Best Practices for Speech Corpora in Linguistic Research

Workshop Programme

21 May 2012

14:00 – Case Studies: Corpora & Methods

Janne Bondi Johannessen, Øystein Alexander Vangsnes, Joel Priestley and Kristin Hagen:
A linguistics-based speech corpus

Adriana Slavcheva and Cordula Meißner:
GeWiss – a comparable corpus of academic German, English and Polish

Elena Grishina, Svetlana Savchuk and Dmitry Sichinava:
Multimodal Parallel Russian Corpus (MultiPARC): Multimodal Parallel Russian Corpus (MultiPARC): Main Tasks and General Structure

Sukriye Ruhi and E. Eda Isik Tas:
Constructing General and Dialectal Corpora for Language Variation Research: Two Case Studies from Turkish

Theodossia-Soula Pavlidou:
The Corpus of Spoken Greek: goals, challenges, perspectives

Ines Rehbein, Sören Schalowski and Heike Wiese:
Annotating spoken language

Seongsook Choi and Keith Richards:
Turn-taking in interdisciplinary scientific research meetings: Using 'R' for a simple and flexible tagging system

16:30 – 17:00 Coffee break

17:00 – Panel: Best Practices

Pavel Skrelin and Daniil Kocharov
Russian Speech Corpora Framework for Linguistic Purposes

Peter M. Fischer and Andreas Witt
Developing Solutions for Long-Term Archiving of Spoken Language Data at the Institut für Deutsche Sprache

Christoph Draxler
Using a Global Corpus Data Model for Linguistic and Phonetic Research

Brian MacWhinney, Leonid Spektor, Franklin Chen and Yvan Rose
TalkBank XML as a Best Practices Framework

Christopher Cieri and Malcah Yaeger-Dror
Toward the Harmonization of Metadata Practice for Spoken Languages Resources

Sebastian Drude, Daan Broeder, Peter Wittenburg and Han Sloetjes
Best practices in the design, creation and dissemination of speech corpora at The Language Archive

18:30 Final Discussion

Editors

Michael Haugh	Griffith University, Australia
Sukryie Ruhi	Middle Easter Technical University, Ankara
Thomas Schmidt	Institut für Deutsche Sprache, Mannheim
Kai Wörner	Hamburger Zentrum für Sprachkorpora

Workshop Organizers/Organizing Committee

Michael Haugh	Griffith University, Australia
Sukryie Ruhi	Middle Easter Technical University, Ankara
Thomas Schmidt	Institut für Deutsche Sprache, Mannheim
Kai Wörner	Hamburger Zentrum für Sprachkorpora

Workshop Programme Committee

Yeşim Aksan	Mersin University
Dawn Archer	University of Central Lancashire
Steve Cassidy	Macquarie University, Sydney
Chris Christie	Loughborough University
Arnulf Deppermann	Institute for the German Language, Mannheim
Ulrike Gut	University of Münster
Iris Hendrickx	Linguistics Center of the University of Lisboa
Alper Kanak	Turkish Science and Technology Institute – TÜBİTAK
Kemal Oflazer	Carnegie Mellon at Qatar
Antonio Pareja-Lora	ILSA-UCM / ATLAS-UNED
Petr Pořízka	Univerzita Palackého
|Jesus Romero-Trillo	Universidad Autonoma de Madrid
Yvan Rose	Memorial University of Newfoundland
Martina Schrader-Kniffki	University of Bremen
Deniz Zeyrek	Middle East Technical University

Table of contents

Author Index

Call for Papers

This half-day-workshop addresses the question of best practices for the design, creation and dissemination of speech corpora in linguistic disciplines like conversation analysis, dialectology, sociolinguistics, pragmatics and discourse analysis. The aim is to take stock of current initiatives, see how their approaches to speech data processing differ or overlap, and find out where and how a potential for coordination of efforts and standardisation exists.

Largely in parallel to the speech technology community, linguists from such diverse fields as conversation analysis, dialectology, sociolinguistics, pragmatics and discourse analysis have, in the last ten years or so, intensified their efforts to build up (or curate) larger collections of spoken language data. Undoubtedly, methods, tools, standards and workflows developed for corpora used in speech technology often serve as a starting point and a source of inspiration for the practices evolving in the linguistic research community. Conversely, the spoken language corpora developed for linguistic research can certainly also be valuable for the development or evaluation of speech technology. Yet it would be an oversimplification to say that speech technology data and spoken language data in linguistic research are merely two variants of the same category of language resources. Too distinct are the scholarly traditions, the research interests and the institutional circumstances that determine the designs of the respective corpora and the practices chosen to build, use and disseminate the resulting data.

The aim of this workshop is therefore to look at speech corpora from a decidedly linguistic perspective. We want to bring together linguists, tool developers and corpus specialists who develop and work with authentic spoken language corpora and discuss their different approaches to corpus design, transcription and annotation, metadata management and data dissemination. A desirable outcome of the workshop would be a better understanding of

- best practices for speech corpora in conversation analysis, dialectology, sociolinguistics, pragmatics and discourse analysis,
- possible routes to standardising data models, formats and workflows for spoken language data in linguistic research
- ways of linking up trends in speech technology corpora with corresponding work in the linguistics communities

Topics of interest include:

- speech corpus designs and corpus stratification schemes
- metadata descriptions of speakers and communications
- legal issues in creating, using and publishing speech corpora for linguistic research
- transcription and annotation tools for authentic speech data
- use of automatic methods for tagging, annotating authentic speech data
- transcription conventions in conversation analysis, dialectology, sociolinguistics, pragmatics and discourse analysis
- corpus management systems for speech corpora
- workflows and processing chains for speech corpora in linguistic research
- data models and data formats for transcription and annotation data
- standardization issues for speech corpora in linguistic research
- dissemination platforms for speech corpora
- integration of speech corpora from linguistic research into digital infrastructures

A linguistics-based speech corpus

Janne Bondi Johannessen[1], Øystein Alexander Vangsnes[2], Joel Priestley[1], Kristin Hagen[1]
Department of Linguistics and Nordic Studies, University of Oslo[1]
CASTL, University of Tromsø[2]
P.O.Box 1102 Blindern, UiO, 0317 Oslo, Norway[1]
Faculty of Humanities, UiT, N-9037 Tromsø, Norway[2]
E-mail: jannebj@iln.uio.no, oystein.vangsnes@uit.no, joeljp@gmail.com, kristiha@iln.uio.no

Abstract

In this paper we focus on the linguistic basis for the choices made in the Nordic Dialect Corpus. We focus on transcriptions, annotations, selection of informants, recording situation, user-friendly search interface, links to audio and video, various viewings of results, including maps.

Keywords: linguistic basis, speech corpus, dialectology, Nordic languages, user-friendliness, transcription, tagging, maps.

1. Introduction

In this paper we will discuss and show the Nordic Dialect Corpus (Johannessen et al. 2009), which was launched in November 2011. The corpus was developed through years of discussions among linguists about such issues as transcriptions and desirable search features. It was subsequently developed in close cooperation between technicians and linguists. The corpus is designed to facilitate studies of variation; this is more costly, but still more rewarding. We've had a focus on user friendliness throughout, and also a focus on recoverability of the data (link to audio, video, two level transcription).

The paper focuses on the choices decided by the linguists and dialectologists.

2. About the Nordic Dialect Corpus

The Nordic Dialect Corpus (NDC) was planned by linguists from universities in six countries: Denmark, Faroe Islands, Finland, Iceland, Norway, and Sweden within the research network Scandinavian Dialect Syntax (ScanDiaSyn). The aim was to collect large amounts of speech data and have them available in a corpus for easy access across the Nordic countries. There were two reasons why this was a good idea: First, the Nordic (North Germanic) languages are very similar to each other, and their dialects can be argued to form a large dialect continuum. Second, dialect syntax had for decades been a neglected field of research in Nordic dialectology, and the hope was that with large amounts of new material, new studies would be possible.

The work started in 2006 and the corpus was launched in 2011. It covers five languages (Danish, Faroese, Icelandic, Norwegian, and Swedish). Most of the recordings have been done after 2000, but some of the Norwegian ones are also from 1950–80. There are altogether 223 recording places and 794 informants.

The overall number of transcribed words is 2,74 million. The corpus has been very costly to build because of the man power needed. As an example, transcribing the Norwegian part alone took 14 people to do, and more than 35 people have been involved in the recording work in Norway only, which included a lot of travel and organising. The Swedish recordings were given to us by an earlier phonetics/phonology project, Swedia 2000, which was a great asset. Along the way, several national research councils, Nordic funds, and individual universities, have contributed. The Text Laboratory at the University of Oslo has been responsible for the technical development.

We know of no other speech corpus that has the combination that the NDC has of two level transcriptions, easy search-interface, direct links to audio and video, maps, result handling, and availability on the web. We refer to Johannessen et al. (2009) for a comparison with other corpora, and likewise to the proceedings of the GSCP 2012 International Conference on Speech and Corpora, Belo Horizonte, Brazil (see URL in reference list).

3. Recordings and informants

Since the goal of the project was to facilitate dialectological research, we put much effort into finding suitable informants. The places were selected carefully to represent all dialect types. Given the differences between the countries, this meant that there are also differences between the necessary number of recording places, for example, in Norway there are 110 (of modern recordings, i.e. from 2006–2011), in Denmark only 15.

The informants were to represent the

traditional dialect in their area, which meant that we were looking for people who had lived at the recording location most of their life, who had parents from the same area, and who had little or no higher education. From this point of view the corpus does not fully capture sociolinguistic variation at each location. On the other hand, we aimed to have the same number of women and men, and also to have both young and old speakers (one of each category, four in total). Since the funding and even the recordings came from different sources, this goal was not always possible to fulfil.

In addition to the recordings made after 2000, we have also been able to include some recordings from earlier projects; thus, for Norwegian there are a number of recordings from 1950–1980. This means that diachronic studies will be possible for the places where we have both old and new recordings. Diachronic studies are of course to some extent also possible where there are recordings of both old and young people, which holds for most locations.

A serious problem when recording people's language is accommodation, and even bidialectality. For this reason we have recordings of dialogues between the informants without the recording assistant being there. In addition we have some short recordings of a more formal interview between informants and assistant. This will enable later studies of how informants may vary their speech according to who they talk to.

4. Transcription and annotation

4.1 Transcription

In order for the corpus to be useful, the conversations had to be transcribed accurately word by word. To be able to search in the corpus, it had to be transcribed to a standard orthography: There is so much individual and dialectal variation that this standardisation is inevitable if the corpus is to be fully searchable. All the recordings have therefore been transcribed orthographically so as to enable automatic processing by existing analytic tools.

However, there are many linguistic purposes, not only phonological, but also morphological and syntactic ones, where it is desirable to have a phonetic transcription. Thus for Norwegian and for the dialect of Övdalian in Sweden, we have also included phonetic transcriptions. Each recording was phonetically transcribed by one person, and the output was then proof-read by another person, who checked the transcription against the audio. Then the text was run through a semi-automatic transliterator, which was trained automatically for each dialect type, and whose output was orthographic transcription. A third person manually checked the output. Finally, a fourth person would proof-read the resulting orthographic transcription, checking it against the audio.

The transcribers have all been linguistics students who have read our extensive guidelines, who have learnt from each other, and who have cooperated and consulted each other along the way. They all sat in the same work place so as to ensure interaction and homogeneity in the transcriptions

The two level transcriptions are extremely valuable. It means that we can do a search for things like the Norwegian negator *ikke* 'not', and immediately get results for all kinds of pronunciations of this word: *ikke, innkje, inte, int, itte, itt* etc. The phonetic transcriptions follow adapted national conventions, not IPA: the Norwegian transcription follow that of Papazian and Helleland (2005), which uses only letters of the standard Norwegian alphabet, and there is thus no need for special fonts.

In addition to proper linguistic transcription, extra-linguistic noises, pauses etc. have also been marked. Figure 1 illustrates the two types of transcriptions.

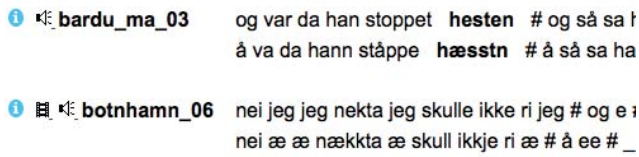


Figure 1: Two types of transcriptions

To our knowledge, no other speech corpus contains two level transcriptions of this kind. However, we would like to mention that a new Finland-Swedish dialect corpus will be adopting our tools; corpus design and interface, and will even apply a two level transcription standard (see Svenska Litteratursällskapet i Finland, in the reference list).

4.2 Annotation

The transcriptions have all been morphologically tagged for part of speech. We have mostly used available taggers, which are of course not optimal for spoken language. Some of the taggers are statistics-based and some rule-based, and some even a combination. Figure 2 shows how the grammatical tags are visible by mousing over the search results.

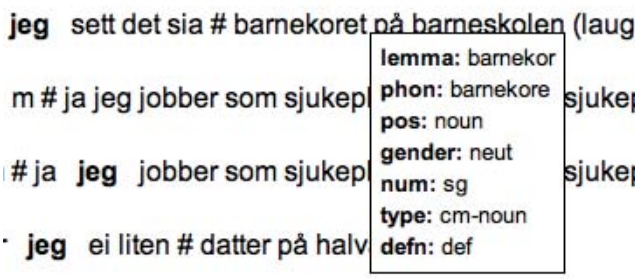


Figure 2: Grammatical tags are visible for each word

But given the written language bias it is fair to say that there is room for improvement with regard to the tagging. Transcriptions may also erroneous. Given these factors, the possibility to check the audio is crucial.

5. Links between audio/video and transcriptions

Even if a part of the corpus is phonetically transcribed, and all of it is orthographically transcribed, it is still important to have access to the audio (and in some cases, video). There are many features that are not marked in the transcriptions, such as intonation and stress. We therefore have a clickable button next to each line in the search result. This gives the linguist direct access to exactly that point in the sound file represented by the transcription. Figure 3 shows the audio and video buttons.

Figure 3: Transcription with audio/video buttons

6. Search interface

The search interface for the corpus is designed to be maximally simple for the linguists who use it. It fits into one screen, and is divided into three parts, see Figure 4. We use the system Glossa (Johannessen et al. 2008) for the search facilities.

Figure 4: The search window

The top part is the linguistic user-interface, where a search can be done specifying word, words, or parts of words, grammatical tagging, phonetic or orthographic search, exclusion of words etc. Any number of words in sequence can be chosen, and any part of a word. We use the Stuttgart CQP system for the text search (Christ 1994, Evert 2005).

The middle part of the screen is used for the desired representation of search results, for example number of results wanted, or which transcription to be displayed.

The bottom part of the screen is used for filtering the search through metadata. Here the user can choose to specify country, area or place, informant age or sex, recordings year etc.

The interface is based completely on boxes and pull-down menus. It is, however, also possible to perform searches using regular expressions, i.e., a formal language used in computer science, if necessary. We will illustrate this here. While the menu-based system allows the user to choose a part of word followed by a word, it does not allow a list of alternative characters. The system allows alternatives by adding one set of search boxes for each, but this can be a very cumbersome solution if there are many alternatives. If a user wants to specify that a word should end in any vowel, she can embed all vowels in a single search using square brackets in the following way:

(1)
.*[aeiouyæøå]

This regular expression will give as results anything that ends in a (Norwegian) vowel.

7. Presentation of search results

We have seen part of the presentations in Figures 1, 2 and 3. In Figure 5 we show more of a search results window, with not just the

transcriptions of each hit, but also the video with audio.

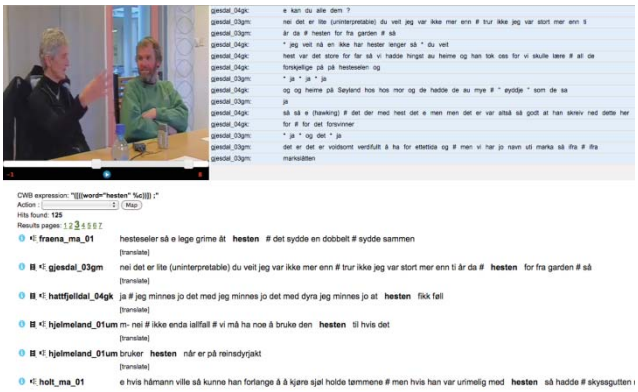


Figure 5: More of a results window

There are several other ways of showing the results. One very useful one is the map view (using Google Map technology). It shows the phonetic variations of one orthographic search. This is a very easy and enlightening way of viewing isoglosses.

We illustrate with Figure 6. Here we have simply looked up the word *ikke* 'not' in Norwegian. There are more than 30 000 hits in the corpus, and obviously impossible to quickly make a manual overview. But the map approach helps us. We have chosen to display three types of the phonetic variants on the map:

1) the versions pronouns with a fricative or affricate /ç,ʃ/ instead of the word-internal stop (red markers), for example: /iʃe/.
2) those that have fricatives and affricates followed by a nasal (yellow markers), for example: /inçe/.
3) those that are pronounced with the stop (black markers), for example: /ike/.

There are many more possibilities with regard to presentation of results, as well as results handling, and downloading.

Figure 6: Map that show results for three different kinds of pronunciations of *ikke* 'not'.

8. Conclusion

The Nordic Dialect Corpus shows the importance of involving the end users in the development of a corpus. In our case, many of the involved linguists were not experienced corpus users beforehand, but could still deliver highly valuable input regarding what would be desirable features of the corpus. In the end, that has led to a highly advanced tool for linguistic research.

9. References

Christ, Oli. 1994. A modular and flexible architecture for an integrated corpus query system. *COM-PLEX'94*, Budapest.

Evert, Stefan. 2005. The CQP Query Language Tutorial. Institute for Natural Language Processing, University of Stuttgart. URL www.ims.unistutgart.de/projekte/CorpusWorkbench/CQPTutorial.

Johannessen, Janne Bondi, Lars Nygaard, Joel Priestley and Anders Nøklestad. 2008. Glossa: a Multilingual, Multimodal, Configurable User Inter-face. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). Paris: European Language Resources Association (ELRA).*

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an Advanced Research Tool.

In Jokinen, Kristiina and Eckhard Bick (eds.): *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4.*
Papazian, Eric and Botolv Helleland. 2005. *Norsk talemål*. Høyskoleforlaget, Kristiansand.

URLs:

Google maps: maps.google.com
Nordic Dialect Corpus: http://tekstlab.uio.no/glossa/html/?corpus=scandiasyn
GSCP 2012 International Conference on Speech and Corpora. http://www.letras.ufmg.br/CMS/index.asp?pasta=gscp2012-eng
Svenska Litteratursällskapet i Finland: http://www.sls.fi/doc.php?category=1
Swedia 2000. http://swedia.ling.gu.se/

GeWiss – a Comparable Corpus of Academic German, English and Polish

Adriana Slavcheva, Cordula Meißner
Herder-Institute, University of Leipzig
Beethovenstraße 15, 04107 Leipzig, Germany
E-mail: {slavcheva, cordula.meissner}@uni-leipzig.de

Abstract

The corpus resources available for research on German academic language are limited, even with regard to the written modality. For spoken academic language they are practically non-existent. To make a first step towards remedying this situation, with *GeWiss* a comparable corpus is being constructed, consisting of spoken academic language data from German, English, and Polish academic contexts. In total it comprises about 120 hours of recording of 447 speakers including native speaker data from Polish, English, and German academics and students, as well as German as a Foreign Language (GFL) data of non-native speakers of German. Data were gathered in two genres (academic papers / student presentations and oral examinations) within one discipline (philology). The *GeWiss* corpus contains detailed metadata which are stored and administered using the EXMARaLDA Corpus Manager (cf. Schmidt/Wörner 2009). The recordings were transcribed using the EXMARaLDA Partitur Editor (ibid.) and the minimal transcription level of the GAT2 transcription conventions (cf. Selting et al. 2009), which were adapted for the multilingual *GeWiss* data. Besides the design of the *GeWiss* corpus, the metadata description and the transcription conventions applied, we describe the workflow from data gathering to corpus publication, which is planned by the end of this year.

Keywords: comparable multimodal speech corpora, academic language, German as a foreign language

1. Introduction

Research on academic language has flourished in recent years, including academic German. The corpus resources available for larger, empirically based research projects remain, however, limited, even with regard to written academic language, and they are practically non-existent for spoken academic language. A detailed, empirical analysis of linguistic conventions and formulaic language used in (oral) academic communication is, however, all the more important in a day and age where our academic landscapes are becoming ever more internationalised. *GeWiss* aims to lay a foundation for such research: With *GeWiss* a comparable corpus is being constructed, consisting of spoken academic language data from German, English, and Polish academic contexts (cf. Fandrych, Meißner & Slavcheva in print).

This paper describes the design of the *GeWiss* corpus, the metadata description of communications and speakers, the transcription conventions applied and the workflow from data gathering to corpus publication.

2. The Design of the *GeWiss* Corpus

2.1 Research Rationale

At present, publicly available and searchable corpora on (spoken) academic German do not exist. Previous studies in the field (e.g. Meer, 1998, Jasny, 2001 or Guckelsberger, 2005) are based on rather small-scale data collections which were compiled according to individual research questions and which have not been made available to the larger research community. The *GeWiss* project aims at making a first step towards remedying this situation by creating an electronically accessible multilingual corpus of academic discourse with academic German at its core. The range of spoken academic data included will allow for studies at least in the following areas:

- analyses of oral academic discourse of native speakers in the German, British and Polish academic settings,
- contrastive analyses of German and other L1 data,
- contrastive analyses of students' and experts' academic language in their respective L1s and in German as a Foreign Language,
- contrastive analyses regarding differences in the discourse practices and conventions of academic German in different academic settings, i.e. in the context of the German academic tradition in comparison to Polish and British academic traditions

In order to obtain a suitable data base, the *GeWiss* corpus is carefully designed. In a nutshell, it can be described as a balanced composition of two prototypical academic discourse genres, a monologic (i.e. academic talk) and a dialogic one (i.e. oral examinations), recorded in comparable disciplines (German, Polish, and English philologies) in a German, a British and a Polish academic context. It comprises L1 and L2 data of German as well as L1 data of English and Polish. As can be seen, English and Polish and their respective academic settings are chosen as points of comparison to German in the first instance. However, the corpus is designed in such a way that it is easily expandable in various ways – more languages and academic traditions can be added, more genres can be included, and further learner data could be added (e.g. data from secondary school settings).

2.2 Parameters of the Corpus Design

There are two key parameters determining the structure of the *GeWiss* corpus: (1) the type of discourse genre chosen, and (2) the language constellation of the speakers recorded.

The two discourse genres included in the *GeWiss* corpus were selected because they were considered to be of great relevance in many academic settings (in various disciplines, academic traditions, and linguistic communities), though, of course, they are far from being universal: Conference papers / student presentations were selected as a key monologic genre. We decided to include oral presentations / conference papers held by expert scholars as well as presentations held by students in seminars to allow for the comparison of different levels of academic literacy. The recordings also include the respective follow-up discussions as we regard them as integral parts of the communicative events as a whole. Oral examinations were chosen as a dialogic genre of prime importance for student success because in many disciplines and countries they are a prerequisite for graduation.

With regard to the parameter 'language constellation of the speakers', we took recordings of L1 productions of German, English and Polish native speakers on the one hand, and L2 productions in German on the other hand. Since our main research interest lies in the study of academic discourse and academic discourse conventions, we broadly define the L1 of our participants as the language in which they (predominantly) received their school education. This still means that in certain cases, more than one L1 can be identified. Since aspects of the participants' language biography are documented as part of the metadata questionnaire (see below), such more complex language constellations are made transparent in the metadata data base. The L2 productions are recorded in three different academic settings (German, British, and Polish), possibly also reflecting influences of different discourse traditions in the sense proposed by Koch and Österreicher (cf. Koch & Österreicher, 2008, 208ff.).

2.3 Corpus Size

The first version of the *GeWiss* corpus will comprise a total of about 120 hours of recording, i.e. 60 hours per genre and 40 hours of data originating from German, English, and Polish academic settings respectively. Table 1 gives an overview of the actual amount of transcribed data.

