisabeth Ahlsén & Magnus Gunnarsson: Department of Linguistics, Göteborg University, Box 200, SE-405 30 Göteborg, Sweden. E-mail: jens@ling.gu.se, leifg@ling.gu.se, eliza@ling.gu.se, mgunnar@ling.gu.se

Peter Juel Henrichsen, Center for Computational Modelling of Language, Copenhagen Business School, Bernhard Bangs Alle 17B, DK-2000 Frederiksberg, Denmark. E-mail: pjuel@id.cbs.dk

1. INTRODUCTION AND PURPOSE

Most linguists would agree that spoken language, from both a theoretical and an empirical point of view, is the basic phenomenon with which linguistics should be concerned. However, it is only with the arrival of audio and video recording, in combination with digital technology, that large-scale empirical investigation of spoken language has become possible. Such investigations are essential, not only for theoretical linguistics, but also for many types of applied linguistics such as speech recognition, provision of materials for teaching language and communication, materials used in speech therapy and as a background for the standardization of diagnostic and other tests.

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

As far as linguistic theory and data for linguistic theory are concerned, it is a common experience that speakers often have uncertain or inconsistent intuitions about spoken language (as distinct from written language). Spoken language corpora may help compensate for this by giving research on spoken language an additional and perhaps more stable empirical basis. As a consequence of this and the development of technology, we have seen the establishment of an increasing number of spoken language corpora in the last ten years. However, there is no agreement on how spoken language should be represented in a corpus. Some spoken language corpora are created with the punctuation conventions of written language. Some researchers on spoken language systematically exclude such spoken language features as overlaps, hesitations, changes of what has been said and interactive feedback signals. Others hold that these features are essential for a correct theoretical understanding of the nature of spoken language. The lack of agreement is also reflected in the conventions for construction of spoken language corpora, making comparisons between corpora difficult. However, corpora are considered useful resources by an increasing number of linguists, and indeed researchers in other fields as well, such as speech communication, sociology and business. The purposes of the corpora are diverse, and this is reflected in the great variety of formats. Several ambitious attempts have been made to create a standard for corpora, specifically concerning transcription format, data storage format, coding conventions, meta data structure, and other aspects. Examples of such projects are the ATLAS project for linguistic annotation (Bird et al. 2000) and the MATE project for multilevel annotation (Dybkjaer et al. 1998). Though it might change in the future, none of these efforts has received wide support and, more importantly, there are many corpora that have already been collected that do not follow any of these standards.

In spite of these discrepancies, comparison of languages and data is essential if progress in our understanding of the nature of spoken languages is to be made. We understand phenomena better through comparison and contrast, and we gain a better understanding of what is represented in our transcriptions by comparing different transcription systems. A more practical reason for comparison of corpora is that it is costly to build up corpora in terms of both economy and time. There are good reasons for resource sharing whenever this is possible. However, even if it is becoming increasingly desirable to be able to compare data from different, already existing corpora, the methodological problems of how to overcome differences in standards and formats still need to be solved. There seem to be two main ways to provide a solution.

The most traditional and probably the most difficult way is to try to impose a uniform standard on the scientific community. According to this view, researchers in the field should meet, agree on a set of transcription standards and then follow them. Although this indeed would be ideal, unfortunately, it is probably not realistic at the present stage. Teuber (1997) gives an account of the wealth of standards available

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

today and the amount of time that would be involved in convincing people to adopt any one of them. The lack of agreement about what is relevant data and what is a relevant theory for spoken language, added to differences in linguistic tradition, indicate that we will have to live with differences in transcription standards for some time to come.

Another way to facilitate comparisons is to accept that there are different standards and then to provide a procedure of translation or transliteration between them. If this translation or transliteration can be done automatically, separate standards can be maintained, but comparisons can still be made, when needed or desired, with the help of a transliteration program. Since we already have a situation of many different types of corpora, it is of general interest to try to develop this methodology.

This paper is written with the second type of solution in mind and 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. Both corpora existed prior to the study reported in this paper and were collected independently of each other. Thus, the problem we were faced with was how to compare them on a large scale, given their already established properties. This means that some ways of making the comparisons – such as (i) teaching interested researchers the tools and transcription system used in one or both of the corpora and subsequently letting them carry out the comparison working with the system and tools or (ii) allowing researchers to retranscribe small parts of one or both of the corpora using a third system, of their own – were not really adequate for the task at hand. Although useful for certain purposes, such methods are not adequate for large scale comparison of linguistic structures in two fairly big corpora.

The procedure we chose was (i) to compare the transcription standards and formats of the two corpora and (ii) to construct a 'translation' or rather a 'transliteration' program for transferring transcriptions which had been made according to the standard used in BySoc to the standard used in the GSLC and vice versa (the GSLC standard is called GÖTEBORG TRANSCRIPTION STANDARD, GTS, and the BySoc standard is called DANSK STANDARD, DS). In reporting on this work below, we will also discuss more generally problems, choices and solutions for corpus transcription and transference between different formats for spoken language corpora. In particular, examples of transliteration originating from the use of two tools for doing automatic transfer, namely *ds2gts* (Dansk Standard to Göteborg Transcription Standard) (applied to transfer from BySoc to the GSLC) and *gts2ds* (Göteborg Transcription Standard to Dansk Standard) (applied to transliteration from the GSLC to BySoc), will be discussed.

The account of our work, as well as the more general discussion, is provided in order to gain more insight into what is generally required when automatically comparing two spoken language corpora, which have been constructed with different goals and standards. Another purpose of the account is to add to a better understanding

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

of what spoken language features are essential to preserve in a spoken language corpus. We go into details of the actual transliteration between the GSLC and BySoc because many of the questions need to be considered in detail to be understood. The insight sometimes lies in the detail. Thus, the paper discusses some of the questions that have to be addressed in transcription and in doing transliteration between corpora transcribed according to different standards.

2. SIMILARITIES

Before we go into the differences between the two corpora, we want to point to the fairly extensive similarities between them. These similarities increase the interest of the comparison, since the number of factors available to explain the differences is smaller, making it easier to separate linguistic differences from other types of differences. In contrast with corpora collected by phoneticians and dialect researchers, dialog and interaction is the main focus of interest. Both corpora, therefore, consist mainly of fairly informal, spoken language interaction between two or more normal, adult, first language speakers (i.e. not children, persons with language disorders, or second language speakers). The corpora have the same size and the main parts were collected during the same period of time. They represent two Scandinavian languages with considerable similarities. They were both collected by linguists who wanted a descriptive non-normative transcription of the interaction, i.e. not the kind of grammatically normative transcription which is common in courts of law, police reports or the parliament. Both corpora were created for basic research purposes, i.e. they were not made for commercial purposes, as are the corpora collected in telecommunications, speech technology or in many user investigations. Both corpora are transcribed according to standards that are compromises between the three purposes of (i) representing spoken language collected in naturalistic circumstances with as little interference from a researcher 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 with spaces between words, but that the transcription standards are specially designed for SPOKEN language (cf. Allwood 1998). Finally, 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.

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

The BySoc corpus was originally recorded and transcribed in 1986–90 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, i.e. with parallel running lines for the different participants (see transcription example in Appendix 2). They have been converted into text files and homogenized/standardized into the present BySoc corpus by Henrichsen (1997, 1998a, b).

The GSLC (the Göteborg Spoken Language Corpus) was mainly recorded in the period 1978–2000 as part of many different projects, with the main purpose of representing many different social activities. (It does, however, also include a few recordings from the 1960s.) The corpus contains around 20 different social activity types (for an overview of activity types, see Appendix 3). It is described in Allwood (1999, 2001) and Allwood et al. (2000, 2002).

This difference in purposes, i.e. that the GSLC puts priority on representing different social activity, whereas BySoc puts priority on representing different individuals in related circumstances, 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. The 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 the GSLC, containing informal interviews and conversations more similar to BySoc.)

3.2 Transcription formats

This section is a fairly detailed comparison of the transcription formats of the two corpora. Over and above the reasons given above, such comparisons are motivated since the transcriptions serve as a 'map' to the actual recorded interactions and is the basis for both automatic and manual searches and comparisons of specific extracts or phenomena. It is therefore important to be aware of and find ways to deal with differences in transcription formats. Different corpora exhibit differences in format, based on the tradition in which they were created and the purpose of the original transcriptions.