Academic setting	German		British		Polish		Total
Language / Genre	German L1	German L2	English L1	German L2	Polish L1	German L2	
Conference paper	603 min.	-	447 min.	321 min.	295 min.	295 min.	32:41 h
Student presentation	385 min.	267 min.	301 min.	315 min.	297 min.	302 min.	31:07 h
Oral exam	625 min.	551 min.	318 min.	666 min.	602 min.	550 min.	55:12 h
Total	26:53 h	13:38 h	17:46 h	21:42 h	19:54 h	19:07 h	119:00 h

Table 1: The actual corpus size of the *GeWiss* corpus

At the moment, the *GeWiss* corpus counts a total of 447 speakers – 128 from the German academic setting, 121 from the British and 198 from the Polish.

2.4 Related Work: *GeWiss* in Comparison to Existing Corpora of Spoken Academic Language

There are currently no other corpora of spoken academic German available. The situation is much better with regard to English where there are three corpora of spoken academic English to which the *GeWiss* corpus may be compared: MICASE (Michigan Corpus of Academic Spoken English, cf. Simpson et al. 2002), BASE (British Academic Spoken English, cf. Thompson & Nesi 2001), and ELFA (English as a Lingua Franca in Academic Settings, cf. Mauranen & Ranta 2008). The first two contain mainly L1 data while ELFA comprises only L2 recordings. *GeWiss* combines the different perspectives taken by MICASE and ELFA with regard to the usage contexts of a foreign language. While MICASE contains English L2 data produced only in a native (American) English context and ELFA contains English L2 data recorded at Finnish universities, i.e. in one specific non-English academic environment, *GeWiss* comprises both types of L2 data, thus allowing for a comparison of academic German used by L2 speakers of German both in a German context and in two non-German academic contexts.

With regard to the range of disciplines and discourse genres covered, the *GeWiss* corpus is much more focussed (or restricted) than both MICASE and ELFA. The latter two both contain recordings of a wide range of spoken academic genres and disciplines. In *GeWiss*, in contrast, the number of genres covered is confined to two (presentations and oral examinations) and that of disciplines to one (philology). With regard to the number of genres covered, *GeWiss* is comparable to BASE which contains only lectures and seminars (which were recorded in different disciplines). The restriction regarding genre and discipline in the *GeWiss* corpus is, however, a gain when it comes to research interests that include pragmatic and textual dimensions. It is also a key feature for any cross-language comparison.

As a comparable corpus, *GeWiss* differs from MICASE, BASE, and ELFA in that it contains L1 data from three different languages (recorded in all the selected discourse genres).

When comparing the *GeWiss* corpus size to that of MICASE (200 hours / 1.8 million tokens), BASE (196 hours / 1.6 million tokens) and ELFA (131 hours / 1 million tokens), it is noticeable that *GeWiss* is somewhat smaller (a total of 120 hours), but if one compares the ratio of data per genre the differences are not generally all that big (cf. Fandrych, Meißner & Slavcheva in print).

In sum, although the first version of *GeWiss* is somewhat smaller than publicly available corpora of English spoken academic discourse, its specific design offers a valuable database for comparative investigations.

3. Metadata

Since the *GeWiss* corpus was designed for comparative studies in spoken academic language, it contains detailed metadata describing the event, the recordings themselves, the transcriptions associated with the speech events, as well as the main participants. The metadata in the *GeWiss* project are encoded using the XML format from EXMARaLDA and stored and administered using the EXMARaLDA Corpus Manager (COMA) (cf. Schmidt & Wörner, 2009).

Especially the *GeWiss* metadata for the speech events and the participants are crucial for the corpus design and allow for the creation of sub-corpora for specific investigations. Below, we describe a few selected data types to illustrate this (cf. Fandrych, Meißner & Slavcheva in print).

3.1 Metadata about the Speech Event

The metadata about the speech event include 13 parameters describing the speech event as a whole, as well as information about the location the communication took place in, the languages used in the communication and some general conditions of the communicative event (partly in their relevance for the recording).

The first relevant design parameter of the *GeWiss* corpus – the type of speech event – is represented in the data type *Genre*.

According to the second design parameter of the *GeWiss* corpus – the language constellation of the speakers – metadata information relating to the nativeness / non-nativeness of the language use of a specific speaker in a given communicative event is entered in the key *L1 Communication*, in order to allow for various comparisons between speakers of different native languages. In addition, an overview of the languages used in a particular communication can be found in the section *Language(s)*. We distinguish between the *Main language of interaction* of a given communicative event and any further languages that may have been used (*Language alternation*). In addition, we characterise the event according to the *Degree of orality* (as freely spoken, read aloud or learnt by heart) based on the evaluation of the participant observer conducting the recording and the additional materials to the speech event like scripts, power point slides or notes.

3.2 Metadata about the Speakers

Apart from some basic socio-demographic information, the metadata about the speakers include information about the education as well as the languages spoken by the speakers. According to the COMA metadata model, particular stages in the education of the speakers are stored as different *Locations*. There are, however, three different types of locations in the metadata set of the *GeWiss* corpus: *Education* (as a general heading for both primary and secondary education of the speaker which is assumed to be significant for the socialisation of the speaker in the educational system of a particular language community and thus might be relevant for the specific and general academic language skills of the speaker), *Study abroad* and *Stay abroad* (for (longer) periods abroad for non-academic purposes). A further set of questions concerns speakers' language competences. Since the *GeWiss* corpus aims to provide a comparison of the academic style of speakers of three different language communities, with a particular emphasis on the distinction between native and non-native speakers of German, metadata on both the *L1* – defined as the language(s) of the educational socialisation (see above) – as well as *L2* – defined as any additional language – were collected. For all cases where German is the L2 there is an additional item *Evaluation of the language competence*. This should allow for comparison between the specific academic language skills, as represented in the recordings of the speech events, and the general language competence.

4. Transcription

All recordings in the *GeWiss* corpus were transcribed manually using the EXMARaLDA Partitur-Editor (cf. Schmidt & Wörner, 2009). The transcriptions were carried out according to the conventions of the minimal transcription level of GAT2, a system of conventions developed more than 10 years ago by German linguists with a mainly conversation analytical background for the specific purposes of the analysis of spoken language data (cf. Selting et al., 2009). According to the GAT2 conventions, words and phrases are transcribed orthographically in the speaker's tier, without any punctuation or capitalisation and using the standard pronunciation as reference norm. Strong deviations from standard pronunciation, however, such as dialectisms and regionalisms as well as idiosyncratic forms are transcribed using a 'literary transcription', e.g. *mitnanner* as an idiosyncratic form of *miteinander*.

Some of the GAT2 conventions were expanded and developed further to improve the searchability of the *GeWiss* corpus, in particular in cases of clitisation and idiosyncratic forms resulting from phonetic processes within the word boundaries. In addition, all such items are listed in a separate document together with their expanded (standardised) equivalents to enable automatic searches and to homogenise transcription practises across the *GeWiss* project (cf. Fandrych, Meißner & Slavcheva in print).

As for the transcription of the Polish and English spoken data, the GAT2 conventions were adapted by the Polish respectively English project group according to the specific spoken language phenomena of each of the two languages (cf. Lange et al. in prep.)

The current version of the *GeWiss* corpus contains mainly orthographic transcriptions of the linguistic actions of the speakers which are entered in a verbal tier. In addition, for every speaker there is a corresponding comment tier for describing non-verbal phenomena of the speakers affecting the communication; it is also used for the standardised versions of abbreviations, dialectisms, regionalisms and idiosyncratic realisations of words and phrases.

5. Annotation

Since the *GeWiss* corpus comprises non-native spoken data of German, too, which may contain instances of code

switching and code mixing, we have included an additional annotation layer for language alternation, annotated with the tag *Wechsel*. In addition, the translation of the passage is given in the comment tier of the particular speaker (cf. fig. 1).

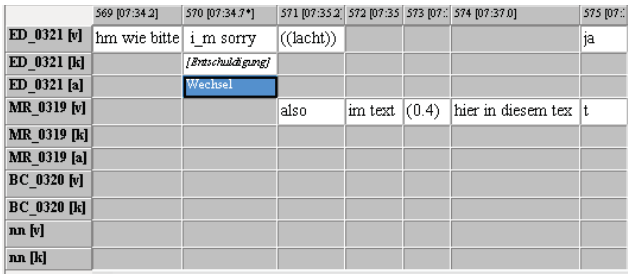


Figure 1: The annotation layer for language alternation in the *GeWiss* corpus

6. The Work Flow of the *GeWiss* Corpus Construction

Creating a well designed corpus, and in particular one comprising (recorded and transcribed) spoken data, is a labour- and cost-intensive project. For the specification of the corpus design it requires clear ideas about the kind of research questions that it might help to answer as well as about the kind of applications it may be used for. In order to build a consistent corpus in a given time-limit and to keep track of the different tasks involved it also needs a straight workflow.

The workflow of data gathering and preparation in the *GeWiss* project includes five complex subsequent steps, all coordinated by a human corpus manager:

1. *Data acquisition and recording* – including the enquiry of recording opportunities, the recruitment of participants, the request for written consent, and finally the recording itself, conducted by research assistants who were also present as participant observers in the speech events in order to identify speakers and collect metadata;
2. *Data preparation* – including the transferring of the recordings to the server, the editing and masking, the assignment of alias and the masking of the additional materials associated with the recording;
3. *Entering the metadata into the EXMARaLDA Corpus Manager* – including the linking of the masked recordings and additional materials to the speech event;
4. *Transcription* – including a three-stage correction phase carried out by the transcriber him-/herself and two other transcribers of the project;
5. *Final processing of the transcript* – including the additional masking of the recording (if needed), the check for segmentation errors, the export of a segmented transcription and finally the linking of the transcription to the speech event in the corpus manager.

At present, the final check and the segmentation of the *GeWiss* transcriptions are in progress and the digital processing of the transcribed data has started. After that, the sub-corpora will be built up and an interface for the online access will be implemented. Through this web interface the *GeWiss* corpus will be publicly available for research and pedagogical purposes after free registration. The release of the first version of *GeWiss* is planned by autumn 2012.

7. Conclusion

We have described the creation process of a comparable corpus of spoken academic language data produced by native and non-native speakers recorded in three different academic contexts, i.e. the German, English and Polish context. We presented the parameters for the design of the *GeWiss* corpus, the types of metadata collected, the transcription conventions applied and the workflow from data gathering to corpus publication. The *GeWiss* corpus will be the first publicly available corpus of spoken academic German. Its specific design offers a valuable database for comparative investigations of various kinds. The successful completion of the phase of data acquisition and transcription is an important prerequisite for the creation of a valuable corpus of spoken data for linguistic purposes. The associated expenditure of time, however, shouldn't be underestimated in the planning stage of such corpora.

8. References

Fandrych, C. ; Meißner, C. and Slavcheva, A. (in print). The GeWiss Corpus: Comparing Spoken Academic German, English and Polish. In T. Schmidt & K. Wörner (Eds), *Multilingual corpora and multilingual corpus analysis*. Amsterdam: Benjamins. (= Hamburg Studies in Multilingualism).

Guckelsberger, S. (2005). *Mündliche Referate in universitären Lehrveranstaltungen Diskursanalytische Untersuchungen im Hinblick auf eine wissenschafts-bezogene Qualifizierung von Studierenden.* München: Iudicum.

Koch, P. ; Österreicher, W. (2008). Mündlichkeit und Schriftlichkeit von Texten. In N. Janich (Ed.) *Textlinguistik*. Tübingen: Narr, pp.199--215.

Jasny, S. (2001). *Trennbare Verben in der gesprochenen Wissenschaftssprache und die Konsequenzen für ihre Behandlung im Unterricht für Deutsch als fremde Wissenschaftssprache*. Regensburg: FaDaF. [= Materialien Deutsch als Fremdsprache 64].

Lange, D. ; Rogozińska, M. ; Jaworska S. ; Slavcheva, A. (in prep). GAT2 als Transkriptionskonvention für multilinguale Sprachdaten? Zur Adaption des Notationssystems im Rahmen des Projekts GeWiss. In C. Fandrych, C. Meißner & A. Slavcheva (Eds.), *Tagungsband der GeWiss-Konferenz vom 27. - 29. 10. 2011*. Heidelberg: Synchronverlag. (= Reihe Wissenschaftskommunikation).

Mauranen, A. ; Ranta, E. (2008). English as an Academic

Lingua Franca – the ELFA project. *Nordic Journal of English Studies*, 7(3), pp. 199--202.

Meer, D. (1998). *Der Prüfer ist nicht der König: mündliche Abschlussprüfungen in der Hochschule*. Tübingen: Niemeyer.

Schmidt, Th. ; Wörner, K. (2009). EXMARaLDA – Creating, analyzing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19, pp. 565--582.

Selting, M. et al. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 10, pp. 353--402.

Simpson, R. : Briggs, S. ; Ovens, J. and Swales, J. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan. Online: http://micase.elicorpora.info

Thompson, P. ; Nesi, H. (2001). The British Academic Spoken English (BASE) Corpus Project. *Language Teaching Research*, 5(3), pp. 263--264.

9. Acknowledgements

GeWiss is funded by the VolkswagenStiftung (Az.: II/83967) as part of its funding scheme *Deutsch Plus – Wissenschaft ist mehrsprachig.*

Multimodal Parallel Russian Corpus (MultiPARC): Main Tasks and General Structure

Elena Grishina, Svetlana Savchuk, Dmitry Sichinava
Institute of Russian Language RAS
18/2 Volkhonka st., Moscow, Russia
rudi2007@yandex.ru, savsvetlana@mail.ru, mitrius@gmail.com

Abstract

The paper introduces a new project, the Multimodal Parallel Russian Corpus, which is planned to be created in the framework of the Russian National Corpus and to include different realizations of the same text: the screen versions and theatrical performances of the same drama, recitations of the same poetical text, and so on. The paper outlines some ways to use the MultiPARC data in linguistic studies.

1. Introduction

It is generally known that the main drawbacks and difficulties in the speech researches are connected with the fact that speech is not reproducible. It seems that we have no possibility to repeat the same utterance in the same context and in the same circumstances. These limitations lose their tension, when we deal with the etiquette formulas, and with other standard social reactions of a fixed linguistic structure. But unfortunately, the standard formulas of the kind are quite specific and may hardly represent a language as a whole. So, we may state that a spoken utterance is unique, in a sense that it takes place on one occasion only, here and now, and cannot be reproduced in combination with its initial consituation.

On the other hand, the question arises what part of this or that utterance is obligatory to all speakers in all possible circumstances, and what part of it may change along with the changes of speakers and circumstances. The only possible way to solve the problem is to let different speakers utter the phrase in the same circumstances. Naturally, the real life never gives us the possibility to put this into practice, laying aside the case of linguistic experiment. But the sphere of art lets us come near the solution.

To investigate the ways of the articulation of the same utterance by different speakers, but in the same circumstances, the RNC[1]-team decides to create a new module in the framework of the Multimodal Russian Corpus (MURCO[2]), which is supposed to be named Multimodal Parallel Russian Corpus (MultiPARC).

2. Three parts of MultiPARC

2.1 Recitation

We suppose that the Recitation zone of the MultiPARC will include the author's, the actor's, and the amateur performances of the same poetic or prosaic text. We plan to begin with the poetry of Anna Akhmatova, who is quite popular among professional actors and ordinary readership; besides, a lot of recordings of Akhmatova's recitations of her own poetry are easily available. There are no comparable corpora of the kind functioning at the present moment, as far as we know.

2.2 Production

MultiPARC will also include the different theatrical productions and screen versions of the same play. For example, we have at our disposal one radio play, three audio books, three screen versions, and seven theatrical performances of the Gogol's play "The Inspector General" ("Revizor"). As a result, the MultiPARC will give us the opportunity to align and compare 14 variants of the 1st phrase of the play: *I have called you together, gentlemen, to tell you an unpleasant piece of news. An Inspector-General is coming*. Naturally, every cue of the Gogol's play may be multiplied and compared in the same matter. And not only the Gogol's play, but also the plays of Chekhov, Vampilov, Rosov, Ostrovsky, Tolstoy, and so on. The only requirement to a play is as follows: it ought to be popular enough to have at least two different theatrical or screen versions.

The comparison of different realization of the same phrase, which is meant to be pronounced along with the same conditions and circumstances, but by the different actors, gives us the unique possibility to define, which features of this or that utterance are obligatory, which are optional, but frequent ones, and which are rare and specific only for one person.

Naturally, here we face the restrictions, which are connected with the artificiality of the theatrical and movie speech. Though, we definitely may come to some interesting and provoking conclusions concerning the basic features of spoken Russian, and probably of spoken communication as a whole.

2.3 Multilingual zone

The above section naturally brings us closer to the most debatable and open to question zone of the MultiPARC, namely the multilingual one. Here we suppose to dispose the theatrical productions and screen versions on the same play/novel, but in different languages (American and Russian screen versions of Tolstoy's "War and Peace",

[1] About the RNC see [RNC'2006, RNC'2009], [Grishina 2007], www.ruscorpora.ru; about the spoken subcorpora of the RNC see, among others, [Grishina 2006, 2007], [Grishina et al., 2010], [Savchuk 2009]).

[2] About the MURCO see, among others, [Grishina 2009a, 2009b, 2010].

French and Russian screen versions of "Anna Karenina", British and Russian screen versions of "Sherlock Holmes", and so on).

This zone of the MultiPARC is intended for the investigation in two fields: 1) comparable types of pronunciation (pauses, intonation patterns, special phonetic features, like syllabification, chanting, and so on), which are often the same in different languages, 2) comparable researches in gesticulation, which has its specificity in different cultures. We think that this zone of the MultiPARC may become the subject of international cooperation.

3. MultiPARC interface

The MultiPARC in total is supposed to have the interface, which is adopted just now for the MURCO. The user's query will return to a user a set of clixts, i.e. a set of the pairs 'clip + corresponding text', the corresponding texts being richly annotated. But the MultiPARC seems to have some specific features. The investigation of movie and theatrical speech has shown that the actors regularly transform the original texts of a play (see [Grishina 2007]). We often meet the transformations of the following types:

1) additions
2) omissions
3) shifts and transpositions
4) synonymic equivalents
5) apocopes
6) restructuring, and some others.

(It should be noted parenthetically that these linguistic events take place also in poetry, though quite rarely.)

As a result, the real cue pronounced on the stage or on the screen may differ considerably from the corresponding cue in the prototypical text. Consequently, the MultiPARC interface ought to provide two types of queries: 1) query for the prototypical cue, 2) query for the real cue (see Pic. 1).

If a user makes a query, which refers to the prototypical cue, then he/she receives the clusters of the real cues (i.e. the complete set of the clixts, which correspond to this very prototypical cue). But if a user makes a query, which refers to the unit (word, construction, combination of letters, accent, and so on) included in a real cue, but missing in the prototypical one, then he/she receives in return only the real cues, which contain this unit.

4. Types of Annotation

Since the MultiPARC is the result of further development of MURCO, it is quite natural that it will be annotated under the MURCO standards. These are as follows:

- metatextual annotation
- morphological annotation
- semantic annotation
- accentological annotation
- sociological annotation
- orthoepic annotation
- annotation of the vocalic word structure

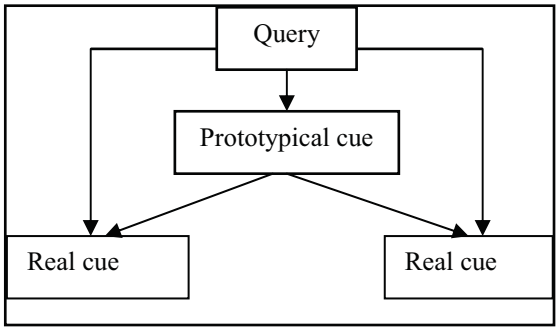
We have described all types of MURCO annotation earlier ([Grishina 2010]), so we need not to return to the question.

5. MultiPARC as Scientific Resource

MultiPARC is meant to be one of the resources for scientific researches, so its main task is the academic one. Being the academic resource, it lets us put and solve the scientific tasks, which concern following fields of investigation.

1. The regularities of the pause disposition in spoken Russian. The types of pauses from the point of view of their
1.1. obligatoriness
1.2. phonetic characteristics
1.3. duration
may be investigated systematically.
2. The regularities of the intonation patterns, which accompany the same lexical and syntactical structures.
3. The correspondence between punctuation marks and pause disposition.
4. The correspondence between the punctuation marks and intonation patterns.
5. The regularities of the change of the word order in spoken Russian in comparison with written Russian.
6. The set and ranking of clitics (proclitics and enclitics) in spoken Russian.
7. The correspondence between the communicative structure of a phrase (theme vs. rheme) and the most frequent manners of its pronunciation from the point of view of phonetics and intonation.

Below we mean to illustrate the above with some interesting observations.



Picture 1

6. Usage of MultiPARC

6.1 Syllabification in Spoken Russian

The trial version of the MultiPARC, which is being prepared just now, let us illustrate some types of its prospective usage in scientific studies. For example, we may investigate the role of some phonetic phenomena in Spoken Russian.

Let us analyze the beginning of the classic Gogol's play "The Inspector General" ("Revizor") from this point of view. The comparison of first 37 fragments gives us the possibility to analyze the main types of meaning of syllabification in Spoken Russian.

6.1.1. The highest degree of quality

Hereinafter the first figure in the brackets refers to the number of the utterances with the syllabification, the second figure refers to the total number of the utterances, and the percentage means the comparative quantity of the syllabicated utterances (it will be recalled that we have compared 14 realizations – the

theatrical performances, movies, audio books – of the same play).

The syllabification is used to mark up the words and word-combinations, which include the component 'the highest degree of quality' in their meaning (hereinafter these words and words-combinations are bold-faced).

The corresponding illustrations are as follows.

It would be better, too, if there ***weren't so many of them***. (5-11-45%)

I have called you together, gentlemen, to tell you an ***unpleasant*** (6-14-43%) *piece of news.*

Upon my word, I never saw the likes of them — black and ***supernaturally*** (6-14-43%) *big*.

The attendants have turned the entrance hall where the petitioners usually wait into a poultry yard, and the geese and goslings ***go poking their beaks*** (5-12-42%) *between people's legs.*

Besides, the doctor would have ***a hard time*** (4-11-36%) *making the patients understand him.*

An ***extraordinary*** (4-13-31%) *situation, most extraordinary!*

He ***doesn't know a word*** (3-11-27%) *of Russian.*

Last night I kept dreaming of two rats — ***regular monsters****!* (4-14-26%)

And I don't like your invalids to be smoking ***such strong tobacco***. (3-10-30%)

You ***especially*** (2-12-17%), *Artemy Filippovich.*

Why, you might gallop ***three years away from here*** (1-14-7%) *and reach nowhere.*

6.1.2. Important information, maxims and hints

The syllabification is used to mark the information of heightened importance. This group includes the suggestions and hints:

Yes, an Inspector from ***St. Petersburg***, (2-14-14%) ***incognito***. (9-14-64%) *And with* ***secret instructions***, (5-14-36%) *too.*

I had a sort of ***presentiment*** (5-14-36%) *of it*.

It means this, that ***Russia — yes — that Russia intends to go to war***, (9-13-69%) *and the* ***Government*** (4-13-31%) *has secretly commissioned an official to find out* ***if there is any treasonable activity anywhere***. (7-14-50%)

On the look-out, or not on the look-out, anyhow, gentlemen, I ***have given you warning.*** (3-14-22%)

In addition, this group includes the maxims. The maxims are the utterances stating something to be absolutely true, without any reference to time, place, and persons involved. Therefore the maxims are accompanied with the syllabification quite often to underline the importance and significance of the conveying ideas:

Treason in this little country town! (= 'It is impossible to have treason in this little country town') (4-14-29%)

The ***Government is shrewd***. (2-14-14%) *It makes no difference that our town is so remote. The Government is* ***on the look-out all the same.*** (3-14-21%)

Our rule is: ***the nearer to nature the better***. (7-12-58%) *We use no expensive medicines*.

A ***man is a simple affair***. (3-13-23%) *If he dies, he'd die anyway. If he gets well, he'd get well anyway.*

6.1.3. Introduction of the other's speech

Third group of syllabification is quite specific. It includes the utterances, which introduce the other's speech or autoquotations. Generally, the introduction precedes the other's speech, but sometimes it summarizes the citation. This group also includes the introductions of one's thoughts and opinions:

"My dear friend, godfather and benefactor — [He mumbles, glancing rapidly down the page.] *— and* ***to let you know*** (4-14-26%)"*— Ah, that's it* [he begins to read the letter aloud]

Listen to what he ***writes*** (3-14-22%)

It means this, (4-13-31%) *that Russia — yes — that Russia intends to go to war*

My opinion is (2-13-15%), *Anton Antonovich, that the cause is a deep one and rather political in character*

I have made some arrangements for myself, and ***I advise you*** (2-12-17%) *to do the same.*

So, the tentative studying of the MultiPARC data has shown that it may give us the possibility to study the semantics and functions of different phonetic phenomena in Russian systematically.

6.2 Types of pauses

The MultiPARC presents the data to investigate the types and the usage of the pauses in Spoken Russian. The preliminary analysis has shown that there are 4 types of pauses as for their frequency:

1) obligatory pauses; frequency 80-100%

I have called you together, gentlemen, to tell you an unpleasant piece of news. ‖ (14-14-100%) *An Inspector-General is coming.*

2) frequent pauses; frequency 50-79%

I advise you to take precautions, ‖ (11-14-79%) *as he may arrive any hour,* ‖ (8-14-57%) *if he hasn't already, and is not staying somewhere* ‖ (8-14-57%) *incognito*.

3) sporadic pauses; frequency 20-49%

Oh, that's a small ‖ (2-11-14%) *matter*.

4) unique pauses; frequency 8-19%.

Oh, as to ‖ (1-13-8%) *treatment, Christian Ivanovich and I have worked out* ‖ (1-13-8%) *our own system*.

Having distinguished the different types of pauses, we may analyze the correlation between

1) the frequency of pauses and the punctuation marks;

2) the duration of pauses and their frequency;

3) the types of pauses and the types of the syntactic boundaries;

4) we may also systematically investigate the expressive features of the unique pauses.

As for the last point, we may notice that breaking up the combination of an attribute and a determinatum (AD) into two parts with a pause is a quite seldom event. In 37 surveyed fragments of the Gogol's play we may see 21 combinations AD without any pauses between A and D, and only 7 combinations with the unique pauses: A‖D. As a result, the pause in the constructions like AD has a great expressivity and underlines the importance of the attribute.

7. Conclusion

We may see that the Multimodal Parallel Russian Corpus (MultiPARC) present the new type of the multimodal corpora. This corpus gives a researcher the possibility to analyze the spoken events from the point of view of their frequency,

singularity, expressiveness, semantic and syntactic specificity, and so on.

Moreover, the MultiPARC presents the data for the gestural investigations. For example, the eye behavior (namely, blinking), which is specific for the professional actors while declaiming poetry, is quite different from this of non-professional performers. Since the MultiPARC is planned to include video, we may obtain the gestural data from different screen versions and theatrical performances. So, the contrastive analysis of the data is available.

8. Acknowledgements

The work of the MURCO group and the authors' research are supported by the program "Corpus Linguistics" of the Russian Academy of Sciences and by the RFBR (The Russian Fund of Basic Researches) (RFFI) under the grants 10-06-00151 and 11-06-00030.

9. References

Grishina, E. (2006). Spoken Russian in the Russian National Corpus (RNC). In *LREC'2006: 5th International Conference on Language Resources and Evaluation. ELRA*, pp. 121-124.

Grishina, E. (2007b). Text Navigators in Spoken Russian. In *Proceedings of the workshop "Representation of Semantic Structure of Spoken Speech"* (*CAEPIA'2007, Spain, 2007, 12-16.11.07, Salamanca*). Salamanca, pp. 39-50.

Grishina, E. (2009a). Multimodal Russian Corpus (MURCO): types of annotation and annotator's workbenches. In Corpus Linguistics Conference CL2009, Universuty of Liverpool, UK, 20-23 July 2009,

Grishina, E. (2009b). Multimodal Russian Corpus (MURCO): general structure and user interface. In NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference, Smolenice, Slovakia, 25-27 November 2009. Proceedings. Tribun, 119-131, http://ruslang.academia.edu/ElenaGrishina/Papers/153531/Multimodal_Russian_Corpus_MURCO_general_structure_and_user_interface

Grishina, E., et al. (2010). Design and data collection for the Accentological corpus of Russian. In In *LREC'2010: 7th International Conference on Language Resources and Evaluation. ELRA* (forthcoming).

Grishina E. (2010) Multimodal Russian Corpus (MURCO): First Steps // 7th Conference on Language Resources and Evaluation LREC'2010, Valetta, Malta. 1

RNC'2006. (2006). *Nacional'nyj korpus russkogo jazyka: 2003–2005. Rezul'taty i perspektivy.* Moscow: Indrik.

RNC'2009. (2009). *Nacional'nyj korpus russkogo jazyka: 2006–2008. Novyje rezul'taty i perspektivy.* Sankt-Peterburg: Nestor-Istorija.

Savchuk, S. (2009). Spoken Texts Representation in the Russian National Corpus: Spoken and Accentologic Sub-Corpora. In *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference, Smolenice, Slovakia, 25-27 November 2009. Proceedings.* Brno, Tribun, pp. 310-320.

Constructing General and Dialectal Corpora for Language Variation Research: Two Case Studies from Turkish

Şükriye Ruhi[a], E. Eda Isik Tas[b]
[a]Middle East Technical University, [b]Middle East Technical University, Northern Cyprus Campus
[a]İnönü Blvd., 06531 Ankara, Turkey; [b]Kalkanlı Güzelyurt Mersin10, Turkey
E-mail: [a]sukruh@metu.edu.tr, [b]edaisik@metu.edu.tr

Abstract

Parallel corpora utilizing a standardized system for sampling, recording, transcription and annotation potentially ensures cross-linguistic analyses for researchers. This paper describes how comparability was achieved in a general corpus and a dialectal corpus in the context of the Spoken Turkish Corpus – STC (Turkish spoken in Turkey) and the Spoken Turkish Cypriot Dialect Corpus – STCDC (Turkish in Northern Cyprus). Section 2 overviews aspects of Variety corpora features that impact variation research. Section 3 describes the corpus design and sampling procedures adopted in STC and STCDC. Section 4 focuses on the corpus construction tools employed in the two corpora. Finally, Section 5 presents the transcription standardization methods devised for STC and STCDC.

Key words: Turkish dialects, discourse corpora, language variation analysis, sampling frames and principles, multi-modality, transcription standardization, annotation parameters

1. Introduction

Clancy (2010: 80-81) makes a useful distinction between "Variety" and "variety" in the construction of corpora. He proposes 'Variety' to mean geographical varieties of language (e.g., Irish English, British English, etc.), while 'variety' refers to discourse genres defined by situational use (e.g. academic discourse, workplace language, etc.). This paper concerns spoken corpora of the Variety kind that are also designed to represent 'varieties'.

In order to portray language variation in a more fine-grained manner at the discoursal level, we argue that the current conceptualization of spoken corpora be replaced with discourse corpora (see also Santamaría-García, 2011), and that genre metadata include socio-cultural activity types, discourse topics, and speech acts so as to achieve fine-grained representation of interaction. The paper develops its argument in the context of two corpora on contemporary dialects of Turkish, namely, the Spoken Turkish Corpus – STC (Turkish spoken in Turkey) and the Spoken Turkish Cypriot Dialect Corpus – STCDC (Turkish in Northern Cyprus). The paper first presents an overview of certain aspects of Variety corpora features that impact variation research, and discusses the importance of developing parallel genre sampling frames that can capture communication as situated interaction. The last part of the paper discusses issues concerning the representation of variation, with a focus on corpus tool desiderata standardizing transcription, and annotation parameters so as to ensure gains in comparative research even on corpora designed for different purposes.

2. Designing 'Variety' Corpora

Spoken corpora as Variety have been compiled in different ways with respect to the nature of interaction types. Those with a particular interest in regional dialectal variation have often relied on elicited speech (e.g., the Freiburg English Dialect Corpus – FRED). Others such as the International Corpus of English – ICE, incorporate Varieties of English and samples of unprompted language in public and private settings. Corpora based on elicited speech have the advantage of allowing the researcher to 'guide' the interaction so as to document inter-speaker variation. However, they would lose out on the capacity to explore intra-speaker variation in naturally occurring communications, unless corpora are designed to include elicited speech. Curating elicited speech arguably has the advantage of controlling for genre, but in line with Čermák (2009), we maintain that spoken corpora –and indeed discourse corpora– necessitate a predominance of natural speech (Ruhi et al., 2010). Thus for discourse-oriented spoken corpora we would highlight the significance of multi-party, spontaneous and high interaction speech (Čermák, 2009: 118) if we are to represent the full quality of discourse in 'real life'. In this regard, elicited speech largely limits the kind of sociolinguistic and pragmatic variation analysis that can be conducted, such as accommodation processes and situation-bound form-function mappings.

Inclusion of a range of genres in Variety corpora is a standard implementation procedure, but a problematic feature at the discourse level concerns the common practice of selecting equal lengths of texts. This means that texts may be cut at appropriate points in the discourse or combined to form "composite[s]" from shorter interactions (e.g. ICE; Greenbaum and Nelson, 1996: 5). Such sampling makes it impossible to conduct cross-varietal research on genres and conversational organization (Andersen, 2010: 556), thereby largely limiting. Other corpora, however, have aimed to include complete interactions wherever feasible (e.g., BNC).

Genre classification in spoken corpora is yet another problematic area (Lee, 2001) that has a bearing on variation analysis. This is partly due to the fact that spoken interaction

is often fluid in terms of communicative goals. Corpus compilers have tackled classification in slightly differing manners and levels of granularity. While BNC lumps casual conversation a single category, the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) differentiates such speech events along two axes: the relational and goal dimensions (McCarthy, 1998:5). One problem with this scheme is the rigid divide it proposes for goal classification and in some cases for speaker relations. Conversations in real life may be either task or idea oriented at different specific times (Gu, 2010) or include differing social relations within the same situated discourse. One way of resolving the multi-dimensionality and fluidity of discourse in corpus files would be to include social activity types in metadata.

The discussion above has raised a number of issues in spoken corpus compilation, focusing on certain design features. In the following, we highlight aspects of speaker demographics, and the employment of domain and social activity criteria for genre classification, which is supplemented with discourse topic and speech act annotation in metadata.

3. Corpus Design, Pragmatic Metadata, and Sampling Procedures in STC and STCDC

The construction of STC and STCDC as web-based corpora started at different but close times at the campuses of Middle East Technical University in Ankara and Güzelyurt (in 2008 and 2010, respectively; http://www.stc.org.tr/en/; http://corpus.ncc.metu.edu.tr/?_lang=en). STC will function as a general corpus for spoken Turkish in Turkey. It thus is a Variety corpus that includes demographic variation and different genres. STCDC is also a Variety corpus that includes different genres, but it has a regional dialect slant. Despite the dialectal focus of STCDC, the two corpora may be used for variation analysis because their sampling frames are similar and because they employ the same speaker metadata and genre classification parameters (see Andersen, 2010: 557).

With respect to speaker metadata, both STC and STCDC include standard demographic categories such as place of birth and residence, age, education, etc. In addition to these, speakers are described according to length of residence in different geographical locations (see Ruhi et al. (2010c) for a fuller description of the metadata features in STC). Both corpora also document the language profiles of the speakers. This will allow sifting of speakers according to these parameters in future dialect research.

STC and STCDC employ two parameters in genre classification: speaker relations and social activity type. The major domains are family, friend, family-friend, educational, service encounter, workplace, media discourse, legal, political, public, research, brief encounter, and unclassified. Finer granularity in metadata is achieved during the corpus assignment and the various steps in the transcription stage in STC through annotation of speech acts (e.g. criticizing), on the one hand, and on the other hand, annotation for conversational topics (e.g. child care), speech events (e.g. troubles talk), and ongoing social activities (e.g., cooking). In the current state of STC, speech acts are annotated as a separate category from topic annotation categories, but Topics as a super-category includes local conversational topics and discoursal topics, which, as evident in the examples given above, concern the meso-level of discourse in that they are more discourse activities than topics in the strict sense (see Levinson (1979) on activity types).

The inclusion of speech acts and topics as two further parameters in metadata is motivated by the assumption that combining generic metadata with more specific pragmatic annotation renders corpora more 'visible' for pragmatics research, broadly construed here to include variation analysis. Conflating local and global topics in a single category may not be the ideal route. However, given that speech event, activities and discoursal topic categorizations are fuzzy notions,[1] we submit that separating them from local topics would not add to the worth of the annotation (Speech act and topic annotation will be implemented in STCDC Project at a later stage.). Figure 1 is a partial display of the metadata for a file in STC:

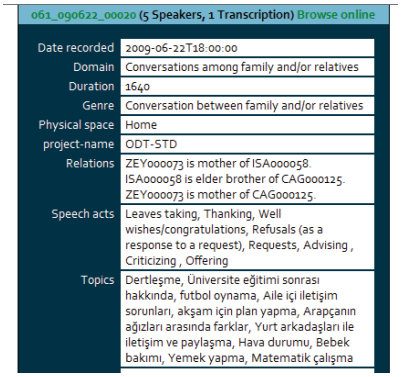


Figure 1. Partial metadata for a communication in STC

This metadata annotation design is an improvement for pragmatics-oriented research on spoken corpora, where scholars have underscored the difficulty of retrieval of speech act realizations through standard searches that more often than not rely on the extant literature on speech acts in various languages (see, e.g., Jucker et al. (2008) for an investigation of compliments in BNC). Identifying speech act realizations during the corpus construction stage can enhance corpus-driven approaches to pragmatics (broadly construed here to include variation analysis) by providing starting points for researchers on where to look for the relevant realizations in the 'jungle' of corpus data. As has been noted scholars that critically assess the possibility of doing corpus-oriented discourse analysis and pragmatics

[1] See Adolphs, Knight & Carter (2011) for related remarks on classifying activities in corpus.

research (e.g., Virtanen, 2009), data retrieval and analysis can be overwhelming. We propose that there is a need to achieve a modicum of practice that garners the best of qualitative and quantitative methodological practices, and that one way to do this is to annotate a minimum of pragmatic and semantic information in corpora (see, also, Adolphs, Knight & Carter (2011) for an implementation of activity type annotation). Below, we expand on the motivation for implementing a minimum level of speech act and topic annotation, taking up four issues regarding general corpora design, researching language in use, and the case of languages that have not been the object of pragmatics research to the extent that we observe in languages such as English and Spanish.

First, to our knowledge, the field of spoken corpus construction is yet to see the publication of large-scale general corpora that are annotated for speech acts with the analytic detail that is displayed in pragmatics research (e.g., full annotation of head-acts, supportive moves, and so on). As noted in Santamaría-García (2011), too, there appears to be little sharing of pragmatically (partially) annotated corpus, with the result that corpus-oriented research is limited in terms of sharing knowledge and experience in the field. Naturally, it is questionable whether general corpora construction should aim for full annotation of speech acts or other discursive phenomena that go beyond the utterance level, given that these are is time-consuming, expensive enterprises, which also carry the risk of weakening annotation consensus (see Leech, 1993). Nonetheless, there is growing interest to employ general or specialized corpora for pragmatics research. We would argue that while listing speech act occurrences certainly does not replace full-scale annotation, rudimentary, list type annotations open the way to enhancing qualitative methodologies that aim to combine them with quantitative approaches.

The second issue concerns corpus compilation research. Topic and speech act annotation aids monitoring for variation in genre, tenor and the affective tone of interaction during corpus construction (Ruhi et al., 2010a). Monitoring for contextual feature variation are too broad notions to achieve representativeness in regard to the 'content' level of language, especially in casual conversational contexts, which exhibit great diversity in conversational topics. While such annotation naturally increases corpus compilation effort, it is useful for studies in pragmatics and dialectology, where it is observed that stylistic variation is controlled by discoursal and sociopragmatic variables (see Macaulay, 2002).

The third issue is related to the fact that pragmatic phenomena (e.g. negotiation of communicative goals and relational management, and inferencing) are not solely observable as surface linguistic phenomena. Despite its limitations, speech act annotation is a viable starting point for exploring these dimensions of language variation. Finally, for the less frequently studied languages in traditional discourse and pragmatics research, there is often little if none by way of research findings that corpus-oriented research can rely on to retrieve tokens of pragmatic phenomena at the linguistic expression level. Turkish is one typical example of such a language, where, for instance, only a handful of studies exist on directives (see Ruhi, 2010). In this regard, speech act and topic annotation can aid in implementing the necessarily corpus-driven methodologies for pragmatics research in such languages. Stocktaking the issues raised concerning the inclusion of what are arguably qualitative metadata features, we maintain that the parameters described above respond to the need for corpora that are pragmatically annotated.

STC and STCDC are similar in regard to sampling procedures. Both projects recruit recorders on a voluntary basis, and selections are based on stratified sampling methods. Even though STCDC is currently much smaller in size and narrower in genre variety, the two corpora will eventually achieve comparability even if at different scales (250, 000 words in STCDC, and 1 million in STC in its initial stage). With respect to the discursive dimensions of interaction, since the two corpora aim at compiling complete conversations, analysis of pragmatic variation will be possible.

4. Corpus Construction Tools in STC and STCDC

Both STC and STCDC are transcribed with the EXMARaLDA Software Suite (Schmidt & Wörner, 2009), which facilitates research on variation owing to the nature of its tools. The transcription tool, Partitur-Editor, enables multiple-level annotation and files are time-aligned with audio and video files. In its simplest form, each speaker in a file is assigned two tiers –one for utterances (v-tier) and one for their annotation (c-tier) (see Figures 2a and 2b for the annotation of dialectal and standard pronunciation of words). Each file also has a no-speaker tier (nn), where background events may be described (e.g., ongoing activities that impinge on conversation). This structure is especially essential in variation analysis, as researchers may themselves consult the original data against possible errors in transcription, or play and watch segments as many times as necessary.

OSK000011[v] **fagat** şimdi ((0,6s)) o... ((0,3s))
OSK000011[c] ***fakat***
[nn] ((noise))

Figure 2a: Transcription of the /k/-/g/ variation in STCDC

HAT000068 [v] ((0.3))
SAT000069 [v] kaba mı katalım?
SAT000069 [c] gaba mı gatalım

Figure 2b: Transcription of the /k/-/g/ variation in STC

With EXAKT, EXMARaLDA's search engine, statistical data such as frequencies can be obtained. However, EXAKT is not only a concordancer but also a device for stand-off

annotation. Researchers have often criticized corpus metadata models that present information only in file headers (e.g. Archer, Culpeper & Davies, 2008). By enabling access to metadata at the site of the tokens retrieved, EXAKT responds to this crucial need in variation analysis. Figure 3 illustrates a token of the word *tamam* 'alright, okay' in STC, along with a select number of metadata categories (Metadata criteria can be selected according to research questions.). Figure 4 displays a stand-off annotation of another token of *tamam* for utterance and speech act function.

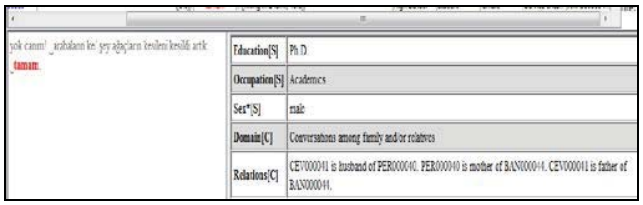


Figure 3: Token and selected conversation features

Figure 4: Utterance and speech act type

5. Transcription Standardization in STC and STCDC

Transcription of language Varieties has been handled in two ways in current corpora and in disciplines such as conversation analysis and discourse analysis, namely,

> Either producing dialectal, orthographic transcription which is as close to the pronunciation as possible, or producing a transcription based on a written standard, but usually allowing for some variation.
> Andersen (2010: 554)

As is current practice in general corpora, STC follows the second approach described in Andersen (2010) (see, e.g., BNC). Standard orthography is employed in the v-tier, except for a limited list of words consistently pronounced in different ways from the so-called careful pronunciation. For instance, the word *hanımefendi* 'lady' is usually pronounced with the deletion of *ım*e. Thus, the word is written either as *hanımefendi* or *hanfendi* in the v-tier depending on speaker pronunciation (for documentation of spelling variation in STC see Ruhi et al. (2010b) and the corpus web site for documents on technical details). Prominently marked dialectal and stylistic variants of words and morphological realizations are transcribed in the c-tier (see Figure 5a). As stylistic variation can interleave with dialect shifts, representing both the actual pronunciation and its standard form provides crucial information for investigating pragmatic variation.

In STCDC, standardized dialectal orthography is employed in the v-tier and standardized orthography of prominent words is written in the c-tier. STC and STCDC transcriptions are thus like mirror images of each other (compare, Figures 2a,b and 5a,b).

STC and STCDC share the same transcription system adapted from HIAT (Ruhi et al., 2010b). A major gain of this procedure will be the possibility to compare dialectal and standard forms in a unitary fashion across Turkey and Northern Cyprus. STCDC will also inform standardization in the compilation of special, dialectal computerized corpora for Anatolian dialects, owing to the fact that dialects in Cypriot Turkish share an extensive number of variation patterns despite variation especially in certain inflectional morphemes (Saraçoğlu, 2004).Figures 5a and 5b illustrate the transcription of the /t/-/d/ and /k/-/g/ variation in STCDC and STC, respectively.

BUR000030 [v] ((0.6)) geç içeri. gel. ((laughs)) ˙
MUS000031 [v] **kız!**
MUS000031 [c] gı
IND000002 [v] ((XXX))
[nn]

Figure 5a: Transcription of /k/-/g/ variation in STC

BUR000002 [v] ((0,8s)) ((OKA)) abiler daşındıydı burdan yoksa
BUR000002 [c] ***taşındıydı ((lengthening))***
FIK000003 [v]
FIK000003 [c]
[nn] *background))*

Figure 5b: Transcription of /t/-/d/ variation in STCDC

For dialectal variation studies, the transcription methodology described above is supported through EXAKT. The tool automatically produces word lists from the v-tier, which can then be selected for search in the corpora. Such searches reveal instances of variation in situ, as can be seen in Figures 2a,b and 5a,b.

Commonality in standardization in STC and STCDC is not restricted to solely the transcription of the word. The two corpora also employ similar annotation conventions for discursively significant paralinguistic features and non-lexical contributions to the conversation (see, e.g., the superscript dot in Figure 5a for laughter). Figure 5b illustrates the lengthening of phonetic units in words, which is described in the c-tier in both corpora. This methodology is expected to ease cross-varietal pragmatic research.

6. Concluding Remarks

This paper has discussed how comparability was achieved in a general corpus and a dialectal corpus in the context of STC and STCDC. Parallel corpora utilizing a standardized system for sampling, recording, transcription and annotation potentially ensures cross-linguistic analyses for researchers. We highlight the importance of moving toward the construction of discourse corpora for language variation research by enriching metadata designs that can enhance corpus-based/corpus-driven cross-varietal research in pragmatics and related fields. At a more specific level, we hope that the corpus design and its implementation in STC and STCDC will move forward the field of corpus construction for Turkish and Turkic languages.

7. Acknowledgments

STC was supported by TÜBİTAK 108K283 between 2008-2010, and is currently being supported by METU BAP-05-03-2011-001. STCDC is being supported by METU NCC BAP-SOSY-10. We thank the reviewers for their comments, which have been instrumental in clarifying our arguments. Needless to say, none are responsible for the present form of the paper.

8. References

Andersen, G. (2010). How to use corpus linguistics in sociolinguistics: In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*. London/New York: Routledge, pp. 547--562.

Archer, D., Culpeper, J. and Davies, M. (2008). Pragmatic annotation. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*, *Vol. I*. Berlin/New York: Walter de Gruyter, 613--642.

Clancy, B. (2010). Building a corpus to represent a variety of a language. In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*: Routledge: London/New York, pp. 80--92.