The two corpora in this study were both standardized prior to their comparison and are as a result of this standardization written in 'two modified standard orthographies', one for Swedish and one for Danish. They share many features,

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

but there are still some notable differences that have to be considered in doing a comparative analysis of them. The differences, in general, exemplify the types of differences that can frequently be found among corpora of transcribed spoken language.

Some of these differences concern the general formats 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 way in which standard orthography has been adapted to spoken language. BySoc is transcribed with Dansk Standard (DS) (Gregersen & Pedersen 1991, Henrichsen 1998a), the 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 'transliteration' between the corpora, and in making comparisons, is given in Tables 1–5 below. Table 1 presents differences concerning some general features of the GSLC and BySoc transcriptions, i.e. the organization of files, the headers of files with information about the recordings and transcriptions, information about time and anonymization. It also treats the actual body of the transcription, how it is organized with respect to speakers, subsections, basic transcription units (words) and utterances.

DS uses score transcription as the basic format. Here every speaker is assigned a speech line that 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 separated by spaces included in the line of a particular speaker (cf. Table 2).

GTS allows for time lines, e.g. #00.30.15 means '30 minutes and 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 the GSLC and the public version of BySoc all names are anonymized.

Thus, the basic transcription of speech looks different in the two corpora. Table 2 illustrates the differences in how speakers are indicated and how overlapping speech is written. Differences of this type (overlap) have to be considered in transfer of data between the two formats and in comparisons of transcriptions from them (see further section 4.4. below).

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

	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 headers
Header containing information about transcription	First part of transcription file	In separate file
Sections (for explanation, see text below)	§ name of subsection	No subsections
Tokenization	Words separated by space	Words separated by space
Utterance delimiter (cf. table 2)	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. Comparison of transcription standards GSLC (GTS) – BySoc (DS).

GSLC	BySoc		
$A: I don't	A>I don't		I never
$B: no	1>	no	John does
$A: I [1 never]1			
$B: [1 John]1 does			

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

In the examples presented in this paper, Swedish and Danish are used mostly and idiomatic translations into English are provided in a simple format, leaving out some of the features of the Swedish and Danish transcriptions which are not translatable. A few longer extracts from the corpora without translations occur in the appendices.

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

In some cases, only examples in English are given, in order to make the illustrations as clear and short as possible.

Table 2 illustrates a difference in how new speakers are indicated. In GTS this is done by '$A:' i.e. a dollar sign for 'speaker', a capital initial for the speaker's name and a colon 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, indicated by digits (1, 2, 3, ...).

3.3 Background information given about the recording and transcription

In DS, background 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, also in GTS, there is 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. In contrast, GTS normally provides more information about the activity that is recorded. However, GTS does have standard fields for social status and several other characteristics of speakers and activity, but these fields are mostly empty due to lack of information. See Appendix 1 and 2 for examples of GTS and DS headers.

3.4 What is transcribed?

What is transcribed can be divided into three parts:

- General features of what is transcribed
- Specific features of the systems of written representation used for Swedish and Danish
- Comments on what is transcribed

Table 4 presents the general features included in the transcriptions.

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 overlapped *yes* or *m*. We can also see that vowel lengthening is done using different symbols in GTS (a colon (:) directly after vowel) and in DS (a tilde (∼), defined as 'hesitation', before or after the word closest to the lengthened vowel). Rising intonation and pause with exhalation have conventionalized transcription characters in DS, but not in GTS, where it can, however, be included as a comment, cf. section 3.5 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, the beginning and end are marked in GTS but only the beginning in DS. In GTS, overlaps are not marked within words, so the precision is limited to the word

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

PARTICIPANT DATA	GTS	DS
Age of participants	Possibly year of birth (not in most)	Age always included
Gender of participants	Included	Included
Social status	Not included	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	Yes	Yes
Name of controller	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	Number of utterances, tokens, overlaps etc.	Not provided
Additional free comments allowed	Yes	Three types: comment concerning participants, interview situation and transcription