Čermák, F. (2009). Spoken corpora design: their constitutive parameters. *International Journal of Corpus Linguistics*, 14(1), pp. 113--123.

Greenbaum, S., Nelson, G. (1996). The International Corpus of English (ICE) Project. *World Englishes*, 15(1), pp. 3-15.

Jucker, A., Schneider, G., Taavitsainen, I. and Breustedt, B. (2008). "Fishing" for compliments. Precision and recall in corpus-linguistic compliment research. In A. Jucker & I. Taavitsainen (Eds.). *Speech Acts in the History of English*. Amsterdam/Philadelphia: Benjamins, pp.273--294.

Lee, David, YW. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3), 37--72.

Levinson, S. (1979). Activity types and language. *Language*, 17, pp.365--399.

Macaulay, R. (2002). You know, it depends. *Journal of Pragmatics*, 34, pp. 749--767.

McCarthy, M. (1998). *Spoken languages and applied linguistics*. Cambridge: Cambridge University Press.

Ruhi, Ş. (2011) 'LOOK' as an Expression of Procedural Meaning in Directives in Turkish: Evidence from the *Spoken Turkish Corpus*. Paper presented at the *Sixth International Symposium on Politeness*, 11-13 July, 2011, Middle East Technical University, Ankara.

Ruhi, Ş., Işık-Güler, H., Hatipoğlu, C., Eröz-Tuğa, B. and Çokal Karadaş, D. (2010a). Achieving representativeness through the parameters of spoken language and discursive features: the case of the Spoken Turkish Corpus. In: Moskowich-Spiegel Fandino, I., Crespo García, B., Lareo Martín, I. (Eds.), *Language Windowing through Corpora. Visualización del lenguaje a través de corpus. Part II.* A Coruña: Universidade da Coruña, 789--799.

Ruhi, Ş., Hatipoğlu, Ç., Eröz-Tuğa, B. and Işık-Güler, H. (2010b). *A guideline for transcribing conversations for the construction of spoken Turkish corpora using EXMARaLDA and HIAT.* ODTÜ-STD: Setmer Basımevi.

Ruhi, Ş., Eröz-Tuğa, B., Hatipoğlu, Ç., Işık-Güler, H., Acar, M. G. C., Eryılmaz, K., Can, H., Karakaş, Ö. and Çokal Karadaş, D. (2010c). Sustaining a corpus for spoken Turkish discourse: Accessibility and corpus management issues. In *Proceedings of the LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*. Paris: ELRA, 44-47. http://www.lrec.conf.org/proceedings/lrec2010/workshops/W20.pdf#page=52

Ruhi, Ş., Schmidt, T., Wörner, K. and Eryılmaz, K. (2011). Annotating for Precision and Recall in Speech Act Variation: The Case of Directives in the *Spoken Turkish Corpus.* In *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011. Working Papers in Multilingualism* Folge B, 96, 203--206.

Santamaría-García, C. (2011). Bricolage assembling: CL, CA and DA to explore agreement. *International Journal of Corpus Linguistics,* 16(3), pp. 345--370.

Saraçoğlu, E. (2004). *Kıbrıs ağzı: Sesbilgisi özellikleri, metin derlemeleri, sözlük***.** Lefkoşa: Ateş Matbaacılık.

Schmidt, T., Wörner, K. (2009). EXMARALDA – creating, analysing and sharing spoken language corpora for pragmatic research**.** *Pragmatics,* 19(4), pp. 565--582.

Virtanen, T. (2009). Corpora and discourse analysis: In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook, Volume 2*. Berlin/New York: Walter de Gruyter, pp. 1043--1070.

The Corpus of Spoken Greek: Goals, Challenges, Perspectives

Theodossia-Soula Pavlidou
Aristotle University of Thessaloniki
GR-54124 Thessaloniki, Greece
pavlidou@lit.auth.gr

Abstract

The purpose of the present paper is to introduce the *Corpus of Spoken Greek,* which has been developed at the Institute of Modern Greek Studies (Manolis Triandaphyllidis Foundation), Aristotle University of Thessaloniki. More specifically, I would like to describe and account for the particularities of this corpus, which is mainly intended for qualitative research purposes from the perspective of Conversation Analysis. I will thus exemplify some of the issues and challenges involved in the development of corpora that consist of naturalistic speech data. As a consequence, I would like to conclude that the idea of "best practices" for speech corpora can only be understood as a function of the research goals, explicit or implicit, that are prominent when compiling a corpus. Moreover, standardization of speech corpora can only be pursued to an extent that allows, on the one hand, comparability with other corpora and usability by a large community of researchers, and, on the other, ensures maintenance of those characteristics that are indispensable for the kind of research that the corpus was originally conceived for.

Keywords: Modern Greek, speech, Conversation Analysis

1. Introduction

While the significance of studying language on the basis of speech has been undisputed for more than a century now and while the necessity for corpora has been long recognized, the international praxis of *speech* corpora is only slowly catching up with that of *written* corpora. This is no coincidence, if one takes into account that the compilation of spoken data is much more time-consuming and expensive as it relies to a greater extent on technical equipment. Moreover, if 'naturalness' of the data is one of the goals, the so-called observer's paradox (more precisely: the attempt to overcome this paradox) often has the consequence that speech corpora comprise public discourse genres, e.g. broadcast lectures, talk shows, news bulletins, etc. Other types of discourse, like everyday conversations or telephone calls, are much more sensitive to the kind of recording (audio vs. video) employed, as regards naturalness or authenticity; moreover, they are much less accessible (e.g. with respect to the participants' consent to record their interaction and use the recorded material). It is no surprise then that speech corpora based on naturalistic data are relatively rare and small in comparison to written corpora.

In this paper, I would like to discuss some of the problems involved in the development of corpora that consist of naturalistic speech data by way of presenting the *Corpus of Spoken Greek* of the Institute of Modern Greek Studies (Manolis Triandaphyllidis Foundation), Aristotle University of Thessaloniki. More specifically, I would like to describe and account for the particularities of this corpus which are closely related to its aims, but also to the specific conditions under which it was developed, and point to some issues and challenges.

2. Background

The compilation of a corpus of naturally occurring talk-in-interaction is one of the aims of the research project *Greek talk-in-interaction and Conversation* Analysis,[1] which is carried out under the author's direction at the Institute of Modern Greek Studies. The project additionally aims at the study of the Greek language from the perspective of Conversation Analysis and the training of researchers in the theory and practice of Conversation Analysis (in the following, CA for short), an ethnomethodologically informed approach to linguistic interaction (for the aims, principles, methodology, etc., of CA, cf. e.g. Lerner, 2004; Schegloff, 2007). Given the CA orientation of the project, the *Corpus of Spoken Greek* is primarily intended for close qualitative rather than quantitative analyses, and it is this objective that lends the corpus its particular characteristics (cf. Section 3). However, as we shall see in Section 5, part of the corpus can also be used for quantitative analyses online.

In its current form (with respect to its conceptualization and more or less stable composition of the team through the employment of part-time assistants), the project *Greek talk-in-interaction and Conversation Analysis* has been running for about four years. However, the *Corpus of Spoken Greek* did not arise out of nowhere nor did it get designed at one shot. Rather, it utilized earlier data collections (and transcriptions) carried out by the author

[1] Website: <http://ins.web.auth.gr/en/ylikoelectr/Corpus.html>

in various research and/or student projects since the 1980s, mostly without any funds. In particular, an earlier corpus of approximately 250.000 words (cf. Pavlidou, 2002) has been incorporated into the current one. It is furthermore important to stress that the *Corpus of Spoken Greek* has not been intended as a "closed" database, but as a dynamic corpus in that it gets enriched with new recordings and transcriptions, while older transcriptions are re-examined and eventually improved.

3. Features

As has already been indicated, the *Corpus of Spoken Greek* is mainly intended for qualitative analyses of the Greek language and, more specifically, for the study of Greek talk-in-interaction from a CA perspective. Accordingly, in the compilation of the Corpus, particular emphasis has been placed on data from everyday conversations in face-to-face interactions or over the telephone. In addition though to informal conversations among friends and relatives, the Corpus also includes other discourse types, which are more institutional, like recordings of teacher-student interaction in high school classes, television news bulletins, and television panel discussions (cf. Table 2 for the size of data from each discourse type). In sections 3.1 to 3.3 further information with regard to the collection of data, transcription, files and metadata is provided.

3.1 Data Collection

The *Corpus of Spoken Greek*, as mentioned, utilized earlier data collections and transcriptions (cf. Section 2). In all cases, however, we have to do with naturalistic data that were collected for the most part by MA or PhD students for their theses or by undergraduate students for their semester work. The informal conversations were tape- or video-recorded by one of the participants. In the case of classroom interaction, it was the teacher her-/himself who made recordings of his/her classes at high school. The equipment used for the recordings has been varying over time, as in the beginning only simple tape- or video-recorders were available, while later on digital recorders could be employed.

Recordings of private conversations and classroom interactions were made openly and after having informed the participants that they would be recorded and getting their consent. In the case of telephone conversations, sometimes this information was given after the telephone calls were conducted. In all cases, the persons who made the recordings were asked to erase anything they wanted and to hand in only those conversations/interactions they would not mind being heard or seen by others. The issue of participants' consent, though, has been handled with much more rigor and detail in the last 15 years, as there is both a growing sensitivity in the Greek society and an official Data Protection Authority for the safeguarding of personal data in Greece. We have therefore been using a written consent form that is signed by all participants in those interactions which are not transmitted publicly.

3.2 Transcription

As is well known, CA lays great emphasis on the detailed representation of spoken discourse via the transcription of the recordings. For CA, transcription is not an automatic, mechanical procedure, of the kind, for example, accomplished by various software packages, nor is it confined to the representation of content, as is usually done in the written form of interviews by journalists (cf. also Ochs, 1979). On the contrary, the 'translation' of sound into written text requires theoretical elaboration and analysis, presupposes training, and demands multiple examination/corrections by several people.

The recordings collected for the *Corpus of Spoken Greek* are therefore meticulously transcribed according to the principles of CA (cf. e.g. Jefferson, 2004; Sacks, Schegloff and Jefferson, 1974; Schegloff, 2007) in several rounds by different people. This is basically an orthographic transcription, in which the marking of overlaps, repairs, pauses, intonational features, etc., is carried out in a relatively detailed manner. To this we add the marking of certain sandhi phenomena and dialectal features. For the disambiguation of prosodic features (e.g. sudden voice uprise), we employ the Praat software. Finally, transcriptions are produced as Word documents, using a table format that allows different columns for marking the participants' names, the numbers of lines (when necessary), etc. An extract of such a transcription is illustrated in Table 1 (to which the English translation has been added):[2]

Dimos	[…] έπαιξε όλα τα:- αυτά που ακούγαμε στα
	[…] he played all the- those songs we listened to when
	δεκαοχτώ, (1.3) και μας κοιτούσε με κριτικό
	we were eighteen, (1.3) and he looked at us with a
	[μάτι. (γιατί δε τ' ακούς αυτά.)]
	critical eye. (why don't you listen to these.)
Afrod.	[Συγγνώμη, τα ίδια ακού[γα]με (όταν ήσασταν)]=
	Excuse me, did we listen to the same songs (when you
Yorgos	[((*γελά*.....]=
	((*he laughs*……..
Afrod.	=[δεκαοχτώ? πού θες να ξέρω τι ακούγατε.]
	((*γελώντας*))
	were) eighteen? how should I know what you listened to.
	((*in a laughing tone*))
Yorgos	=[.....))]
	.....))

[2] 'Afrod.' stands for Afroditi, the name of a female participant. For the symbols used in the transcription please cf. the Appendix in Section 8.

	Δηλα[δή (όπως .....)]=
	You mean (like .....)
Dimos	[E: έκφραση είναι.]=
	U:h it's just an expression.
Yorgos	=[(σ' αυτή τη) δουλειά, τι εννοείς με κριτι]κό μάτι?
	(in this) job, what do you mean with a critical eye?
Manos	=[Α υ τ ά τα: ποιοτικά (π ο υ α κ ο ύ ν).]
	Those quality songs (they listen to).
	(.)
Afrod.	.hh Ο Δημήτρης [ας πούμε άκουγε >τι άκουγε.<]
	((*γελαστά*.....))
	.hh Dimitris for instance listened to what did he listen to.
	((*in a laughing tone*……..))

Table 1: Extract from Conversation I.14.A.20.1

As can be seen from Table 1, phenomena like overlaps, pauses and other prosodical features are marked in a relatively detailed manner. On the other hand, annotating, for example, the 'sequential organization' of the above excerpt, i.e. what the basic 'adjacency pair' is, what its 'expansions' are, and so on, but even things that one might think of as simpler (e.g. 'turns', 'increments', etc.) would require even deeper theoretical analysis than the process of transcribing itself entails.[3] It goes without saying that this would exceed the scope and aims of a speech corpus; after all the particular Corpus is compiled in order to serve as a tool for such an analysis.

On the other hand, the annotation or tagging of easily quantifiable linguistic categories (e.g. parts of speech) is not a relevant objective for our project at the moment. As for the tagging of social/pragmatic categories (e.g. speech acts), such a practice would run counter to the CA principles, unless it is the result of an in-depth theoretical elaboration that establishes, among other things, the relevance of these categories for the participants themselves. But as already noted the aim of the Corpus is to serve as a tool for such qualitative analyses, rather than provide the analysis itself.

3.3 Files and Metadata

The *Corpus of Spoken Greek* comprises five different types of digital files, stored on CDs or DVDs:

1) audio-recordings,
2) video-recordings,
3) transcriptions of the recordings without any metadata (as Word documents),
4) transcriptions of the recordings with all the metadata (as Word documents),
5) transcriptions of certain recordings, without any metadata, in html format for the online word search (cf. Section 5).

In addition, printouts of the transcribed texts are kept in separate folders.

The reference code for each file involves five variables indicating: the type of discourse, the year in which the recording was produced, the type of recording (i.e. audio vs. video), the person who made the recording, and the number of the recording (if more recordings by the same person are available). In the case of telephone calls, classroom interaction and TV news bulletins or panel discussions, a sixth variable is added to indicate, e.g. whether the call is made on a cell phone or what high school level ('gymnasium' vs. 'lyceum') the interactions come from or which TV channel broadcast the news bulletin, and so on.

As for the metadata, their exact nature varies depending on the type of discourse involved. For example, for the face-to-face conversations among friends and relatives the metadata comprise:

a) the names of the participants and the pseudonyms employed in the transcription,
b) their age,
c) the relationship to one another,
d) occupation,
e) their place of residence during the last five years at the time of the recording,
f) the place they come from,
g) any special features in their linguistic behavior (e.g. dialectal accent),
h) where/when the recording was made,
i) name of the person that made the recording,
j) names of the original plus subsequent transcribers.

4. Current Size and Types of Discourse

The current size of the *Corpus of Spoken Greek* amounts to approximately:

72,853 MB of audio-recordings and
105,309 MB of video-recordings.

The transcribed files exceed 1,7 million words, distributed across the following discourse types:

DISCOURSE TYPES	**NUMBER OF WORDS**	**%**
conversations among friends and relatives	566,977	33.0%
TV news bulletins	535,421	31.2%
teacher-student interactions	305,222	17.8%
telephone calls	189,349	11.0%
TV panel discussions	119,466	7.0%
TOTAL	**1,716,435**	**100%**

Table 2: Discourse Types and Number of Words

[3] An example of such an annotation can be found in Pavlidou (in press).

As previously mentioned, the degree of detail and quality of the transcription varies, depending on the number of people working on the data and their training, but also according to the research needs of the project and the necessities that have arisen (cf. e.g. the preparation of part of the Corpus for word search online described in the next Section). Consequently, not all transcribed texts appearing in Table 2 are on the same level. In particular, most transcriptions of face-to-face conversations have been reworked on the average by three to four different persons. TV panel discussions, on the other hand, have only undergone transcription by one to two persons.

The *Corpus of Spoken Greek* is, thus, unique both in its aims/conception but also in its make-up, as compared to the other four existing Greek corpora (of which only one contains spoken discourse files), since a great part of it (ca. 44%, i.e. more than 750,000 words) comes from informal (face-to-face or telephone) conversations.

5. Extensions

As already mentioned, the *Corpus of Spoken Greek* has been primarily designed for the qualitative analysis of language and linguistic communication from a CA perspective. However, a tool has recently been developed for the search of words and phrases, concordances, statistics, etc., in a database consisting of informal conversations among friends and relatives that amounts to 200,000 words. This tool, which is available online,[4] allows the user to track a word, irrespective of modifications of its form due to the transcription conventions. Figure 1 illustrates the main page of the tool for quantitative analyses:

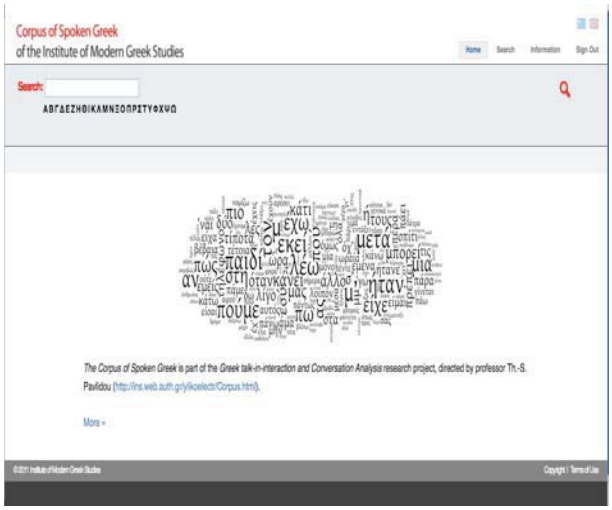


Figure 1: Main page of the tool for quantitative analyses

When searching for a certain word, e.g. *καλά* ('good', 'well'), the results are listed in contextual environments consisting of three lines of transcribed text. These lines may contain utterances belonging to different speakers (symbolized in different color icons before every line). Figure 2 illustrates two such contextual environments for the word *καλά*. It also shows the total number of its occurrences –617 in this case.

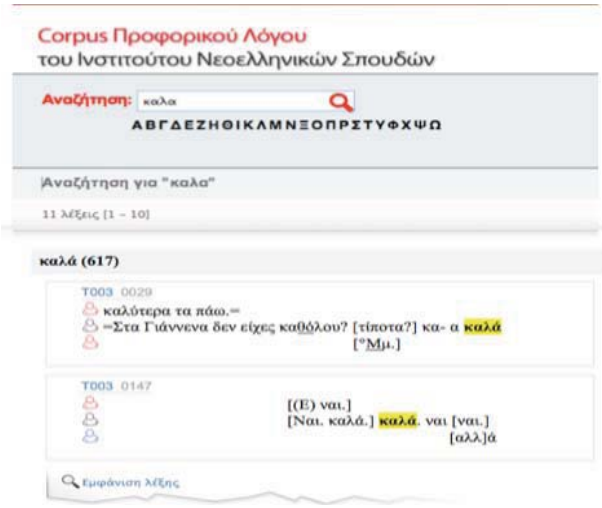


Figure 2: Search results

The search for a word, e.g. *καλά* ('good', 'well'), will also yield all the words beginning with the letter sequence *κ-α-λ-α*, e.g. *καλάμι* ('reed'), *καλάθι* ('basket'). By using the asterisk (*), e.g. **καλά** or **καλά*, the search will also yield all the words containing the letter sequence *κ-α-λ-α* or ending in it. Finally, the search for pairs of words, e.g. *καλά καλά* or *πολύ καλά*, is also possible.

Finally, some statistics can be provided, as for example the ten most frequent words in this part of the Corpus shown in Figure 3:

το	6105
να	5827
και	5807
ναι	4193
είναι	3335
δεν	3232
ε	3226
θα	3022
τα	2457
που	2146

Figure 3: The ten (10) most frequently used words

In addition to the tool for quantitative purposes, we are currently in the process of assessing alternative approaches to coding multimodal data (e.g. CLAN, ELAN, EXMARaLDA), so that the available video recordings can be adequately transcribed and managed (cf. e.g. Mondada, 2007; Schmidt & Wörner, 2009) in ways best suited to the purposes of the *Corpus of Spoken Greek*.

[4] Website: <http://corpus-ins.lit.auth.gr/corpus/index.html>

6. Availability

The *Corpus of Spoken Greek* is available to scholars for research purposes, but not online. Scholars interested in qualitative analysis can contact the author for the conditions pertaining to access and use of parts of the Corpus.

What is available online is the tool for word search, etc. (cf. Section 5). In other words, part of the corpus, consisting of informal conversations among friends and relatives can be accessed online (cf. footnote 4) for quantitative analyses.

7. Conclusion

The discussion above has hopefully shown how certain features (naturalistic data, detailed transcriptions of everyday conversations) of the *Corpus of Spoken Greek* derive from the purposes for which it was developed (qualitative analysis, CA perspective) and how they get modified by the real-life conditions (funding, technical affordances, experienced staff) under which they have to be implemented.

In conclusion, then, I would like to suggest that the idea of "best practices" for speech corpora can only be understood as a function of the research goals –explicit or implicit– that are prominent when compiling a corpus. It is these goals that inform the features and particularities of a specific corpus, thus rendering it eventually a/the "best" tool for specific purposes under specific conditions, but less suitable for others. Consequently, standardization of speech corpora can only be pursued to an extent that allows, on the one hand, comparability with other corpora and usability by a large community of researchers, and, on the other, ensures maintenance of those characteristics that are indispensable for the kind of research that the corpus was originally conceived for.

8. Appendix

List of Transcription Symbols used in Table 1

[**[**	left brackets: point of overlap onset between two or more utterances (or segments of them)
] **]**	right brackets: point of overlap end between two or more utterances (or segments of them) To prevent potential confusion over the temporal sequence of the overlapping segments double brackets are used in some cases.
=	The symbol is used either in pairs or on its own. A pair of *equals signs* is used to indicate the following: 1. If the lines connected by the equals signs contain utterances (or segments of them) by different speakers, then the signs denote 'latching' (that is, the absence of discernible silence between the utterances). When latching occurs in overlapping utterances (or segments of them) of more than two speakers, then an equals sign is added to every additional line containing those utterances (or segments of them). 2. If the lines connected by the equals signs are by the same speaker, then there was a single, continuous utterance with no break or pause, which was broken up in two lines only in order to accommodate the placement of overlapping talk. The single *equals sign* is used to indicate latching between two parts of a same speaker's talk, where one might otherwise expect a micro-pause, as, for instance, after a turn constructional unit with a falling intonation contour.
(0.8)	Numbers in parentheses indicate silence, represented in tenths of a second. Silences may be marked either within the utterance or between utterances.
(.)	micro-pause (less than 0.5 second)
punctuation marks	indication of intonation, more specifically:
.	the period indicates falling/final intonation
?	the question mark indicates rising intonation,
,	the comma indicates continuing/non-final intonation
:	Colons are used to indicate the prolongation or stretching of the sound just preceding them. The more colons, the longer the stretching.
-	A hyphen after a word or part of a word indicates a cut-off or interruption.
>word<	The combination of 'more than' and 'less than' symbols indicates that the talk between them is compressed or rushed.
.h	If the aspiration is an inhalation, then it is indicated with a period before the letter h.
((*laughs*))	Double parentheses and italics are used to mark meta-linguistic, para-linguistic and non-conversational descriptions of events by the transcriber.
(word)	Words in parentheses represent a likely possibility of what was said.
[…]	Dots in brackets indicate a strip of talk that has been omitted.

9. References

Corpus of Spoken Greek, Institute of Modern Greek

Jefferson, G. (2004). Glossary of Transcript Symbols with an Introduction. In G.H. Lerner (Ed.), *Conversation Analysis: Studies from the First Generation*. Amsterdam/Philadelphia: John Benjamins, pp. 13-31.

Lerner, G. (ed.) (2004). *Conversation Analysis: Studies from the First Generation*. Amsterdam/Philadelphia: John Benjamins.

Mondada, L. (2007). Commentary: Transcript Variations and the Indexicality of Transcribing Practices. *Discourse Studies* 9(6), pp. 809-821.

Ochs, E. (1979). Transcription as Theory. In E. Ochs & B. Schieffelin (Eds.), *Developmental Pragmatics*. New York: Academic Press, pp. 43–72.

Pavlidou, Th.-S. (2002) [in Greek]. GR-Speech: Corpus of Greek Spoken Discourse Texts. *Studies in Greek Linguistics* 22, pp. 124-134.

Pavlidou, Th.-S. (in press). Phases in Discourse. In A. Barron & K. Schneider (Eds.), *Handbooks of Pragmatics (HoPs) Vol. 3: Pragmatics of Discourse*. Berlin: Mouton de Gruyter.

Sacks, H., Schegloff E.A. and Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-taking for Conversation. *Language,* 50, pp. 696-735.

Schegloff, E. (2007). *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge: Cambridge University Press.

Schmidt, T., Wörner, K. (2009). EXMARaLDA – Creating, Analysing and Sharing Spoken Language Corpora for Pragmatic Research. *Corpus-based Pragmatics*, 19 (4), pp. 565-582.

Annotating spoken language

Ines Rehbein∗, Sören Schalowski∗, Heike Wiese∗†

German Department†, SFB 632 "Information Structure"∗
Potsdam University
irehbein@uni-potsdam.de, soeren.schalowski@uni-potsdam.de, wiese@uni-potsdam.de

Abstract

The design of a linguistically annotated corpus of spoken language is crucial for the future usefulness of the resource, and should thus be carefully tailored towards the needs of the corpus users and the characteristics of the data while, at the same time, paying attention to existing standards to support comparability and interoperability of language resources. In this paper, we outline important issues for the design of a syntactically annotated corpus of spoken language, focussing on standardisation/interoperability and segmentation, and put our proposals up for discussion.

Keywords: Syntactic annotation of non-canonical language, spoken language corpora, Kiezdeutsch

1. Introduction

Syntactically annotated corpora, also known as treebanks, have long found their way into theoretical linguistics, as they provide a valuable resource for the investigation of specific linguistic phenomena and for the verification of linguistic theories. Most treebanks, however, have been developed for written language. This is due to the enormous expenditure of time needed for collecting, digitalising and annotating spoken language data.

This paper discusses best practices for the syntactic annotation of non-canonical language, using Kiezdeutsch (*'hood German'*) as a test case. Kiezdeutsch is a variety of German spoken by adolescents from multiethnic urban areas (Wiese, 2009; Freywald et al., 2011) and can be considered non-canonical in two ways. First, it is a spoken variety of German, thus deviating strongly from canonical written German. It shares all the properties of spoken language which pose a challenge for syntactic annotation, like filled pauses, self-corrections and aborted utterances. In addition, it displays a number of phenomena which make Kiezdeutsch distinct from other varieties of spoken German, such as bare noun phrases, or a word order which, according to standard German, is considered ungrammatical.[1]

In the remainder of the paper we present preliminary work on building a treebank for a non-canonical variety of spoken language. We report on our first experiences with annotating Kiezdeutsch and outline the main problems we encountered during the annotation process. We compare our solutions with the ones chosen in other syntactically annotated corpora of spoken language, and address issues we consider crucial for good treebank design.

The paper is structured as follows. In section 2. we describe the design and architecture of the Kiezdeutsch-Korpus. Section 3. reviews related work on syntactic annotation of spoken language and situates our work in the research context, and Section 4. discusses best practices for annotating language data. We present our design desicions for the KidKo corpus in Section 5. and conclude in Section 6.

2. KidKo – The Kiezdeutsch-Korpus

In our project, we work with Kiezdeutsch (*'hood German'*), a variety of German spoken by adolescents from multiethnic urban areas.

The data was collected in the first phase of project B6 "Grammatical reduction and information structural preferences in a contact variety of German: Kiezdeutsch" as part of the SFB (Collaborative Research Centre) 632 "Information Structure" in Potsdam. It contains spontaneous peergroup dialogues of adolescents from multiethnic Berlin-Kreuzberg (around 48 hours of recordings) and a supplementary corpus with adolescent speakers from monoethnic Berlin-Hellersdorf (around 18 hours of recordings). The current version of the corpus contains the audio signals aligned with transcriptions. We transcribed using an adapted version of the transcription inventory GAT basic (Selting et al., 1998), including information on primary accent and pauses.

We are especially interested in new grammatical developments in Kiezdeutsch (Wiese, 2009; Wiese, 2011) and their interaction with information structure and discourse phenomena. We thus decided on a multi-layer architecture with different levels of annotation, including the audio files and the transcriptions, part-of-speech tags and syntax.

To enable investigations of prosodic characteristics of the data, our transcription scheme has an orthographic basis but tries to closely capture the pronounciation, including pauses and encoding disfluencies and primary accents. In addition, we are adding a level of orthographic normalisation where non-canonical pronunciations and capitalisation are reduced to standard German spelling. This annotation layer enables us to use standard NLP tools for semi-automatic annotation.[2] It also increases the usability of the corpus as it allows one to find all pronounciation variants of a particular expression.

[1] Many of these phenomena also occur in other varieties of spoken German. In Kiezdeutsch, however, they seem to be more frequent than in other varieties. The exact extent of these differences have yet to be determined.

[2] We still have to adapt these tools to our data. However, without the normalisation this would be infeasible.

In addition to part-of-speech tags and phrase structure trees, we also plan to encode topological fields (Drach, 1937; Höhle, 1998) in the corpus. The topological field model is a descriptive model which captures the semi-free German word order. Standard German accepts three possible sentence configurations (verb first, verb second and verb last) by providing fields like the prefield, the middle field and the final field. The fields are positioned relative to the verb, which can fill in the left or the right sentence bracket. The ordering of topological fields is syntactically constrained but serves information-structural functions.

In the next section, we introduce other syntactically annotated corpora of spoken language as points of reference for our work.

3. Related work

Treebanks exist not only for resource-rich languages like English, but also for low-resource languages like Basque (Aduriz et al., 2003), Bulgarian (Osenova and Simov, 2003), or Urdu (Abbas, 2012), to name but a few. Most of these resources, however, are based on standard written text and are often limited to one single genre, namely newspaper text. Noteworthy exceptions are the Christine corpus (Sampson, 2000) and, of much larger size, the Switchboard corpus and the Verbmobil corpus.

3.1. The Christine Corpus

The Christine corpus (Sampson, 2000) was developed in the late 90ies and was one of the first treebanks of spoken language data. Christine includes extracts from the spoken part of the BNC and other sources, totalling more than 80,000 words. The annotation scheme of the Christine corpus provides very fine-grained information based on functional dependencies. The annotation scheme accounts for pauses, speech repairs, and other phenomena of spoken language.

3.2. The Switchboard Corpus

The Switchboard corpus (Godfrey et al., 1992)[3] is a corpus of spontaneous conversations of telephone bandwidth speech which was collected at Texas Instruments. The complete corpus includes about 2,430 conversations averaging 6 minutes in length, which results in over 240 hours of recorded speech with about 3 million words of text, spoken by over 500 speakers of both sexes from every major dialect of American English. The Switchboard corpus is syntactically annotated according to the guidelines of the Penn Treebank (Marcus et al., 1993), with additional guidelines for annotating disfluencies (Meteer and others, 1995), covering phenomena such as non-sentential elements, slash-units[4] and restarts.

[3] Also see the NTX-format Switchboard corpus (Calhoun et al., 2010) which brings together the several annotation layers of the Switchboard corpus and unites them in one single XML format.

[4] Slash-units are units which can (maximally) correspond to a sentential unit, but can also map to incomplete sentences which nevertheless constitute a complete utterance in the discourse.

3.3. The Verbmobil Corpus

Most relevant to our work is the Tübingen Treebank of Spoken German (TüBa-D/S) (Stegmann et al., 2000) which was created in the Verbmobil project (Wahlster, 2000).

The German part of the Verbmobil corpus (which is identical to the TüBa-D/S and henceforth called the Verbmobil corpus) is a syntactically annotated corpus based on spontaneous dialogues between native speakers of German role-playing business partners. The topic of conversation in the data is restricted to scheduling. This is due to the fact that the Verbmobil corpus was created with an eye towards applications for machine translation of spontaneous dialogues. Restricting the domain and limiting the vocabulary used in the corpus was intended to make the task more feasible.

All the dialogues have been transcribed and annotated manually. They include around 38,000 sentences (360,000 token). The annotation provides phrase structure trees enriched with grammatical function labels (dependency relations). The Verbmobil corpus also encodes topological field information as part of the phrase structure trees. We will come back to the treatment of repetitions and speech errors in Verbmobil in section 5.4.

We will compare our work mostly to the Verbmobil corpus which was also created for German, but will also make references to other corpora, when suitable.

4. Design principles for language corpora

This section discusses general principles for good design of (spoken) language corpora which we consider crucial and thus try to implement in KidKo.

4.1. Standardisation/Interoperability

An important issue for building corpora is compatibility with already existing annotation schemes and formats. Using standard annotation schemes to encode linguistic information in the new corpus has some major advantages. First, it allows for comparing the linguistic structure of the new resource to other corpora, and is thus indispensable for studying language variation across different corpora. For NLP purposes, this approach is also favorable, as it allows data users to add new data to existing training sets for system development and domain adaptation.

On the other hand, most existing corpora and annotation schemes for syntactic annotation have been developed with standard written text in mind and fail to capture many of the characteristics of spoken language.[5] A minor problem is the lack of part-of-speech (POS) categories for specific phenomena of spoken language like filled pauses (*uhm, uh*). Such categories can easily be added to the tag set without having a negative impact on the comparability of the corpus, as they do not skew the distribution of existing tags. A

[5] There are, of course, projects concerned with the annotation of speech and spoken language data, developing corpora and annotation standards for spoken language. Most of them, however, are focussed on transcription, prosodic tagging, or part-of-speech annotation. Up to now, only few projects building spoken language corpora provide syntactic annotations beyond the part-of-speech level (but see Deulofeu et al. (2010) for a dependency-based annotation of spoken French).

more severe problem is caused when the annotation scheme of a corpus is based on concepts from written language which cannot be transfered easily to spoken language.

One case in point is the sentence concept, which is the basic unit of analyis in nearly all corpora of written language. When analysing spoken language data, we have to deal with non-sentential utterances, sometimes caused by interruptions, more often by statements consisting of, e.g., only an NP or a PP. These utterance are self-contained utterances in the discourse and should, despite not being sentence-equivalent, be treated as a basic unit. This poses the question of how to segment spoken language, and how to define its basic syntactic unit of analysis. We will come back to this question in section 5.3.

Questions like these have to be considered carefully, as they determine the future usability of the resource with respect to its comparability to (and interoperability with) other language resources.

4.2. Theory-driven vs. data-driven annotation

In this section we want to contrast a *theory-driven* approach to annotation with a *data-driven* approach. We argue that the two approaches will result in completely different resources and will thus have a strong impact on the linguistic results one will get when working with the corpus.

The theory-driven approach takes a particular grammar framework as its starting point and uses it as the basis for analysis. It is thus based on solid grounds (presuming the theory is valid) and can resort to existing solutions for different kinds of phenomena in the data. In general, the theory-driven approach to linguistic annotation does support the consistency of the annotations which, in fact, is highly desirable.

The problem starts when encountering phenomena which so far have not been captured within the theory. The annotators, who have to resolve these issues on the fly using the set of analyses licensed by the theory, then often resort to an analyis valid within the theoretical framework which does not really fit the data. Even more important, they are tempted to ignore linguistic phenomena simply because there is no valid analysis at hand. This is a major problem when working with non-canonical data where one will encounter many new phenomena not licensed within the theoretical framework, which most probably was tailored to the phenomena of standard, canonical language.

To illustrate our point, let us consider the case of multiple frontings in German. The Verbmobil annotation guidelines state that only one constituent is allowed in the prefield (Stegmann et al. (2000), p.24). As a result, elements like *dann* in Example (1) are not placed within the prefield in the German Verbmobil Corpus but instead are annotated as isolated phrases and attached to the virtual root node (Figure 1).

(1) [...] , dann ich sehe jetzt Don-Giovanni von Mozart .
[...] , then I watch now Don-Giovanni by Mozart .
"... , then I'll go see Don Giovanni by Mozart."

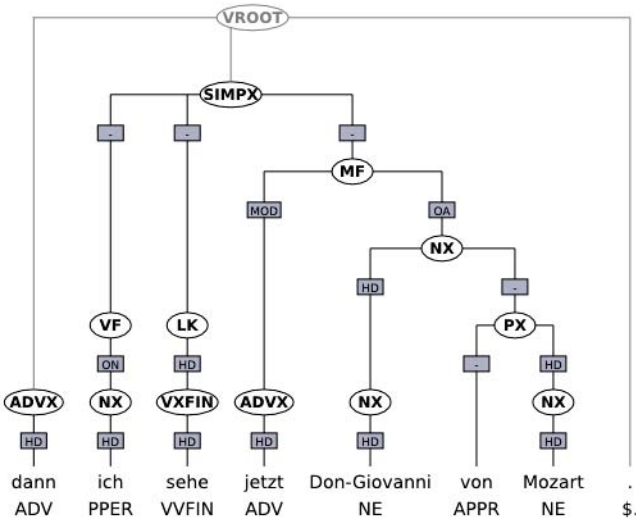


Figure 1: Multiple prefield in the Verbmobil corpus (VF: prefield, MF: middle field, LK: left sentence bracket)

Although their existence is not acknowledged in the Verbmobil guidelines, those constructions are quite frequent in spoken language. In the Verbmobil Corpus we found 157 cases like the one in (1) where an adverbial phrase, attached to the root node, is followed by another preverbal constituent. We argue that the theory-driven approach runs the risk of ignoring phenomena which are not licensed by the theory, as done for the multiple frontings, as these are not allowed in the grammar and thus not annotated as such. Instead, we propose a data-driven approach where all constructions which occur in the data are taken at face value and described exactly where they occur at surface level. We thus define the prefield as the constituent(s) in the left periphery (meaning everything occuring to the left of the left sentence bracket), without restricting the number and type of constituents allowed here. This, of course, leads us back to the question of segmentation, as this will determine what exactly we will find in the left periphery.

5. Syntactic annotation in KidKo

We will now report on our preliminary experiences with annotating Kiezdeutsch. To familiarise ourselves with the problems arising for annotating non-canonical language data, we selected a small sample with 1,265 token from the corpus. The sample is a recording of five teenage girls having an informal conversation about school, friends, and similar topics. We automatically provided syntactic analyses for the data using the Berkeley parser (Petrov et al., 2006) trained on the TIGER treebank (Brants et al., 2002) and manually corrected the parser output trees.

5.1. POS Annotation

In Section 4. we argued for using standard annotation schemes to support interoperability and comparability and extended them to suit the needs of the data at hand. Following this guideline, we use an extended version of the Stuttgart Tübingen Tag Set (STTS) (Schiller et al., 1995) for part-of-speech annotation. The STTS is the standard

POS	description	example	transliteration
PTK	*particle, unspecific*	Ja/PTK kommst Du denn auch ?	yes come you then too ?
PTKFILL	*filler*	Ich äh/PTKFILL ich komme auch .	I er I come too .
PTKREZ	*backchannel signal*	A: Ich komme auch . B: Hm-hm/PTKREZ .	A: I come too . B: Uh-huh .
PTKONO	*onomatopoeia*	Das Lied ging so lalala/PTKONO .	The song went so lalala .
PTKQU	*question particle*	Du kommst auch . Ne/PTKQU ?	You come too . No ?
XYB	*unfinished word*	Ich ko/XYB #	I ko #
UI	*uninterpretable*	(unverständlich)/UI #	(uninterpretable) #
$#	*unfinished utterance*	Ich ko #/$#	I ko #

Table 1: Additional POS tags in KidKo (the # is used to mark incomplete utterances or utterances not interpretable due to low quality of the audio)

part-of-speech tagset for German and was also used (with minor variations) to annotate POS tags in the TIGER treebank and in the Verbmobil corpus.

The original STTS provides a set of 54 different tags. Our extended version comprises additional tags for filled pauses, question particles, backchannel signals, onomatopoeia, unspecific particles, breaks, uninterpretable material, and a new punctuation tag to mark unfinished sentences (Table 1). This allows us to do a more fine-grained analysis than the one in the Verbmobil corpus, which uses the tags from the original STTS only.[6] Our part-of-speech annotation of interjections, discourse markers and fillers is also more fine-grained than the one in the Switchboard corpus which combines them all into one tag (INTJ).

In addition, we propose including an extra layer of annotation which maps our language-specific POS tagset to a coarse-grained, universal POS tagset (Petrov et al., 2011) which already provides a mapping for 25 different treebank tagsets for 22 languages. This will enable a comparison (admittedly on a very coarse-grained level) of the numbers for part-of-speech annotation across different corpora.

5.2. Syntactic Annotation

Our syntactic annotation is adapted to the one in the TIGER treebank (Brants et al., 2002), which is characterised by its flat tree structure. Like the Verbmobil corpus, TIGER encodes syntactic categories as well as grammatical functions in the tree. TIGER uses a set of 25 syntactic category labels and distinguishes 44 different grammatical functions.

Unlike in Verbmobil, non-local dependencies are expressed through crossing branches. TIGER neither annotates unary nodes, nor does it provide annotations for topological fields.[7]

In the remainder of this section we will go into the problem of segmentation and will describe the syntactic representation of disfluencies in the KidKo corpus.

5.3. Unit of analysis

Many proposals have been made in theoretical linguistics for defining the most adequate unit of analysis for spoken language, in analogy to the sentence concept for standard written text (for an overview see (Crookes, 1990; Foster et al., 2000)). The ideas of what to use as a basis for linguistic analysis differ considerably, depending on the theoretical background of the respective researchers. Although the issue has been discussed for a long time, there is no general agreement as to what should be used, and even different definitions of the same individual unit show substantial variation. In a thorough review of literature on spoken language, Foster et al. (2000) looked at more than 80 studies published in four leading journals on applied linguistics and second language acquisition, and found that identical units were either not defined in the same way or not defined at all, or that the instructions given in the article were not sufficient to handle real-world, messy spoken data. This trend is alarming, as the choice of unit will have a significant impact on the results of the analysis, and also on the comparability of the results to other studies.

Our main ideas for defining a proper unit of analysis are as follows. First, we want to maintain interoperability and comparability with corpora of written language on sentence-like utterances in the data. Second, we want to capture different types of non-sentential units which are frequent in spoken language and which are not, in general, fragments, even when not following the rules of a standard grammar for written language. Third, the guidelines on what should be annotated as a unit of utterance should be based on a theoretical basis suitable for describing spoken language, and should enable the annotators to apply them in a consistent way. We thus base our unit of analysis on structural properties of the utterance and, if not sufficient, also include functional aspects of the utterance. The latter results in what we call the principle of the *smallest possible unit*, meaning that when in doubt whether to merge lexical material into one or more units we consider the speech act type and discourse function of the segments.[8] Example (2) illustrates this by showing an utterance including an imperative (*Speak German!*) and a check question (*Okay?*). It would be perfectly possible to include both in the same unit, separated by a comma. However, as both reflect different speech acts, we choose to annotate them as separate

[6] Our annotation is not as fine-grained as the one in the Christine corpus which, for example, distinguishes pauses filled with nasally produced noises (erm) from vocally filled pauses (er).

[7] As stated earlier, we plan to annotate topological fields in the corpus. This information will be added on a separate annotation layer instead of including it in the phrase structure tree.

[8] Our classification of speech act types and discourse functional units is inspired by work on shallow discourse-function annotation (Jurafsky et al., 1997) and on non-sentential units in dialogue (Fernandez and Ginzburg, 2002; Fernández et al., 2007).

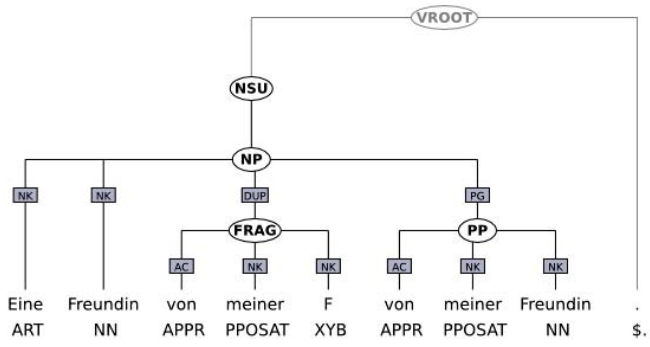


Figure 2: Syntactic representation of a non-sentential unit (NSU) in KidKo "A friend of my F of my friend."

units of analysis.

(2) Rede auf Deutsch ! Okay ?
Speak on German ! Okay ?
"Speak German! Okay?"

This treatment is opposed to the one in the Verbmobil corpus, called the *longest match principle*. The longest match principle demands that as many nodes as possible are combined into a single tree as long as the resulting tree structure is syntactically as well as semantically well-formed.
Based on our considerations outlined above, our basic unit of analysis for syntactic annotation can be defined as an utterance either corresponding to the sentence (S) in standard written language[9], or to a non-sentential unit (NSU). Our classification of non-sentential units distinguishes different question types (e.g. backchannel questions, clarification questions, check questions, WH questions), non-sentential imperatives, different answer types (short answers, plain affirmative answers), different types of modification (adjectival modifiers, propositional modifiers, bare modifiers), non-sentential exclamatives, onomatopoeia, vocatives, and more.
Figure 2 shows an example from the corpus. The utterance is a non-sentential elaboration of an answer given in response to the question "Who is that girl?", adding more information to the original answer. Figure 2 also illustrates our treatment of unfinished words (see 5.1., Table 1) and the representation of repetitions (DUP). We will come back to this topic in section 5.4.

5.4. Syntactic representation of disfluencies

A severe problem for the syntactic analysis of spoken language phenomena concerns the representation of disfluencies in the syntax tree, such as repetitions, hesitations, self-repair or filled pauses (fillers).

Self-repair Example (3) illustrates a self-repair marked by a filled pause. Our terminology follows the one of Shriberg (1994) which is based on (and modifies) Levelt (1983), and which was also used in the annotation of the Switchboard corpus.

[9]We operationalise the annotation of S as the set of all utterance including a finite verb.

(3) Was machs äh was mache ich falsch ?
$\underbrace{What\ do.2.Sg}_{reparandum}$ $\underbrace{uh}_{interregnum}$ $\underbrace{what\ do.1.Sg}_{repair}$ I wrongly ?
"What's wrong with what I do?"

First, the material to be corrected (*the reparandum: Was machs*) is produced, followed by the filled pause (*the interregnum: äh*). The interregnum refers to the time span from the end of the reparandum to the onset of the repair. In Example (3), this temporal region is filled, but this is not necessarily the case. Sometimes there are unfilled pauses which serve the same reason, namely replanning the utterance abandoned by the speaker. Then we have a new start, the *repair*, displaying the intended utterance (*was mache*). Disfluencies in spoken language often result in fragments which do not correspond to proper constituencies and where the grammatical status of the fragment is by no means clear. Off-the-shelf NLP tools, being trained on standard written text, are not able to handle those structures. The removal of the reparandum makes it easier to apply off-the-shelf NLP tools to the data. In addition, *reparandum* and *repair* often give rise to duplicate arguments. E.g. in Example (3), the self-repair results in an utterance with two direct objects and two predicates. Figures 3-6 show four (out of many) possibilities for representing the syntactic structure in a constituency tree.

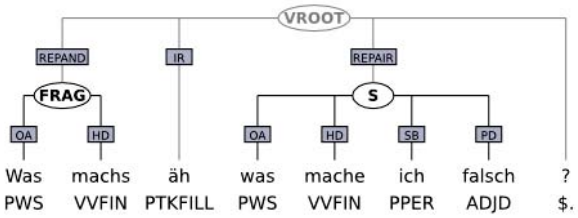


Figure 3: Different syntactic analyses for self-repairs I

In Figure 3 the reparandum is annotated as a fragment (FRAG), while the functional label identifies it as the reparandum (REPAND). The representation in 3 reflects the common view to regard filled pauses as a performance problem and thus does not integrate them in the sentence but attaches them to the virtual root node (VROOT). A more radical approach consists in eliminating all fillers from the corpus, as done in the Verbmobil corpus. This strategy does not allow one to directly encode the function of the *interregnum* in Example (3), namely to mark the interruption point and to indicate the self-repair. The second strategy is even worse for linguistic purposes, as it makes it infeasible to use the corpus for investigations of disfluencies and strategies in speaking.[10,11]
Figure 4 annotates both reparandum and repair inside of FRAG nodes which are both attached to the same sentence

[10]See, e.g., the studies by (Clark and Wasow, 1998; Clark and Fox Tree, 2002).