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

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

	GSLC (GTS + MSO)	BySoc (DS)
What vocal information is included	Everything said that is conventional, including hesitation and feedback	Only what can be represented in standard orthography
How vocal information is represented	Standardized by MSO	Standard orthography, supplemented by a list of reserved special words (e.g. ik', hva' (= not, what))
Hesitation	OCM-morpheme, like äh, eh etc. (OCM = Own Communication management)[1]	~ (~ may be attached to a word, but need not)
Specification of Feedback (FB) expressions	Many variants, like ja, jaa, ja:, a, a: (= different variants of yes) -standardized by MSO	Only ja, nej, jo, næ, næh, mm, nå (= yes, no, yes-no, mm, "well") and a few more
Rendering of numbers	Letters: e.g. två (= two)	Letters: e.g. to (= two)
Lengthening of vowel	ö:l (= beer) bi:len (= the car)	øl~ (= beer) bilen~ (= the car)
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, e.g. A: john NEVER smokes	Not indicated
Overlap	Start and end marked (only complete words) A: what [2 do]2 you say B: [2 let's go]2	Start but not end marked A> what do you say 1> let's go
Pause + time	3 degrees /, //, /// (short, normal, long)	3 degrees £ ££ £££ (unmarked pause, long, very long)
Interrupted word	spo+	spo-
Incomprehensible	(. . .)	(uf) (= incomprehensible)
Uncertain transcription	(XYZ)	[XYZ]

Table 4. General features of what is transcribed in GSLC and BySoc (XYZ = letters in transcribed words).

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

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) (= incomprehensible) (ler) (= smiles) (latter) (= laughter)
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

Table 5. Comments in GTS and DS.

level. In GTS, overlaps are indicated with square numbered matching brackets and in DS 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 using two different symbols (GTS uses + and DS uses -). Incomprehensible or uncertain transcription is marked using parentheses ( ) in GTS and square brackets [ ] in DS. XYZ are always variables over letters in transcribed words in the present text and tables.

3.5 Transcription comments

In Table 5, we give an overview of the comments used in GTS and BySoc. 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 formats, (XYZ) in the text line and K> XYZ for comments above the speaker line. GTS has a manual of standardized comments (Nivre 1999b), but also allows non-standard 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 marks in DS. In GTS, double 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.

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

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. As has been mentioned above, both corpora use standard orthography with the word as the basic transcription unit and a number of modifications to capture spoken language features better than ordinary standard orthography (see, for example, Table 4 above). There are several reasons for choosing this type of orthography (rather than, for example, the IPA) in a large corpus of spoken language transcriptions:

1. There has to be high consistency among transcribers. Over the 20 years that we have worked with spoken language, we have come to the conclusion that as a general tendency, the richer the transcription standard or coding schema is, the harder it is to achieve inter-transcriber or inter-coder reliability. Dressler & Kreuz (2000) have come to the same conclusion when working with transcribing discourse.
2. It has to be possible to learn to transcribe in a fairly short time and with an explicit and unambiguous manual.
3. Transcription (and transcription check) should be possible in a reasonable amount of time, given the high costs involved.
4. Correspondents to written language words should be identifiable, in order to enable comparisons with written language corpora, but specific features of importance to spoken language should also be retrievable, in order to describe interactive spoken language.
5. Consistency in the transcription of words and other phenomena is necessary to enable automatic search and analysis of the corpora.
6. Finally, it is assumed that any type of in-depth or microanalysis study, especially of specific functions in context, will be based on transcriptions only in combination with recordings, thus directly providing some of the specific features captured in the IPA.

A major drawback in using GTS or DS is that they are not, like the IPA, widely known or used. This means that comparison with other spoken language corpora will always involve extra work.

3.6.1 GTS

In GTS, MSO (Modified Standard Orthography) – a standard allowing three levels of specification – is used. It includes the following three levels which enable disambiguation from IDT (see below) to the level of ambiguity in written language.

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

IDT: Non-disambiguated speech transcription (IckeDisambiguerat 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 differentiated from *jag* 'I'.