[11]The approach taken in the Verbmobil corpus is, of course, motivated by its main purpose to serve as a resource for machine translation of spontaneous dialogues, not as a resource for linguistic research.

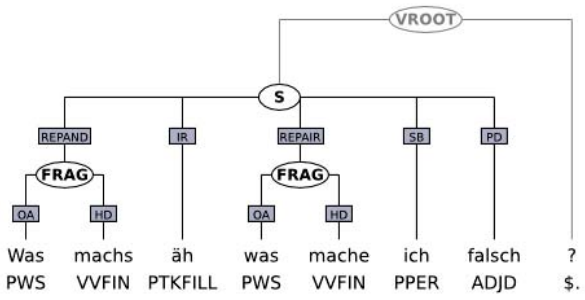


Figure 4: Different syntactic analyses for self-repairs II

node, as is the interregnum. This reflects the idea that all should belong to the same unit of analysis. This particular representation, however, makes it hard to recover the well-formed sentence from the utterance and also gives misleading search results for linguistic queries looking for, e.g., the number of direct objects governed by a sentence node.

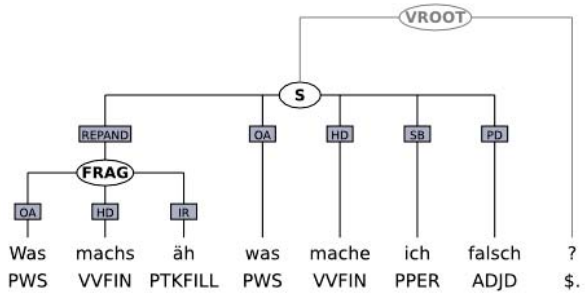


Figure 5: Different syntactic analyses for self-repairs III

Figure 5 solves this problem by attaching the repair directly to the sentence node (S). The interregnum is annotated as part of the reparandum, which (same as in 3) is annotated as a fragment (FRAG). Filled pauses, however, can also occur without a self-repair. Thus we decided to treat all fillers the same and not to include them in the FRAG node containing the reparandum. This results in our prefered representation, shown in Figure 6. Our solution is similar to the treatment of self-repair in the Switchboard corpus (Figure 7).

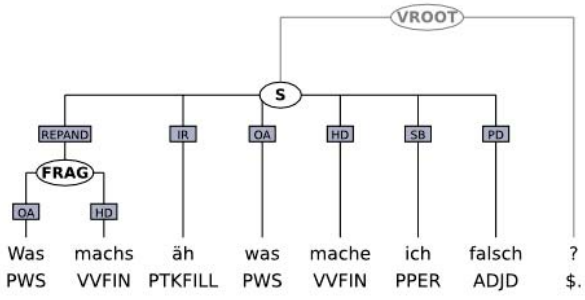


Figure 6: Different syntactic analyses for self-repairs IV

In the Switchboard corpus, square brackets are used to mark the beginning (RM) and end (RS) of a self-repair sequence (Figure 7), and the interruption point (IP), which occurs directly before the interregnum, is assigning the plus (+) sym-

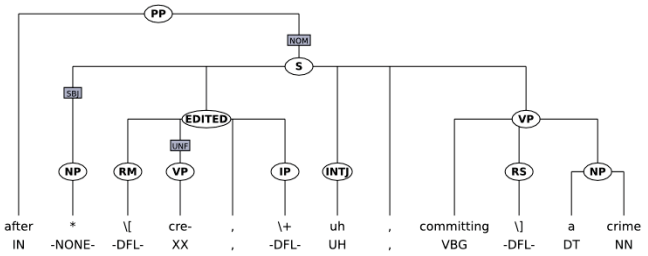


Figure 7: Syntactic representation of a self-repair in the Switchboard corpus (RM: start of disfluency, RS: end of disfluency; IP: interruption point

bol. The reparandum is attached to a node with the label EDITED (comparable to our FRAG node). Removing this EDITED node, which includes all material from the start of the reparandum up to the interruption point, results in recovering a well-formed version of the utterance.

Our final representation (Figure 6) allows us to consider *reparandum* and *repair* as a functional unit and to conduct corpus searches for filled pauses inside of specific constituents, e.g. comparing the occurences of filled pauses inside NPs to ones on the sentence level. The annotation in Switchboard also explicitely encodes the end of the repair, which our annotation does not. We agree that it would be desirable to have this information but, considering the additional effort for manual annotation, refrain from doing so. Unlike Switchboard, we also annotate unfilled pauses in the corpus, differentiating beween short pauses ($< 1sec$), medium pauses ($1 - 3sec$) and long pauses ($> 3sec$). Our annotation enables the user to identify and discard the fragments and recover only the well-formed part of the utterance, if need be.

Repetitions of (sequences of) words without a self-repair also occur frequently in spoken language data. The repeated material can occur at any position in the utterance and, like the self-repair, also inserts extra material which causes problems for syntactic analysis.

The treatment of repetitions in the Switchboard corpus is illustrated in Figure 8 and looks similar to the treatment of self-corrections. The repeated material is attached to the EDITED node, as is the interruption point. The square brackets mark begin and end of the repetition, and through the deletion of the EDITED node the utterance can be transformed into a well-formed sentence.

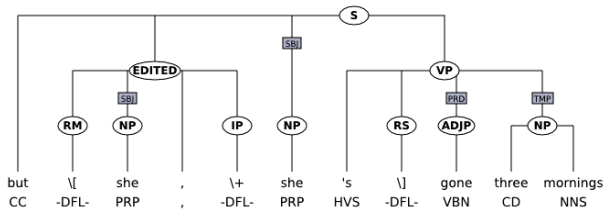


Figure 8: Representation of repetitions in the Switchboard corpus

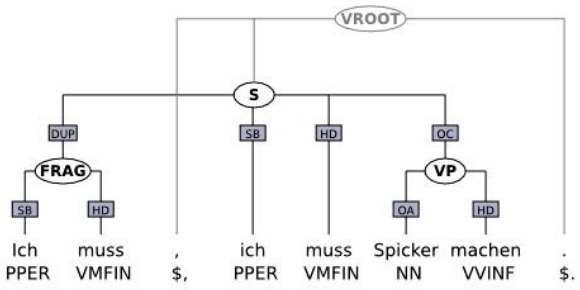


Figure 9: Representation of repetitions in KidKo: "I have to make a cheatsheet."

The syntactic representation of repetitions in the Kiezdeutsch corpus (Figure 9) is again a slimmed-down version of the one in the Switchboard corpus. Again, we neither mark the start or end of the repetition nor the interruption point, but follow the Switchboard annotations in attaching the extra material to a special node (FRAG) to mark it as non-canonical and to allow the user to recover the well-formed utterance. The function label in KidKo encodes the information that this is duplicate material (DUP).

In contrast to our analysis, repetitions in the Verbmobil corpus are not attached to the same constituent as the duplicate material, but are rather treated as an independend unit. Figure 10 shows an example. Here, the repetition at the start of the utterance (*Ich habe*) is attached to an extra sentence node (SIMPX). The sentence is neither recognisable as a false start (which is also a plausible interpretation), nor is it marked as a repetition of the material in the following utterance. There is no observable relation between the two sequences, disregarding that they are, in fact, closely related.

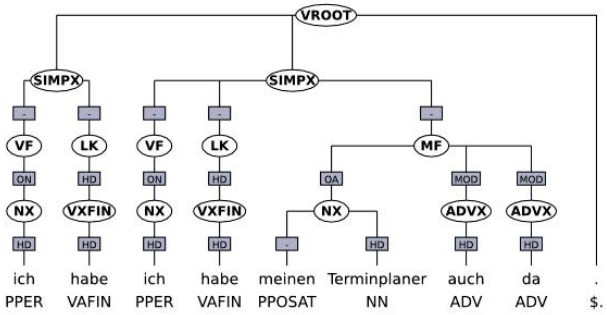


Figure 10: Representation of repetitions in the Verbmobil corpus: "I have my scheduler here, too."

In conclusion, we strongly argue for including disfluencies (which are often, but not exclusively caused by self-repairs) in the corpus, even if this will result in more manual effort for correcting automatically predicted POS tags during preprocessing. The negative impact of repetitions on the accuracy of automatic POS tagging was the main reason for the BNC (Burnard, 2007) to discard them from the spoken part of the corpus. We address the problem by inserting an additional annotation layer where we manually mark all utterances with self-corrections during transcription, which will allow us to identify and efficiently handle them during the POS tagging step.

6. Conclusion

In the paper we discussed major issues for the design of a syntactically annotated corpus of spoken language. We highlighted the importance of standardisation and interoperability, which we accomodate by using and expanding an existing annotation scheme developed for standard written language.

To accomodate conflicting needs and research interests, we propose an encoding of disfluencies and other phenomena of spoken language which allows users to either include or exclude these pieces of information, depending on their needs and research question. We argue for a theory-neutral, data-driven approach to linguistic annotation which describes spoken language phenomena at the surface level and which will allow researchers to build and test linguistic theories on real-world data.

7. Acknowledgements

This work was supported by a grant from the German Research Association (DFG) awarded to SFB 632 "Information Structure" of Potsdam University and Humboldt-University Berlin, Project B6: "The Kiezdeutsch Corpus". We acknowledge the work of our transcribers and annotators, Anne Junghans, Jana Kiolbassa, Marlen Leisner, Charlotte Pauli, Nadja Reinhold and Emiel Visser, We would also like to thank the anonymous reviewers for helpful comments.

8. References

Qaiser Abbas. 2012. Building a hierarchical annotated corpus of Urdu: The URDU.KON-TB Treebank. In *13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012)*, pages 66–79.

Itzair Aduriz, María Jesús Aranzabe, José María Arriola, Aitziber Atutxa, Arantza Díaz De Ilarraza, Aitzpea Garmendia, and Maite Oronoz. 2003. Construction of a Basque dependency treebank. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT 2003)*, pages 201–204, Växjö, Sweden.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 24–42.

Lou Burnard. 2007. Reference guide for the British National Corpus XML edition. Technical report, http://www.natcorp.ox.ac.uk/XMLedition/URG/.

Sasha Calhoun, Jean Carletta Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speech. *Cognition*, 84:73–111.

Herbert H. Clark and Thomas Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology*, 37:201–242.

Graham Crookes. 1990. The utterance and other basic units for second language discourse analysis. *Applied Linguistics*, 11:183–199.

José Deulofeu, Lucie Duffort, Kim Gerdes, Sylvain Kahane, and Paola Pietrandrea. 2010. Depends on what the French say. Spoken corpus annotation with and beyond syntactic functions. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 274–281. Association for Computational Linguistics.

Erich Drach. 1937. Grundgedanken der Deutschen Satzlehre.

Raquel Fernandez and Jonathan Ginzburg. 2002. Non-sentential utterances in dialogue: A corpus-based study. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Pauline Foster, Alan Tonkyn, and Gillian Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3):354–375.

Ulrike Freywald, Katharina Mayr, Tiner Özcelik, and Heike Wiese. 2011. Kiezdeutsch as a multiethnolect. In Friederike Kern and Margret Selting, editors, *Ethnic Styles of Speaking in European Metropolitan Areas*, pages 45–73. Amsterdam, Philadelphia: Benjamins.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 517–520, San Francisco, California, USA.

Tilman Höhle. 1998. Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340, Göttingen, Germany.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL. Shallow-discourse-function annotation. Coders manual, draft 13. Technical report, University of Colorado, Boulder. Institute of Cognitive Science.

Willem J.M. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14:41–104.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the penn treebank. *Compututational Linguistics*, 19(2):313–330.

Marie Meteer et al. 1995. Dysfluency annotation stylebook for the Switchboard corpus. Linguistic Data Consortium. `ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps`. Revised June 1995 by Ann Taylor.

Petya Osenova and Kiril Simov. 2003. The Bulgarian HPSG treebank: Specialization of the annotation scheme. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT 2003)*, pages 81–93, Växjö, Sweden.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *The 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-COLING)*, pages 433–440.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. In *ArXiv*, April.

Geoffrey Sampson. 2000. Christine corpus, stage i: Documentation. Technical report, Sussex: University of Sussex.

Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart, Universität Tübingen.

Margret Selting, Peter Auer, Birgit Barden, Jörg Bergmann, Elizabeth Couper-Kuhlen, Susanne Günthner, Uta Quasthoff, Christoph Meier, Peter Schlobinski, and Susanne Uhmannet. 1998. Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte*, 173:91–122.

Elizabeth E. Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California at Berkeley.

Rosmary Stegmann, Heike Telljohann, and Erhard W. Hinrichs. 2000. Stylebook for the german treebank in verbmobil. Technical Report 239, Seminar für Sprachwissenschaft, Universität Tübingen.

Wolfgang Wahlster, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin.

Heike Wiese. 2009. Grammatical innovation in multiethnic urban europe: New linguistic practices among adolescents. *Lingua*, 119:782–806.

Heike Wiese. 2011. The role of information structure in linguistic variation: Evidence from a German multiethnolect. In Frans Gregersen, Jeffrey Parrott, and Pia Quist, editors, *Language Variation – European Perspectives III*, pages 83–95. Amsterdam: Benjamins.

Turns in interdisciplinary scientific research meetings: Using 'R' for a simple and flexible tagging system

Seongsook Choi, Keith Richards

Centre for Applied Linguistics
University of warwick
S.Choi@warwick.ac.uk, K.Richards@warwick.ac.uk

Abstract

This paper presents our initial step towards identifying and mapping functions (of utterances/turns) and actions (a series of connected actions managed over the course of a sequence of turns) inherent in authentic spoken language data using a simple and flexible tagging system in R. Our ultimate goal is to capture the patterns of dynamic practices through which interactants produce and understand talk-in-interaction both qualitatively and quantitatively. The procedure involves annotating the transcripts with tags that blends elements of CA (conversation analysis) and DA (discourse analysis), which we can then analyse quantitatively. The paper addresses the challenge of developing and annotating a CA and DA integrated tagging system and demonstrates graphical representation of a quantitative analysis that can be derived from it.

Keywords: conversation analysis, discourse analysis, R, turn-taking

1. Introduction

This paper presents our initial attempt at mapping functions (of utterances/turns (Crookes, 1990)) and actions (a series of connected actions managed over the course of a sequence of turns) inherent in authentic spoken language data using a simple and flexible tagging system in R (R Development Core Team, 2011). Our ultimate goal is to capture the patterns of dynamic practices through which interactants produce and understand talk-in-interaction both qualitatively and quantitatively. We carry this out by annotating the transcripts with tags that blends elements of CA (conversation analysis) and DA (discourse analysis), which we can then analyse quantitatively.

While DA allows the use of priori categories as analytical resources, this is explicitly rejected in CA methodology, which has led to the assumption that the two are irreconcilable, though CA findings have been used as the basis for subsequent quantitative analysis (e.g. Mangione-Smith, Stivers, Elliott, McDonald and Heritage, 2003). We seek a closer integration of the two, using a CA analysis of sequences of talk to reveal aspects of participant design which would remain hidden in a DA approach, then identifying discourse patterns within these which can be mapped, using DA, across large data sets in order to reveal ways in which relevant features of talk play out in different interactional contexts. This paper focuses on issues of representation in the quantitative dimension of our work, focusing specifically on turn-taking.

2. The challenge of CA/DA synthesis

The overarching aim of this project is to combine the analytic penetration of CA with the potential applications of DA to large databases of spoken interaction. In this section we identify the challenges which this presents and the ways in which we intend to address these.

Although DA embraces a much wider analytical spectrum than CA, their very different conceptual foundations make procedural synthesis inherently problematic. While ethnography and CA have been widely accepted as complementary approaches (e.g. Miller and Fox, 2004), as have ethnography and DA (e.g. Sarangi and Roberts, 2005) and, more broadly, ethnography and linguistics (Rampton, Tusting, Maybin, Barwell, Creese and Lytra, 2004), the differences between CA and DA, at least in the form that we draw on this project, were exposed in debates that began over a quarter of a century ago (e.g. Levinson, 1983; Schegloff, 2005; Van Rees, 1992). Since then each has followed its own lines of development, CA developing in applied areas and connecting with Membership Categorisation Analysis, DA engaging with contextual aspects to the point where researchers would now claim that 'at its heart DA remains an ethnographically grounded study of language in action' (Sarangi and Roberts, 2005, p. 639). To our knowledge, no attempt has been made to bring them together as part of the same analytic enterprise and our attempt to do so draws on forms of DA which are susceptible to coding and quantitative analysis.

The reason for this is clear from the following comment on quantitative studies from a CA perspective (Heritage, 1995, p. 406):

> Quantitative studies have not, so far, matched the kinds of compelling evidence for the features and uses of conversation practices that have emerged from 'case by case' analysis of singular exhibits of interactional conduct. It does not, at the present time, appear likely that they will do so in the future. For quantitative studies inexorably draw the analyst into an 'external' view of the data of interaction, draining away the conduct-evidenced local intelligibility of particular situated actions which is the ultimate source of security that the object under investigation is not a theoretical or statistical artefact.

The 'externality' of DA arises from its willingness to identify specific features of talk (classically, particular speech acts) and apply these as what CA would see as a priori categories. The advantage of such an approach is that it allows the sort of coding that makes extensive data sets accessible to the analyst; the disadvantage is that it does so at the expense of failing to capture aspects of the construction of the talk, with the result that it may all too easily miss what is actually getting done through the talk.

It is this focus on action that characterises CA and explains its insistence on the importance of the sequential unfolding of interaction. Schegloff (1991, p. 46) captures this essential relationship well:

> . . . the target of its [CA's] inquiries stands where talk amounts to action, where action projects consequences in a structure and texture of interaction which the talk is itself progressively embodying and realizing, and where the particulars of the talk inform what actions are being done and what sort of social scene in being constituted.

The advantages of this approach is that it enables the analyst to understand what is being achieved through the talk in a way that is not open to the discourse analyst using coding to analyse large data sets; the disadvantage is the demands it places on the analyst in terms of time and resources. The transcription system itself demands close attention to minutiae of delivery and the process of collecting data is often a slow and painstaking process, which might be described in terms of 'tracking the biography of the phenomenon's emergence' (Jefferson, 1983, p. 4) or 'having accumulated a batch of fragments' (*ibid* p. 16) .

Our basis for bringing these apparently incommensurate approaches together lies in finding a way of applying CA in order to identify an 'action', then using DA in the form of a coding system based on pragmatic features to identify patterns across stretches of talk that identify this action, then applying CA to the instances thus identified in order to check the accuracy of the specified pattern. While no CA practitioner would accept this as a legitimate analysis in itself, since it will always be possible that other things are being accomplished through the talk, it does allow specific actions to be identified and thereby makes it possible to develop increasingly rich pictures of how particular actions are distributed through the talk. What follows focuses on the tools that can be used to maximise the benefits derivable from an action -based analysis.

3. Data

The data used for this project are drawn from audio recorded interdisciplinary scientific research project meetings ranging from large collaborative funded projects with at least 6 participants in each meeting to interdisciplinary PhD supervision meetings consisting of two supervisors and a student. The disciplines represented in these meetings consist of mathematics, statistics, biology and bioinformatics. The data have been collected since March 2011, producing about 120 hours of audio recordings to date as part of a collection that will continue to grow as we follow a number of research projects to completion. What we are presenting here is based on only small part of the data that have been transcribed (amounting to 20 hours to date).

It is also necessary to emphasise that this represents an early stage in the project. Development is currently focused on an action in which the speaker introduces a question and then follows this with a series of turns leading to a suggestion. What makes this a particularly attractive starting point for our analysis is that the sequence is marked by turns with *so* in the turn-initial position ('*so*-clusters'), which are easily identifiable in the data. Extract 1 provides an example this action (all names are pseudonyms):

Extract 1

01	ALF	was it a strict criterion for it
02	ROY	or no not very strict
03	ALF	so you don't think it's normal
04		it's (xxx) fourteen that's
05		really the
06	ROY	I probably could find more yes
07		I mean I didn't use any strict
08		criteria just I applied some
09		(xxx) two or three (xxx) also
10		my eye each if I believe.
11	ALF	yeah so it should be about the
12		number.
13	GARY	yeah
14	ROY	depends you know if I make it
15		less strict (xxx) fifty (xxx).
16	ALF	so if say eighty percent of them
17		are thought to be affected by
18		the wash if we did the whole
19		mock wash micro array data it
20		would allow us to identify
21		twenty genes that are affected
22		by pulse so we don't know
23		whether that's relevant it's
23		worth its worth finding out (4.0)

In terms of interdisciplinary talk, once the analysis is complete it will be interesting to see how this action is distributed in the data. If, for example, quantitative analysis reveals that such exchanges are inter-disciplinary (as opposed to intra-disciplinary), this would provide prima facie evidence of genuine interdisciplinary exchanges. It would also enable us to collect examples of this across different data sets in order to understand more about how such sequences work towards the building of shared understanding and action.

At this stage we are working with basic transcriptions of the sort illustrated above and limiting more delicate transcription to examples of the relevant action, though the differences between these can be considerable, as a comparison of Extract 2 with its 'equivalent' in lines 05 and 06 in Extract 1 demonstrates:

Extract 2

```
01  ALF   really:
02        (0.4)
03  ALF   the (norm).
04        (0.5)
05  ALF   ·hh
06        (3.0)
07  EMMA  °mm°
08  ALF   hh:°::°=
09  ROY   =I probably could fi::nd
10        more yes::
```

While the application of CA to sequences identified through tagging allows us to cross-check the accuracy of the latter and thereby overcomes some of the problems associated with coding, such as 'specifying and categorizing the meaning of utterances that are far more ephemeral, malleable and negotiable than coding schemes allow' (Fairhurst and Cooren, 2004, p. 134), a number of other challenges remain in terms of how the data is transcribed and coded, and how a threshold level of relevant features is to be determined.

4. Tags

4.1. Description of tags

In order to carry out a quantitative analysis, we annotate the transcript by using a simple tagging system in which turns are 'tagged' manually in order to capture any information of interest. We are in the process of developing a blended CA-DA tagging system that captures the pragmatic functions of each turn and sequences of turns, which, following Van Dijk (1981), we label episodes. At the level of each turn, each tag displays primary and secondary pragmatic functions of each turn. For example a tag 'agr-qual' describes agreement followed by a qualification (e.g., A: 'we should do X.' B: 'Yes, but we need to limit the number.'):

Extract 3

```
PAUL:  yeah maybe what you really need
       is error bars.  /agr-sugg, ep3/
MARY:  erm yeah haven't worked out how
       to do that percentage of input
       but if I repeat them as well
       then I can start easily putting
       error bars on.  /agr-qual, ep3/
```

Mary's response to Paul is an agreement but she says that she hasn't yet worked out how to do it, thus qualifying her agreement.

At the level of episode, each turn within one episode is tagged (as 'ep3' in the above example). This is to capture patterns of turn sequences that exhibit topic changes over the course of conversation. Each episode is defined in terms of one topic or issue that may be initiated by a question that ends with an answer or a solution that is agreed by the interactants.

These tags are comma-separated and can be of any number; the property they define is identified by their position. Thus, if one is interested in topic-switching and pragmatic function of each turn, two tags are used, as in Extract 3. This approach does not impose any limitation to the user and can suit a broad range of analyses.

4.2. Tagging process

The functions and actions are annotated manually by first identifying episodes of topics. Then each turn of the speaker is tagged by its pragmatic function. So the tag for each turn is pragmatically linked with the preceding turn, representing a pragmatic token of talk-in-interaction. Any attempt to code data must face problems arising from the context and situation sensitivity of each turn, making possible alternative interpretations (taking an example from Extract 3, as Drummond and Hopper (1993) have shown, *yeah* cannot simply be treated as indicating agreement), so the process of designing an adequate coding system is a long and complex one, balancing analytical utility with interpretive adequacy. For this reason it has to be carried out manually and consequently it is a time consuming process, as capturing pragmatic function of each turn is not the same as annotating parts of speech or syntactic components of utterances. Currently, we are filtering through the initial set of tags that we have created and are developing systematic tagsets that will enable us to capture interactional dynamics. When we have arrived at a workable set of descriptors for relevant categories, these will be tested using different raters and inter-rater reliability measures. However, the success of the system will ultimately be determined by whether specific patterns can be used to predict the presence of particular actions confirmed by CA procedures.

5. Computational approach

We have developed a simple R script that automatically parses a tagged transcript file to a data frame (the equivalent of a spreadsheet in Excel). R is free and open-source software which offers a gamut of statistical tests and procedures. The resulting data frame has a number of columns that include the speech part itself, the speaker, the line number and any tags of interest to the user (see Figure 1). For our initial analysis, we focused on the turn-initial *so* and its function within the episode.

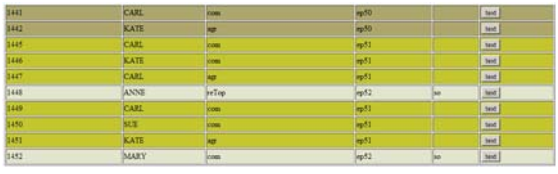


Figure 1: Timeline of *so*

Using this information, we can then apply a number of statistical techniques available in R to analyse the data. In order to more easily browse the data and identify patterns, we generated html files to represent the transcript. The text itself is not shown (but can be displayed with one click), only the value of the tag and whether so has been used. As shown in figure 1, each turn is colour-coded by the episode it belongs to, making it easy to spot topic-switching, for example if a topic is abandoned for a short period of time, only to be returned to at a later stage.

6. Data analysis

Each turn is annotated by one of the following types of tags: agr(eement), expl(anation), sugg(estion), sum(mary), q(uestion), ch(ecking), disagr(eement), recap(itulation), ups(hot), com(ment). These are not the exhaustive list of tags that we are developing but here we present our analysis using these tags only.

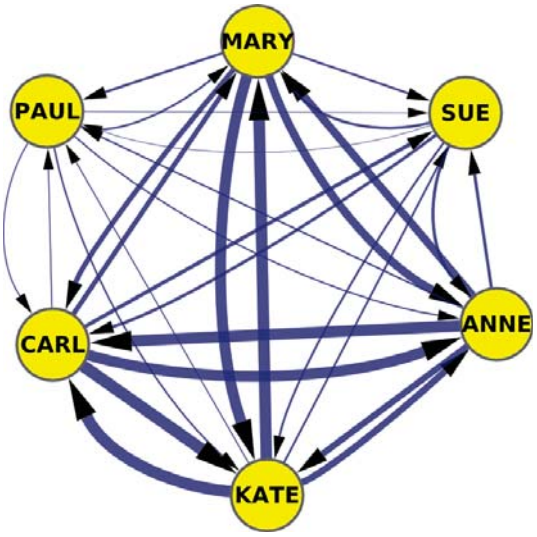


Figure 2: Speaker interaction

This figure, generated with Cytoscape (Smoot, Ono, Ruscheinski, Wang and Ideker, 2011), presents a graphic summary of the interactions between speakers. The thickness of the edges conveys the number of occurrences of the pair (speaker 1, speaker 2). This representation can be understood as a probabilistic model of the meeting: starting from one speaker, the next speaker is chosen according to the observed frequency of turn-taking from that speaker. The figure reveals immediately that the meeting is dominated by four persons, Anne, Carl, Kate and Mary, but it is also interesting to note that exchanges between all the participants are symmetrical. Had fewer of Kate's turns, for example, followed those of Mary, the arrowed edge from Mary to Kate would have been thinner than that from Kate to Mary. Representations of this sort allow the analyst to identify features of potential interest in the talk, some clear (e.g. whether, for example, participants from a particular discipline are dominating the talk or the meeting shows evidence of cross-disciplinary exchanges), others suggestive (e.g. in the case of unequal edges, asymmetries in the relevant relationship).

It is also possible to complement this speaker network by developing a representation that captures the nature of the responses made to the prior turn. Figure 3 below provides information about the nature of the speaker's utterance, given by its tag. For each edge 'speaker 1 to speaker 2' in the network, a corresponding bar in the figure shows the tags that were used by speaker 2 in that turn-taking occurrence. For example, we know from the network (figure 2) that Carl's turns frequently follow those of Anne. The wheel below then allows us to see at a glance that in those cases roughly 10% of the responses were agreements and a further 30% were comments, while disagreement was very rare.

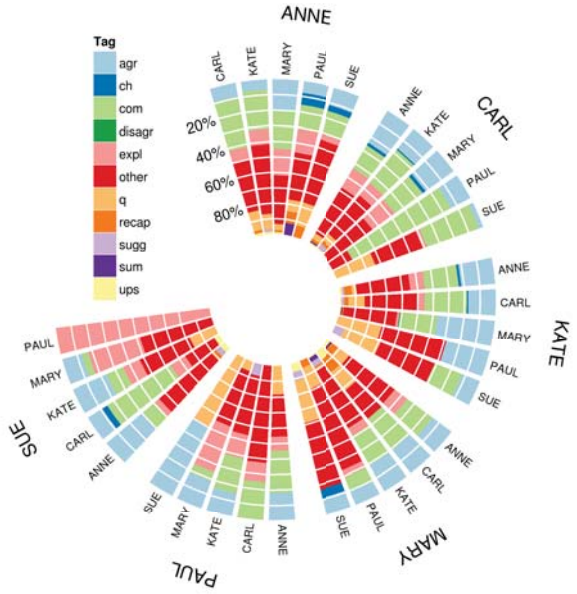


Figure 3: Speaker-Tag distribution

Having a consistent and instantly accessible way of representing the interactions between participants will enable us to compare different meetings and see, for example, if any pattern emerges from a status, gender and discipline angle. A probabilistic approach will allow us to systematise and quantify those comparisons.

For a more general picture of relationships between pragmatic functions, it is possible to generate a graph based on counting the annotations for each turn in order to visualise the frequency with which a comment follows a question for example. This is shown in matrix form figure 4.

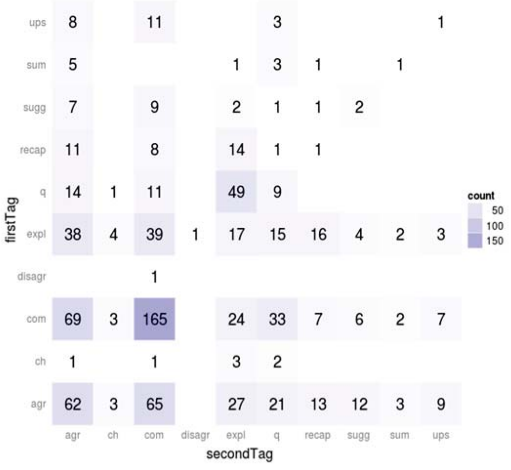


Figure 4: Tag pattern

Here we can see that a comment is most often followed by

another comment (165 counts) or less often by an agreement (69 counts). In fact, comments and agreement constitute the bulk of the interactions. Predictably, questions are mostly followed by explanations (49), more rarely by an agreement or another question. Where they do not follow questions, explanations are most likely to follow either comments (24) or agreements (27). While this representation, like that in figure 2, provides the analyst with a valuable overview of patterns within the talk, it also has the potential to be adapted at a later stage in the project so that pragmatic features can be replaced by and/or related to actions within the data set.

7. Conclusion

The approach presented here is the initial step towards identifying and mapping interactional patterns incorporating different theoretical approaches. The proposed tagging system is simple and flexible enough to add as many specific tags as necessary in order to enable us to capture the interactional patterns that underlie common set of methods or procedures that interactants employ in establishing a mutual understanding of what is getting done through their talk. While we would argue that the system is valuable in itself as a means of raising questions about, and generating insights into, the nature of the interactions taking place, it will not fully come into its own until we are able to move from the representation of individual pragmatic features to the capture of the actions of individuals participating in the talk.

8. References

Crookes, G.: 1990, The utterance, and other basic units for second language discourse analysis, *Applied Linguistics* **11**(2), 183–199.

Drummond, K. and Hopper, R.: 1993, Some uses of *Yeah*, *Research on Language and Social Interaction* **26**(2), 203–212.

Fairhurst, G. T. and Cooren, F.: 2004, Organizational language in use: Interaction analysis, conversation analysis and speech act schematics, *in* C. O. G. Grant, C. Hardy and L. Putnam (eds), *The Sage Handbook of Organizational Discourse*, Berlin: de Gruyter, pp. 131–152.

Heritage, J.: 1995, Conversation analysis: methodological aspects, *in* U. M. Quasthoff (ed.), *Aspects of Oral Communication*, Berlin: de Gruyter, pp. 391–418.

Jefferson, G.: 1983, *Notes on a systematic deployment of the acknowledgement tokens 'Yeah' and 'Mm hm'*, Tilburg University.

Levinson, B.: 1983, *Pragmatics*, Cambridge: Cambridge University Press.

Mangione-Smith, R., Stivers, T., Elliott, M., McDonald, L. and Heritage, J.: 2003, Online commentary during the physical examination: a communication tool for avoiding inappropriate antibiotic prescribing?, *Social Science and Medicine* **56**(2), 313–320.

Miller, G. and Fox, K. J.: 2004, Building bridges: The possibility of analytic dialogue between conversation analysis, ethnography and foucault, *in* D. Silverman (ed.), *Qualitative Research: Theory, Method, Practice*, London: Sage, pp. 35–54.

R Development Core Team: 2011, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: `http://www.R-project.org`

Rampton, B., Tusting, K., Maybin, J., Barwell, R., Creese, A. and Lytra, V.: 2004, UK linguistic ethnography: A discussion paper. uk linguistic ethnography forum. (accessed 27 March 2012).
URL: `http://www.lancs.ac.uk/fss/organisations/lingethn/documents/discussion_paper_jan_05.pdf`

Sarangi, S. and Roberts, C.: 2005, Theme-oriented discourse analysis of medical encounters, *Medical Education* **39**, 632–640.

Schegloff, E. A.: 1991, Reflections on talk and social structure, *in* D. Boden and D. H. Zimmerman (eds), *Studies in Ethnomethodology and Conversation Analysis*, London: Polity Press, pp. 44–70.

Schegloff, E. A.: 2005, Presequences and indirection. applying speech act theory to ordinary conversation, *Journal of Pragmatics* **12**, 55–62.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. and Ideker, T.: 2011, Cytoscape 2.8: new features for data integration and network visualization, *Bioinformatics* **27**(3), 431–432.

Van Dijk, T. A.: 1981, Episodes as units of discourse analysis, *in* D. Tannen (ed.), *Analyzing Discourse: Text and Talk*, Georgetown: Georgetown University Press, pp. 177–195.

Van Rees, M. A.: 1992, The adequacy of speech act theory for explaining conversational phenomena: A response to some conversation analytical critics, *Journal of Pragmatics* **17**(1), 31–47.

Russian Speech Corpora Framework for Linguistic Purposes

Pavel Skrelin, Daniil Kocharov

Department of Phonetics, Saint-Petersburg State University,
Universitetskaya emb., 11, 199034, Saint-Petersburg, Russia
skrelin@phonetics.pu.ru, kocharov@phonetics.pu.ru

Abstract

The paper introduces a comprehensive speech corpora framework for linguistic purposes developed at the Department of Phonetics, Saint Petersburg State University. It was designed especially for phoneticians, providing them access to speech corpora and convenient tools for speech data selection and analysis. The framework consists of three major parts: speech data, linguistic annotation and software tools for processing and automatic annotation of speech data. The framework was designed for the Russian language. The paper presents the underlying ideas of framework development and describes its architecture.

Keywords: corpora annotation, transcription, Russian, corpora application

1. Introduction

Most of the large speech corpora used in speech technology are intended for automatic collection and processing of statistical data and not for linguistic analysis of the speech data. For these purposes, it is enough to have sound recordings and their orthographic transcription or a simple phonetic transcription.

A linguist, however, often uses the corpus to test her/his hypothesis, to explain errors of automatic speech processing. The interrelationship between different levels of the language system and the way it shows up through corresponding sound patterns are of particular interest.

Obviously the corpus should contain high quality annotated speech data that provides researchers with a wide range of linguistic information. Good example of such a resource is the corpora developed for Dutch (Grønnum, 2009).

Some data processing results could be so essential or useful that it may turn out to be desirable to add them to the corpus as new annotation data for further use. Thus, the corpus annotation scheme should be scalable to enable adding new annotations to the speech data.

When studying a specific linguistic or speech phenomenon, a user would like to deal only with the parts of the corpus that have something to do with the subject of her/his research and not the whole content. To make this possible, the speech corpus needs to be accompanied with software enabling customizable search and extraction the segments with specific annotation by the given criteria.

The framework developed at the Department of Phonetics, Saint Petersburg State University consists of three major parts: speech data, linguistic annotation and software tools for processing and automatic annotation of speech data as well as for a complex search of relevant speech data using multiple search criteria. At present, two corpora of fully annotated Russian speech are used within this framework. The first one was used a material for cross-linguistic phonetic study of spontaneous speech (Bondarko, 2009). The second is used for a study of read-aloud speech (Skrelin et al., 2010). The paper presents the underlying ideas of framework development and describes its architecture.

2. General Architecture

The framework consists of three major modules. The first is speech data. The second and the most essential one is speech annotation: segmentation information and data labeling. The third one is a set of built-in tools for speech processing, basic feature extraction, statistical processing and extending corpus with linguistic and automatically generated annotation, for searching within a corpus for specific data and extracting slices of these data.

2.1. Annotation Scheme

The annotation captures the maximum amount of phonetically and prosodically relevant information. Our primary objective to ensure that the annotation of the corpus covers a wide range of information that may be of interest to those involved in most areas of linguistic research and phonetics in particular. For example, the linguistic goal is to determine spectral characteristics of [u] pronounced by a female speaker with a high pitch on a prosodic rise consistent with question intonation patterns. When selecting experimental data for this task, it is necessary to take into account various levels of annotation:1) the canonical phonetic transcription, 2) the manual phonetic transcription, 3) the word level and the position of stressed syllable as vowel quality in Russian depends on its place relative to word stress and may be the reason behind the discrepancy between phonemic and phonetic transcription, 4) intonation transcription level with the type of tone group and position of head, where the maximum rise may be expected, 5) fundamental frequency level that allows to generate a melodic curve and thus determine the parameters of the melodic rise.

There are two kinds of annotation. The first one is segmentation, i.e. information about boundaries between segmental units and their transcription labels. There are 8 main levels of segmentation, which are arranged in hierarchical order (see figure 1):

1. pitch marks;
2. manual phonetic events;

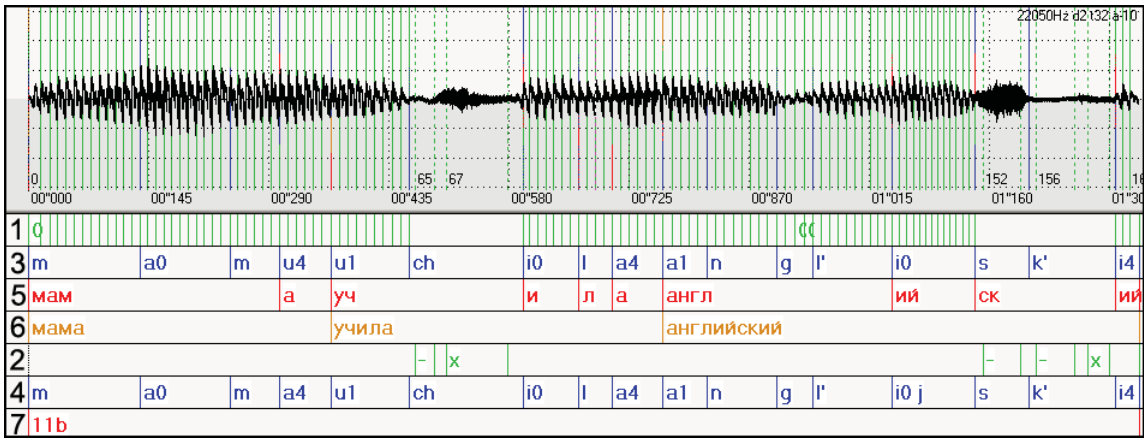


Figure 1: Annotation Scheme.

3. manual phonetic transcription (this reflects the sounds actually pronounced by the speakers);
4. rule-based phonetic transcription (this level is automatically generated by a text transcriber);
5. morphemes;
6. orthographic words;
7. prosodic units and pauses;
8. utterances.

There could be two types of manual phonetic transcription: based on acoustic analysis and based on perceptual analysis. Our experience shows that these two differ from each other. Orthographic and prosodic transcription labels, phonetic transcription labels of word-initial allophones, and fundamental frequency labels in case of voiced signal are automatically aligned with each other.

The other kind of annotation is a set of attributes of segmental units, i.e. words could be attributed as perceptually prominent words and also have other attributes showing part of speech and grammatical information.

There are two interconvertible formats we use for storing annotation. The first one is text format similar to TextGrids for Praat. The difference is that every annotation level is stored in separate text file. The other is XML format, where the annotation is stored as a hierarchical tree structure in one XML document (Tananayko et al., 2011).

Clear and unified format of files with annotation data enables use of external software for data processing and analysis. Besides, we provide tools for exporting annotation to TextGrid files for further processing with the help of Praat (Boersma and Weenink, 2012).

The annotation that could be done automatically we did by means of automatic procedures. We use manual expert annotation for two types of transcriptions: phonetic and prosodic. All the other types of transcription are done by automatic tools with a small amount of further manual correction. Experts could make mistakes, thus manual annotation is automatically validated by the help of automatic tools. Every manual annotation is processed by specially designed automatic validators before being added into annotation scheme. This procedure prevents some human mistakes and increases overall annotation quality. Validator mainly check if there are misplaced segmentation marks.

The wrongly named labels are hard to validate automatically. Sometimes the expert annotator responds to the communicative function of the speech unit and neglects objective acoustic information. In this situation a qualified phonetician acts as a native speaker (listener) and even scrupulous verification does not help.

Recently we used a ASR system to recognize sounds in the corpus. The result of automatic identification was different from expert's manual transcription in about 20% of cases. Further analysis showed that in some cases automatic identification was correct. An annotator was guided by phonological perception of a sound but not by its actual quality.

2.2. Built-in Tools for Data Processing

There are three types of built-in tools within the framework. First, tools for automatic data processing, feature extraction and analysis. Second, tools for intelligent data search and extraction. Third, tools for automatic segmentation and addition of new information to annotation.

2.2.1. Tools for automatic data processing and analysis

There is a set of command line tools written in Perl for annotation data processing and analysis tasks including:

- extraction of the essential segmental unit features, e.g. duration, amplitude, melody, and timing;
- statistical calculations concerning duration, amplitude and timing of segmental unit parameters;
- statistical calculations concerning realization of various segmental units and paralinguistic events within a given context;
- comparative analysis of various speakers, styles of speech and discourse, various contexts and collocations;

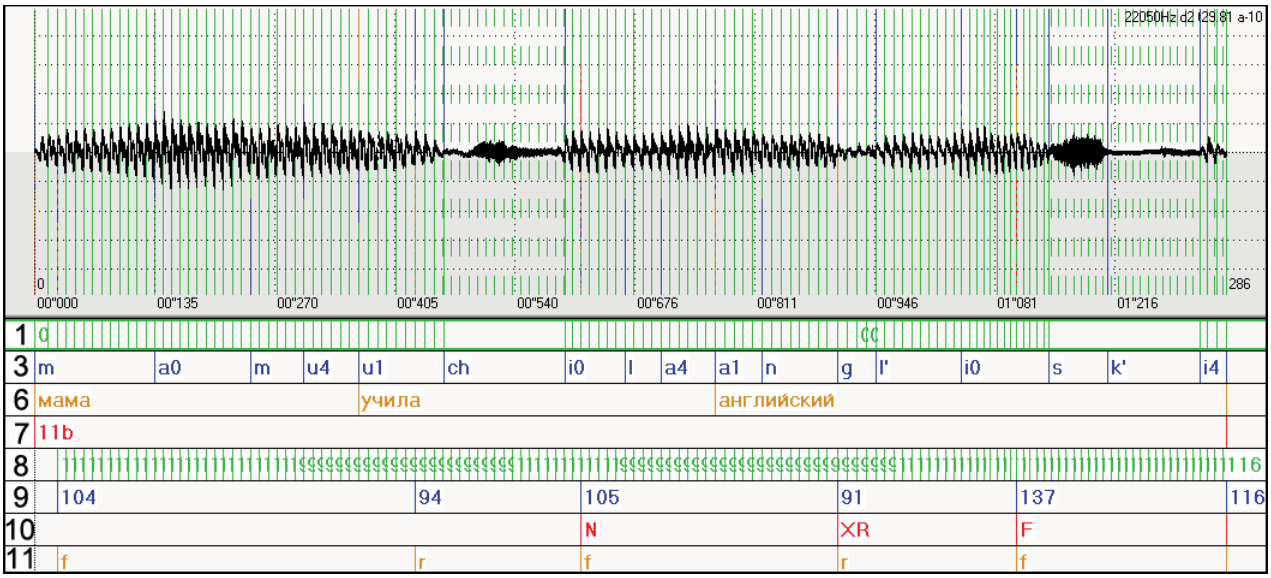


Figure 2: Illustration of additional annotation introduced to the corpus.

2.2.2. Tools for automatic extension of annotation

Automatic processing results could be incorporated into the framework both as additional attribute information of the corresponding segmental units and as new segmentation level. The flexible and scalable structure of the annotation framework allows to extend a corpus with additional annotation data which is a result of data processing and analysis. New annotation data could be processed further together with the original one in more complex cases.

For instance, the system has a tool for calculating the average duration of a specific phoneme within the selected part of the corpus. Therefore, it is very easy to determine the length of each sound relative to this average duration. A phonetician studying pre-boundary lengthening or prominence features needs to know relative sound durations in a given speech sample. Storing this attribute after it has been calculated and adding to the annotation allows accessing this information without processing the whole corpus all over again, on the one hand, and on the other hand, it makes it possible to see the duration values when manually going through the sample in Praat, for example, as it will show up as one of the annotation levels (grids).

This was the reasoning behind the option of adding new data to the annotation. If the data are features of the already existing segments, they are added as attributes. If the data introduces a new kind of segments, this information is added as a new segmentation level into appropriate place in the hierarchy of linguistic annotation levels.

2.2.3. Tools for speech data search and extraction

A corpus user often deals with the situation when s/he needs not all the data but some very specific data relevant to her/his research. For example, a phonetician studying the reasons behind prosodic prominence of specific words needs to analyze only the utterances with such prominent words. Going through the whole corpus contents in search of such samples would be inefficient, considering that only about 5 % of utterances contain prominent words.

That could be a big problem especially if s/he has no programming skills. We have solved this problem by introducing a tool for search and extraction of the segmental units specified by long context within the same segmentation level (up to five consecutive units) and specified by a context of units of the higher segmentation levels. We have command-line Perl written tools for data search and extraction. Meanwhile we are working on a tool with graphical interface for convenient data search.

That is easy if the results are for further automatic processing. But it is very important to give flexible interface in case of manual expert analysis. An expert should have a possibility to choose a data scale of the output (e.g. phoneme, morpheme, word, tone unit) and a length of the context.

A linguist studying the strategy used for the selection of introductory phrases during discourse needs to take the context into account, if not that of the whole discourse then at least that of several speech acts. At the same time, a phonetician interested in the acoustic qualities of introductory phrases would be content with taking into account only a couple of neighboring words. For this reason, we have included the tools that enable to determine the length of relevant context during the search for the segments of interest.

3. Applications of the Proposed Speech Corpora Framework

Current scientific and research tasks that use our framework include:

- phonetic research concerning acoustic and perceptual variability of speech sounds depending on events at different language levels;
- research of intra-speaker variability;
- comparative linguistic research of different styles of speech and discourse;

- comparative linguistic research of strategies of reading and speaking the same texts by different speakers.
- research of prosodic realization variability of expressing syntactic structures and semantic concepts;
- comparative research of the cues for detecting expressive and emotional speech;
- research of morphonological variability.