SSM: Written language correspondent (SkriftSpråksMotsvarighet)
The words are written with standard written language orthography.

DT: Disambiguated transcription (Disambiguerat Tal)
DT is the basic format for transcription in GTS, and includes the information of IDT as well as that of SSM, and thus it can be converted to either of the two other formats. IDT and SSM cannot be converted to DT, 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* → *å*0 'and'.

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

3.6.2 DS

The basic format for transcription in DS is standard orthography, which is most similar to the GTS format SSM. This means that in transfer between DS and GTS, SSM should always be preferred to other levels.

The strictly orthographic style was introduced in the proof reading and restructuring of BySoc in 1996–97. The original 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, the following phenomena in DS lack regular equivalents in GTS: some sociobiographical information, score format, names, very long pauses,

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

rising intonation, and pauses with inhalation, while the following phenomena in GTS lack regular equivalents in DS: information about transcriber, controller, activity, subsections, time indications, anonymization, some own communication management (OCM) and Feedback (FB)[2] morphemes, contrastive stress, end of overlap, and conventionalized deviations from standard orthography.

The solutions in general are the following:

- Leave out from the second transcription phenomena which are not indicated in both transcriptions, i.e. loss of information.
- Provide a 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>
- Provide a facility for deriving missing information; cf. a discussion below on how the 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 the GSLC and in BySoc. Since BySoc uses only 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

Both the DS and the GTS standards distinguish two kinds of data, referred to here as 'background' and 'transcription'. BACKGROUND data (cf. Table 3 above) include participants' personal data, information about the recording (identification number duration, quality, date, etc.), transcribers' personal data, and information about the structure of the transcription (the number 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 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. For consequences of differences in background information, see below.

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

Headers, then, are the loci of background information. The DS header and the GTS header both consist of two 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 standards, however, do not agree on which particular information types are to be conventionalized. For example, TRANSCRIBER'S NAME is a dedicated field in both GTS and DS on a par, while TRANSCRIPTION DATE is a dedicated field only in GTS, and PARTICIPANT'S NAME is a dedicated field 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 date is 99-03-16

EVTT is EVenTuelt vedr. Transskription, meaning 'Miscellaneous, Transcription level'. 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 only in the uncontrolled form of comments, in which case they cannot be recovered automatically (comment lines having uncontrolled syntax). Consider a case of information transfer from a GTS header to a DS header leading to lack of information.

Participant: A = ?

The gender of participant A cannot be read off the code '?', leaving KOEN-field ('sex') unspecified in the DS-format too:

DELTAGER: A
KOEN: ???

In such cases, default strategies (qualified guessing, default values, heuristic methods) have to be applied so that obligatory data will not be missing in the header produced. Examples of DS- and GTS-headers are found in Appendices 1 and 2.

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

The differences in the recording of background information entail some limitations on how background data can be accessed and, more importantly, what types of causal analysis can be applied in a comparison of the two corpora.

4.3 The problem of false similarity

Another type of conflict arises where the two formats are superficially similar, but differ in intended meaning. As an example of this, we will discuss the transliteration of pauses 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 straightforward translation scheme.

Pause translation scheme 1			
GTS <=> DS			
/	<=>	£	
//	<=>	££	
///	<=>	£££	
//t	=>	£	for t < 1 sec
//t	=>	££	for 1 sec < t < 2 sec
//t	=>	£££	for t > 2 sec

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.

Pause translation scheme 2		
GTS <=> DS		
/	=>	£
//	<=>	£
///	<=>	££
//t	=>	£, ££, or £££ (depending on t)

Pause translation scheme 3		
GTS <=> DS		
/	=>	(nothing)
//	<=>	£
///	<=>	£££
//t	=>	£, ££, or £££ (depending on t)

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

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%)

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

However, both scheme 2 and scheme 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 conclusion, scheme 1 is the only feasible alternative. The remaining question is: How bad is this?

As seen in Table 6, // 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 on empirical grounds. A conclusive answer, however, cannot be given without consulting the sound recordings, something which is not possible for BySoc. We have therefore chosen to keep the original pause markings in transfer between the two formats. It might be objected here that it would have been better to return to the original recordings and compare pause lengths on this basis. Unfortunately, this would not have been possible, since the BySoc recordings are not available for such control, and our exercise is based on transliterating transcriptions.

4.4 The problem of incompatible timing of information

- In GTS, overlaps are marked both at the beginning and at the 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 speakers involved we will get 16 combinations. Below, some of these are given with possible interpretations:

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

- Final (A) + Initial (B) — B could, for example, interrupt 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 both end at the same time

Some cases are impossible:

- Initial + Initial
- Final + Final
- Medial + Medial
- Medial + Initial
- Medial + Final

The distinctions between the above cases are impossible to make in the BySoc corpus, but by adding underscores (_) in the files created by gts2ds (Dansk Standard to Göteborg Transcription Standard), this distinction can be upheld

```
A> ändå mella{n} natur å0 kultur i vårt s+
B> ne:j jo: det ____                          tro{r}   ja{g}
(English translation:
A> still between nature and culture in our    s±
A> no: ye:s that ____                              I think)
```

In this example, *ne:j jo: det* overlaps with *ändå mella{n} natur*.

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
(English translation:
$A: yes there is of course a difference in // the culture in the different societies / but I don't think that there has to be any opposition [1 still between nature and culture in our s+]1
$B: [1 no: ye:s that]1 I think there must be)

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

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

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 an utterance that interrupts A. From this representation, it looks more like two utterances. A transfer back to GTS with ds2gts (Dansk Standard to Göteborg Transcription Standard) 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 looks like another utterance, which 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 thus the problem will 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 the word *two* or 1's uttering the word *what*. It looks as if all three words start at the same time but, since there is a correspondence between A and 1 only 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.

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

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 the GSLC to DS was done with the tool gts2ds, which takes GTS into DS. (The orthographic spelling of words is kept in the transfer between the corpora.) Below we will discuss some actual problems and solutions we have found in doing transfer from BySoc to GTS and from the GSLC to DS.

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

One 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 the 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 the following:

```
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.

5.2 Superfluous pauses

By definition, '/', '//' and '///' denote PAUSES. Intuitively, the term 'pause' could be ambiguous between two readings: either (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. (The original type of pause marking (/ or £) is kept in the translation, since we cannot know if the pause markings correspond to each other (see section 4.3 above).

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

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
(English translation:
A> is it bergstedt there...excuse me...god how cold it is /
I remember when we (...)...
X> THANKS ann Kristin //... no // it lies (...) //... excuse me
// ... yes sure)
```

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

$PG: hej [10 //]10 ja{g} vi{ll} tanka på / [11 gå{r} de{t} bra]11
$C: [10 hej]10 // [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>      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
(English translation:
C>     hi //                      i[t] sk+___ i(t)
PG> hi //  I wa(nt) to fill up on / i(s) tha(t) ok)
```

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

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

In this example, both occurrences of '//' conform to definition (ii), and '/' to (i). Such inconsistency is quite disturbing, since it corrupts the timing information of the transcription. What good is it to know that the GSLC 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. Correctly transcribed in the GTS format, this example will be as follows:

$PG: hej
$C: hej
$PG: ja{g} vi{ll} tanka på / [11 gå{r} de{t} bra]11

5.3 Transcribing complex overlapping

Many instances of complex overlapping structures occurring in the GSLC are clearly unintentional. Thus, 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 the cases of more than two segments with the same overlap index, the FIRST TWO 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 the following 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 noted earlier, crossing scopes are hard to administer and often lead the transcriber to errors of great complexity.

6. CONCLUSIONS

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

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

- Neglect: The difference can be ignored, in some cases, since all properties of spoken language are perhaps not equally sensitive to activity variation (Allwood 2001).
- Comparison of subcorpora: For linguistic properties which are activity-sensitive, a subcorpus of the GSLC, consisting of 'interviews' and 'conversations', can be used for a comparison with BySoc (Allwood et al. 2000).

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. We have presented above a fairly complete survey of such differences and constructed transliteration procedures for:

- transcription standard
- header
- what is transcribed
- allowable comments
- 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.

When faced with incompatibilities between standards, the problem is to decide what is and what is not essential in a transcription. We also have to consider if transcriptions should be subdivided into an obligatory part and an optional part that, in principle, can always 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 determine the nature of the differences reasonably well.

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.

NOTES

1. Own Communication Management (OCM): Morphemes and operations for choice and change used by a speaker for planning and managing his/her own speech. Typical examples

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

are: hesitation sounds, such as *eh* and self-repetition or mechanisms for changing what has been said.

2. Feedback (FB): A part of interactive communication management, typically small, unobtrusive words indicating contact, perception, understanding and other attitudinal reactions. Typical examples are *yes*, *no* and *m*.

REFERENCES

Allwood, Jens. 1998. Some frequency based differences between spoken and witten Swedish. In Timo Haukioja (ed.), *Papers from the 16th Scandinavian Conference of Linguistics*. Turku: Department of Finnish and General Linguistics, Turku University, 18–29.

Allwood, Jens (ed.). 1999. Talspråksfrekvenser – ny och utvidgad upplaga [Spoken language frequencies – new and extended edition]. *Gothenburg Papers in Theoretical Linguistics* **S 21**. Göteborg: Department of Linguistics, Göteborg University.

Allwood, Jens. 2001. Capturing differences between social activities in spoken language. In István Kenesei & Robert M. Harnish (eds.), *Perspectives on Semantics, Pragmatics and Discourse*. Amsterdam: John Benjamins, 301–319.

Allwood, Jens, Maria Björnberg, Leif Grönqvist, Elisabeth Ahlsén & Cajsa Ottesjö. 2000. The spoken language corpus at the department of linguistics, Göteborg University. *FQS – Forum Qualitative Social Research* **1.2**, December 2000.

Jens Allwood, Leif Grönqvist, Elisabeth Ahlsén & Magnus Gunnarsson. 2002. Annotations and tools for an activity based spoken language corpus. In Jan van Kuppevelt & Ronnie W. Smith (ed.), *Current and New Directions in Discourse and Dialogue* (Proceedings from SIGDial workshop Aalborg Aug. 2002). Dordrecht : Kluwer Academic Publishers, 1–18.

Bird, Steven, David Day, John Garofolo, John Henderson, Chrisophe Laprun & Mark Liberman. 2000. ATLAS: A flexible and extensible architecture for linguistic annotation. Paper presented at LREC 2000, Second International Conference on Language Resources and Evaluation. <http://www.lrecconf.org/lrec2000/>, <www.xanthi.ilsp.gr/lrec/conference/papers-2.htm>

Dressler, Richard A. & Roger J. Kreuz. 2000. Transcribing oral discourse: a survey and a model system. *Discourse Processes* **29.1**, 25–36.

Dybkjaer, Laila., Niels Ole Bernsen, Hans Dybkjaer, David McKelvie & Andreas Mengel. 1998. MATE Delivarable D1.2. The MATE Markup Framework. <http://mate.nis.sdu.dk/information/matepubl.html>

Gregersen, Frans & Inge Lise Pedersen (eds.). 1991. *The Copenhagen Study in Urban Sociolinguistics*. Copenhagen: Reitzel.

Henrichsen, Peter Juel. 1997. Talesprog med ansigtsløftning. Utilisering af et stort dansk talesprogskorpus [Spoken language with a face lift. The utilization of a large Danish spoken language corpus]. *Instrumentalis* **10**. Copenhagen: IAAS, University of Copenhagen.

Henrichsen, Peter Juel. 1998a. Talesprog med netstrømper, Internet-adgang til et stort dansk talesprogskorpus [Spoken language with net stockings, Internet access to a large spoken language corpus]. *Instrumentalis* **12**. Copenhagen: IAAS, University of Copenhagen.

Henrichsen, Peter Juel. 1998b. Peeking into the Danish living room. In *NODALIDA '98: Proceedings/The 11th Nordic Conference on Computational Linguistics, Copenhagen, 28–29 January 1998*. Copenhagen: IAAS, University of Copenhagen, 109–119.

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

Nivre, Joakim. 1999a. Modifierad Standardortografi [Modified Standard Orthography]. Göteborg: Department of Linguistics, Göteborg University.
Nivre, Joakim. 1999b. *Transcription Standard Version 6.2*. Göteborg: Department of Linguistics, Göteborg University.
Teubert, Wolfgang. 1997. Language resources for language technology. in Dan Tufis & Poul Andersen (eds.), *Recent Advances In Romanian Language Technology*. Bucarest: Academiei Române, 23–34.

APPENDIX 1. GSLC HEADER + TRANSCRIPTION

GSLC-transcription V8203011.MS6 (complete)
@ Activity type, level 1: Travel Agency
@ Activity type, level 2: Face To Face
@ Activity type, level 3: Göteborg
@ Recorded activity title: Travel Agency, Face To Face, Göteborg, Dialog 5
@ Recorded activity date: 981126
@ Recorded activity ID: V820301
@ Transcription name: V8203011
@ Transcription System: MSO6
@ Duration: 00:02:16
@ Short name: Travelagencyfacegbg5
@ Participant: F = F1552 (Fiona)
@ Participant: R = F1540 (Rita)
@ Participant: T = F1553 (Tintin)
@ Anonymized: yes
@ Kernel: yes
@ Transcriber: <anonymized>
@ Transcription date: 990316
@ Checker: <anonymized>
@ Checking date: 991016
@ Project: multimodal project
@ Comment: fiona is talking with a foreign accent
@ Time coding: yes
@ Transcribed segments: all
@ Tape: v8203, kv8203
§ 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

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

$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
$R: london / < ja'a har vi bara platser så / / >
@ < 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

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

$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 >

APPENDIX 2. BYSOC HEADER + TRANSCRIPTION

Sample header (BySoc interview 600000620a)

HEADER:	EXPLANATION:
(..)	
INTERVIEW: 60000620	interview id
BDNR: 6032-4-61, 6032-4-62	recording id
ITLE: 102	duration (min.s)
ADEL: 4	no. of participants
ATRS: 1	no. of transcriptions
BSTY: pers	type of activity
EVTI:	misc. (interview level)
DELTAGER: A	particip. A
(interviewer)	
BSID: 997	A's id
BSGR:	A's category (in BySoc terms)
ROLL: itv	A's role in interview
NAVN: Jens Andersen	A's name
INIT: JA	A's initials
ALDR: 33	A's age
KOEN: M	A's sex
KLAS:	A's social class
TILH: ikke Nyboder; fra Nørrebro	A's origin
EVTD:	misc. (participant level)

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

DELTAGER: 1	participant 1 (1st informant)
BSID: 62	(do.)
BSGR: IIa	
ROLL: inf	
NAVN: Pernille Ferner	
INIT:	
ALDR: 32	
KOEN: F	
KLAS: MK	
TILH: Nyboder	
EVTD:	
DELTAGER: 2	participant 2 (2nd informant)
(. . .)	
TRANSSKRIPTION: a	transcription a (main transcription)
BS97: /60000620/60000620a	a's location in file tree
TRDK: T	a's recording coverage
ITTR: 102	a's coverage (min.s)
TRAN: JA	a's transcriber
EVTT:	misc. (transcription level)
(. . .)	

Transcription sample from interview 60000620

The score is slightly edited. Person names are changed/masked (e.g. B%%%%%, preserving only initial letter and 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
-------------------------------------------------------------
```

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

```
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
```

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

APPENDIX 3. ACTIVITY TYPES IN GSLC

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

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?

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

APPENDIX 4. A SAMPLE TRANSLATION

Below is presented a fragment of GSLC-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 gottsamförstånd
$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 gottsamförstånd
-----------------------------------------------------------
D>i den finns det fler såna här kårn /
X>                                    kårn
```

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press

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 // 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 gottsamförstånd
$D: i den finns det fler såna här kårn /
$X: kårn

The only differences are that some overlap ending marks have moved slightly and that the two utterances by X in the middle of the example are collapsed into one. The fact that these are transcribed as two utterances indicates that the transcriber thinks that it may be two different speakers, but this is not preserved through the conversions.

https://doi.org/10.1017/S0332586505001307 Published online by Cambridge University Press