Recently we applied this framework for the task of automatic prosodic modeling of utterance and its prosodic type identification (Skrelin and Kocharov, 2009). To solve that task we used all available information in the corpus excluding the canonical phonetic transcription. We automatically processed the speech and annotation data thus obtaining various melodic features. Features essential for our analysis were added to the original annotation scheme (see figure 2):

- smoothed and interpolated fundamental frequency values (level 8);
- extreme values of fundamental frequency (level 9);
- boundaries of melodic movements (level 11);
- main melodic movements within the utterance corresponding to the largest drop, the largest rise, the movement reaching the global minimum, the movement reaching the global maximum (level 10).

This year we launched a project on creation the articulatory data corpus of the Russian speech. The speech data will include speech signal and articulatory data expressed by both EMA and video data. The scalable framework allows using multimedia data as the annotation is following the general ideas described above. We are able to combine annotation of different media in one annotation scheme.

4. Conclusion

The comprehensive framework for linguistic research is presented in the paper. The major features of the framework are as follows. The annotation is strictly hierarchical, scalable and allows the assignment of any number of annotation attributes to segmental units. This makes it possible to easily extend the speech corpus by the individual automatically produced annotation. There is a possibility of complex search and extraction of precise relevant slices of speech data. The output of processing result is linguistically sensible and could be individually set up in different cases.

The speech corpora framework is successfully used for many various linguistic tasks including those concerning simultaneous processing of different levels of language.

5. Acknowledgements

The authors acknowledge Saint-Petersburg State University for a research grant # 31.37.106.2011.

6. References

Paul Boersma and David Weenink. 2012. Praat: doing phonetics by computer (version 5.3.04) [computer program].

Liya Bondarko. 2009. Short description of russian sound system. In Viola de Silva and Riikka Ullakonoja, editors, *Phonetics of Russian and Finnish. General Introduction. Spontaneous and Read-aloud Speech*, pages 23–37. Peter Lang GmbH.

N. Grønnum. 2009. A danish phonetically annotated spontaneous speech corpus (danpass). *Speech Communication*, 51:594–603.

Pavel Skrelin and Daniil Kocharov. 2009. Avtomaticheskaya obrabotka prosodichekogo oformleniya viskazivaniya: relevantnye priznaki dlya avtomaticheskoj interpretatsii intonatsionnoj modeli. In *Trudy tretiego mezhdistsiplinarnogo seminara Analiz razgovornoj russkoj rechi (AR3-2009)*, pages 41–46, Saint-Petersburg.

Pavel Skrelin, Nina Volskaya, Daniil Kocharov, Karina Evgrafova, Olga Glotova, and Vera Evdokimova. 2010. Corpres – corpus of russian professionally read speech. In *Proceedings of the 13th International Conference on Text, Speech and Dialogue*, pages 386–393, Brno, Czech Republic. Springer Verlag.

Svetlana Tananayko, Daniil Kocharov, and Ksenia Sadurtinova. 2011. Programma statisticheskoj obrabobtki korpusa rechevih dannih. In *Proceedings of the 14th International Conference on Speech and Computer*, pages 457–462, Kazan, Russia. Moscow State Linguistic University.

Developing Solutions for Long-Term Archiving of Spoken Language Data at the Institut für Deutsche Sprache

Peter M. Fischer, Andreas Witt
Institut für Deutsche Sprache
R5 6-13, 68161 Mannheim, Germany
peter.fischer@ids-mannheim.de, witt@ids-mannheim.de

Abstract

This document presents ongoing work related to spoken language data within a project that aims to establish a common and unified infrastructure for the sustainable provision of linguistic primary research data at the Institut für Deutsche Sprache (IDS). In furtherance of its mission to "document the German language as it is currently used", the project expects to enable the research community to access a broad empirical base of working material via a single platform. While the goal is to eventually cover all linguistically relevant digital resources of the IDS, including lexicographic information systems such as the IDS German Vocabulary Portal, OWID, written language corpora such as the IDS German Reference Corpus, DeReKo, and spoken language corpora such as the IDS German Speech Corpus for Research and Teaching, FOLK, the work presented here predominantly focuses on the latter type of data, i.e. speech corpora. Within this context, the present document pictures the project's contributions to the development of standards and best practice guidelines concerning data storage, process documentation and legal issues for the sustainable preservation and long-term accessibility of primary linguistic research data.

Keywords: Best-Practice, Long-Term Archiving, Spoken Language Data

1. Introduction

This document presents ongoing work related to spoken language data within the project called *Zentrum für germanistische Forschungsprimärdaten* (Center for Primary Research Data in German Linguistics), funded by Germany's largest funding organization, the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) and based at Germany's principal research facility in modern German linguistics, the *Institut für Deutsche Sprache* (IDS, German Language Research Institute), in Mannheim, Germany.

In furtherance of IDS's mission to "document the German language as it is currently used", this project aims to establish a common and unified infrastructure for the sustainable provision of germane primary research data, thus to enable the research community to access a broad empirical base of working material via a single platform. While the goal is to eventually cover all linguistically relevant digital resources, including lexicographic information systems such as the IDS German Vocabulary Portal, OWID, written language corpora such as the IDS German Reference Corpus, DeReKo, and spoken language corpora such as the IDS German Speech Corpus for Research and Teaching, FOLK, the work presented here predominantly focuses on the latter type of data, i.e. speech corpora. Within this context, the present document pictures the project's contributions to the development of standards and best practice guidelines concerning data storage, process documentation and legal issues for the sustainable preservation and long-term accessibility of primary linguistic research data.

2. Spoken Language Data at the IDS

IDS maintains a variety of spoken language data resources, painstakingly collected, processed, archived and published by the Department of Pragmatics over the last few decades. Concomitantly, IDS has developed and continuously refined strategies and techniques both to optimize, facilitate the work on the data involved and to ensure appropriate and sustaining access to these resources, resulting in high proficiency in storage matters, documentation tasks and legal issues. Today, separate projects carry out different tasks of this elaborated workflow. Generally speaking, the project *Archiv für Gesprochenes Deutsch* (AGD, German Spoken Language Archive) is the first point of contact when it comes to general processing and principal archiving of corpora and all kinds of files related to them. It is also responsible for rendering them serviceable enabling the *Datenbank für Gesprochenes Deutsch* (DGD, German Spoken Language Database) to provide access to selected parts of the archive through an interactive web-based interface (cf. Fiehler and Wagener, 2005).

2.1 The archive project AGD

The AGD maintains a constantly growing portfolio of German speech data, currently comprising 44 corpora containing approx. 6,000,000 tokens and 6,000 hours of audio recordings. They include collections of numerous German dialects, colloquial and standard language and different types of conversations, which were acquired and processed by various internal and external research projects. To this end, the ADG continually accepts data from institutions, external donor projects and individual researchers who conduct language surveys or otherwise contribute spoken language data.

2.2 The database project DGD

A part of the data managed by the AGD is accessible through an interactive web-based interface maintained by the DGD. The provided content consists of manifold data collections such as recordings, documentations or aligned transcriptions. However, the corpus data accessible through DGD are limited and restricted due to their original user agreements, mostly due to the privacy rights of individuals in the recordings.

In the course of recent modernization efforts a new DGD release as German National Speech Corpus is scheduled for 2012 (cf. Deppermann and Hartung, 2011) and will be accompanied by a significant increase of accessible data, among others as part of the FOLK corpus which is included in the set of corpora available via the DGD interface.

2.3 The FOLK corpus

The *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (FOLK, German Speech Corpus for Research and Teaching; cf. Deppermann and Hartung, 2011) strives to document the German language as it is spoken today by giving insight into the reality of social communication in present-day Germany and other German-speaking regions, on the basis of a wide-ranged yet balanced collection of speech data from different areas of social life such as work, leisure, education and media. As a reference corpus for the spoken language, it aims to render the recurrent process of data acquisition in many linguistic and discourse analysis research projects unnecessary, and to provide illustrative examples of today's spoken German for academic education and language teaching performances. Hence, FOLK seeks to discover new opportunities in the fields of research and teaching.

The available resources comprise not merely the recordings but also their respective textual transcriptions as well as fine-grained alignments between transcriptions and their recording. These additional resources are created with the transcription editor FOLKER (cf. Schmidt and Schütte, 2010) following the GAT 2 transcription conventions (cf. Selting et al., 2009) that specify a modified orthography providing a blend of both pronunciation resemblance and proper (retrievable) spelling, along with rules for typical conversational phenomena such as pauses, breathing, coughing, laughing, etc. and the handling of uncertain or incomprehensible passages. Extensive metadata is also available for all resources, including the conversational circumstances and socio-demographic data on the speakers.

3. Long-Term Archiving and Best Practice

The main goal of this project is the establishment of a common and unified infrastructure for the sustainable provision of primary research data at IDS. In order to achieve this, many obstacles on different levels have to be overcome. This section of the document addresses these levels individually discussing their challenges and sketches the project's current progress in developing a solution which, since the project has just launched in December 2011, is often merely an outline albeit always with a clear line of approach. It should be noted that, although the solutions given here are narrowed down to spoken language resources, in particular to those introduced in the preceding section, however, the issues themselves in general hold equally for other kinds of language resources and presumably for other sustainability-oriented infrastructures as well.

The ground strategy for installing a long-term archiving environment is fourfold:

- Workflow models need to be developed that define concrete practical measures on how to prepare and preprocess resources archived in a "living" storage solution like the AGD, in order to appropriately channel them into the archives of a long-term storage solution like the one to be developed by this project. This particularly involves much-debated issues on the definition of adequate comprehensive and stable formats (cf. Witt, 1998; Rehm et al., 2010; Schmidt, 2011).

- Various strategies regarding the archive's accessibility must be developed, among others to guarantee that legal usage regulations are met which is a pervasive concern to speech corpora (as recordings usually raise privacy issues) and to tackle the findability problem which primary research data collections commonly used to face. The latter is closely associated with a standardized, consistent and reliable referencing system for resources.

- After technically setting up a permanent, reliable and secure storage infrastructure, the repository may then outgrow its simple data-serving functionality by particularly supporting input and output strategies as developed in the two preceding points, respectively.

- In order to sustain the environment's longevity, one must ensure that the technical systems are continuously adapted to ongoing developments and that the devised workflows become firmly established. As for the latter, intensive contact with users applying these guidelines in their own environments and feeding their experiences back to the community immediately supports the long-term effort to streamline best-practice procedures.

The following sections will expand upon the aforementioned points covering the development of best-practice guidelines.

3.1 Standards for Primary Data

As long-term archives comprise more than just their underlying repository, by committing primary research data to such an environment, users expect additional

features such as searching the data, referencing the findings, defining and accessing subsets of them or means for discovering unknown data, alongside their albeit permanent, reliable and secure but plain storage. Yet, the ability of an archive to fully support such features heavily depends on the interpretability of the data committed.

This project therefore encourages the utilization of standard data types and formats in the process of acquisition or preparation of the primary data by developing and defining a list of low-level but obligatory requirements to be met when committing data to the archive. Figuratively speaking, this can be understood as an admission ticket certifying the archivability of a resource.

3.2 Standards for Metadata

In principle, the requirement for standards for primary data holds equally for metadata, except that the range of reasonable metadata, especially in rich-metadata environments as with spoken language recordings, is entirely too broad, in order to have a comprehensive yet expedient cover of standards for it. As a consequence, a sufficiently dynamic metadata schema with a small fixed core set of obligatory items is the preferred approach. The Component Metadata Infrastructure CMDI offers possibilities to implement such a dynamic schema (cf. Broeder et al., 2011).

The aim of the project here is to define such a minimal core set of descriptors that have pan-corpus relevance along with their sets of corresponding possible values. This is important in order to render the metadata harvestable and thus the resource searchable. This is supported by the fact that the CMDI specifications fully comply with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

3.3 Identifying Resources

In 2011, the International Organization for Standardization (ISO) adopted an international standard for persistent reference to electronic language resources (ISO 24619) prepared by their Technical Committee TC37 (Terminology and other language and content resources), Subcommittee SC4 (Language resource management). It introduces a method to uniquely and persistently identify language resources by issuing persistent identifiers (PI) that can be assigned (once) to single resources rendering them referenceable (cf. Broeder et al., 2007).

This functionality is indispensable in the context of long-term archiving as the preservation of primary research data and their consequent availability necessitate their permanent referenceability and retrievability. Therefore, this project will apply the standard to all data committed to the archive. By doing so, it will have to implement a resolver as part of the repository to properly handle the translation to the location where a resource is actually stored.

3.4 Citing Resources

Another aim in the field of standard procedures for bibliographically cataloguing resources is the introduction of metadata covering the linking between primary data and the publications citing them. It is important for researchers to get acknowledged for compiling primary data, because publications are often the only impact factors. Moreover, it can be useful to find out what research has been conducted on a specific resource. On the other hand, in publications, the information about an empirical base is often given in a very informal way, occasionally encoded within the text. Making this information more explicit is therefore highly desirable.

This project recently entered into cooperation with the project called *Integration von Forschungsdaten und Literatur in den Sozialwissenschaften* (InFoLiS, Integrating Research Data and Literature in the Social Sciences), a fellow project that is based at Germany's largest infrastructure institution for the Social Sciences, GESIS, and aims to automatically detect and cross-link publications.

3.5 Technical awareness

The software system that conceptually underlies such a repository must be considered very carefully, as its sufficiently modular architecture is crucial to achieve technical sustainability. In particular, the encapsulation of storage-related matters from internal business logics and, in turn, from front-end applications prepares for incessant adjustments that inevitably come with technical progress over time.

This project has evaluated some available open-source repository systems that (sometimes in combination and to their respective degree) meet the aforementioned requirements. Fully stand-alone developments include OPUS4[1] (licensed under the GNU GPL) and DSpace[2] (shared under a BSD license), whereas systems based on Fedora Commons[3] (licensed under a Creative Commons License) include eSciDoc[4] (distributed under the CDDL) and the interaction of the content management platform Drupal[5] with the digital asset management system Islandora[6] (both licensed under the GNU GPL). At this stage, this project favours the latter solution.

3.6 Community awareness

DFG has acknowledged that long-term preservation of primary research data is necessary and important, and is funding a multidisciplinary list of various projects that are engaged in this direction. However, as solutions are being developed, two opposing implementation principles

[1] http://www.kobv.de/opus4
[2] http://www.dspace.org/
[3] http://www.fedora-commons.org/
[4] http://www.escidoc.org/
[5] http://www.drupal.org/
[6] http://www.islandora.ca/

emerge: *in situ* repositories primarily processing subject-specific data with highly specialized applications in their respective fields on the one hand, and fairly interdisciplinary, usually centralized repositories with a focus on catholicity and general interoperability on the other hand.

This project attempts to effect a compromise by implementing a technically centralized storage solution with leaving the control over the data (like structuring, documentation, accessing policies, etc.) as far as possible to the data providers. Also, the repository specializes on linguistic research data to its most general extent, as one goal of the project is to eventually cover all linguistically relevant IDS-internal digital resources, including lexicographic information systems, written language and spoken language corpora alike. To this end, the project has also entered into cooperation with some external partners like the *Hamburger Zentrum für Sprachkorpora* (HZSK, Hamburg Center for Speech Corpora), based at the University of Hamburg, Germany, and the *Institut für Deutsche Sprache und Linguistik* (German Language and Linguistics Research Institute), based at the Humboldt University of Berlin, Germany.

4. Perspective

While proximate development clearly focuses on means to allow for linguistic resources other than speech-related data to be included in the preservation process, in the long term the project is contemplating opening up the platform for third-party participation by providing some kind of upload mechanism in order to enable external partners to have their non-IDS data preserved.

5. References

Broeder, Daan; Declerck, Thierry; Kemps-Snijders, Marc; Keibel, Holger; Kupietz, Marc; Lemnitzer, Lothar; Witt, Andreas; Wittenburg, Peter (2007). Citation of Electronic Resources: Proposal for a new work item in ISO TC37/SC4. ISO TC37/SC4-Document N366. http://www.tc37sc4.org/new_doc/ISO_TC37_SC4_N366_NP_CitER_Annex.pdf.

Broeder, Daan; Schonefeld, Oliver; Trippel, Thorsten; Van Uytvanck, Dieter; Witt, Andreas (2011). A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI). In: Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies, vol. 7. doi: 10.4242/BalisageVol7.Broeder01.

Deppermann, Arnulf; Hartung, Martin (2011). Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des 'Forschungs- und Lehrkorpus Gesprochenes Deutsch' (FOLK) am Institut für Deutsche Sprache (Mannheim). In: Felder, Ekkehard; Müller, Marcus; Vogel, Friedemann (Eds.). Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. Berlin, New York: de Gruyter, pp. 414-450.

Fiehler, Reinhard; Wagener, Peter (2005). Die Datenbank Gesprochenes Deutsch (DGD) – Sammlung, Archivierung und Untersuchung gesprochener Sprache als Aufgaben der Sprachwissenschaft. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion. 6/2005, pp. 136-147. http://www.gespraechsforschung-ozs.de/heft2005/px-fiehler.pdf.

Rehm, Georg; Schonefeld, Oliver; Trippel, Thorsten; Witt, Andreas (2010). Sustainability of Linguistic Resources Revisited. In: Proceedings of the International Symposium on XML for the Long Haul. Issues in the Long-term Preservation of XML. Balisage Series on Markup Technologies, vol. 6. doi:10.4242/BalisageVol6.Witt01.

Schmidt, Thomas; Schütte, Wilfried (2010). FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction. In: Calzolari, Nicoletta et al. (Eds.). Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta: European Language Resources Association (ELRA), pp. 2091-2096. http://www.lrec-conf.org/proceedings/lrec2010/pdf/18_Paper.pdf.

Schmidt, Thomas (2011). A TEI-based approach to standardising spoken language transcription. In: Journal of the Text Encoding Initiative, 1. http://jtei.revues.org/142

Selting, Margret; Auer, Peter; Barth-Weingarten, Dagmar; Bergmann, Jörg; Bergmann, Pia; Birkner, Karin; Couper-Kuhlen, Elizabeth; Deppermann, Arnulf; Gilles, Peter; Günthner, Susanne; Hartung, Martin; Kern, Friederike; Mertzlufft, Christine; Meyer, Christian; Morek, Miriam; Oberzaucher, Frank; Peters, Jörg; Quasthoff, Uta; Schütte, Wilfried; Stukenbrock, Anja; Uhmann, Susanne (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung (10), pp. 353-402. http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf.

Witt, Andreas (1998). TEI-based XML-Applications: Transcriptions. In: ALLCACH98, Joint Conference of the ALLC and ACH, Debrecen, pp. 170-174.

Using A Global Corpus Data Model for Linguistic and Phonetic Research

Christoph Draxler

BAS Bavarian Archive of Speech Signals
Institute of Phonetics and Speech Processing
Ludwig-Maximilian University Munich, Germany
draxler@phonetik.uni-muenchen.de

Abstract

This paper presents and discusses the global corpus data model in WikiSpeech for linguistic and phonetic data used at the BAS. The data model is implemented using a relational database system. Two case studies illustrate how the database is used. In the first case study, audio recordings performed via the web in Scotland are accessed to carry out formant analyses of Scottish English vowels. In the second case study, the database is used for online perception experiments on regional variation of speech sounds in German. In both cases, the global corpus data model has shown to an effective means for providing data in the required formats for the tools used in the workflow, and to allow the use of the same database for very different types of applications.

1. Introduction

The Bavarian Archive for Speech Signals (BAS) has collected a number of small and large speech databases using standalone and web-based tools, e.g. Ph@ttSessionz (Draxler, 2006), VOYS (Dickie et al., 2009), ALC (Schiel et al., 2010), and others. Although these corpora were mainly established to satisfy the needs of speech technology development, they are now increasingly used in phonetic and linguistic basic research as well as in education. This is facilitated by the fact that these speech databases are demographically controlled, well-documented and quite inexpensive for academic research.

The workflow for the creation of speech corpora consists of the steps *specification*, *collection* or *recording*, *signal processing*, *annotation*, *postprocessing* and *distribution* or *exploitation* – each step involves using dedicated tools, e.g. SpeechRecorder (Draxler and Jänsch, 2004) to record audio, Praat (Boersma, 2001), ELAN (Sloetjes et al., 2007), EXMARaLDA (Schmidt and Wörner, 2005), or WebTranscribe (Draxler, 2005) to annotate it, libassp (`libassp.sourceforge.net`), sox (`sox.sourceforge.net`) or Praat to perform signal analysis tasks, and Excel, R and others to carry out statistical computations (Figure 2).

All tools use their own data formats. Some tools do provide import and export of other formats, but in general there is some loss of information in going from one tool to the other (Schmidt et al., 2009). A manual conversion of formats is time-consuming and error-prone, and very often the researchers working with the data do not have the programming expertise to convert the data. Finally, if each tool provides its own import and export converter, the number of such data converters increases dramatically with every new tool.

We have thus chosen a radical approach: we store all data independent of any application program in a database system in a *global corpus data model* in our WikiSpeech system (Draxler and Jänsch, 2008).

This global data model goes beyond the general annotation graph framework for annotations proposed by Liberman and Bird, which provides a formal description of the data structures underlying both time-aligned and non time-aligned annotations (Bird and Liberman, 2001). In our global corpus data model, annotations consist of tiers which in turn contain segments; segments may be time-aligned with or without duration, or symbolic, i.e. without time data. Within a tier segments are ordered sequentially, and between tiers there exist 1:1, 1:n and n:m relationships.

Besides annotations, our global corpus data model also covers audio and video recordings, technical information on the recordings, time-based and symbolic annotations on arbitrarily many annotation tiers, metadata on the speakers and the database contents and administrative data on the annotators (Figure 1.).

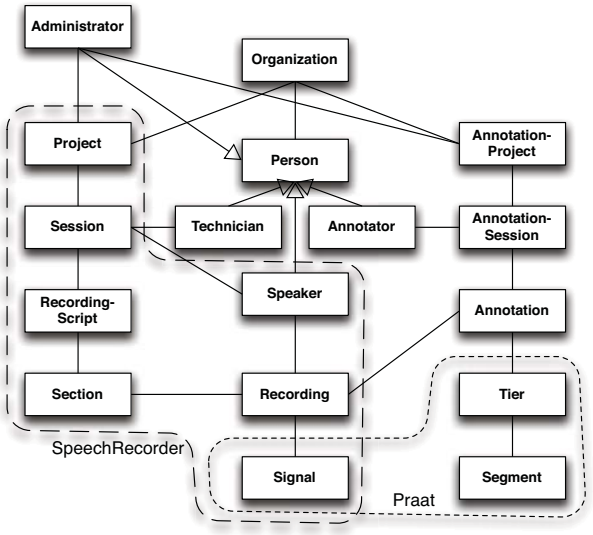


Figure 1: Simplified view of the global corpus data model in WikiSpeech. The dashed lines show which parts of the data model are relevant to given tools. For example, Praat handles signal files and knows about annotation tiers and (time-aligned) segments on these tiers. SpeechRecorder knows recording projects that contain recording sessions which bring together speakers and recording scripts; a script contains sections which in turn are made up of recording items which result in signal files.

The global corpus data model is a superset of the individual tools' data models. Hence it is, within limits, possible to import data from and export data to any of the tools in the workflow. This import and export is achieved via scripts that convert from the database to the external application program format and back.
Finally, the global corpus data model has evolved over the years and has now reached a stable state. It has shown to be sufficiently powerful to incorporate various speech databases. These databases were either created directly using WikiSpeech, or existing speech databases, whether produced by BAS or elsewhere, were successfully imported into WikiSpeech.
A relational database using the PostgreSQL database management system implements the global corpus data model. Using a relational database system has several advantages: there is a standard query language (namely SQL), many users can access the data at the same time, query evaluation is efficient, external applications can access the database via standard APIs, data migration to a different database vendor is straightforward, and data can be exported into plain text easily.

2. Workflow

Figure 2 shows the general workflow for phonetic and linguistic corpus creation and exploitation.

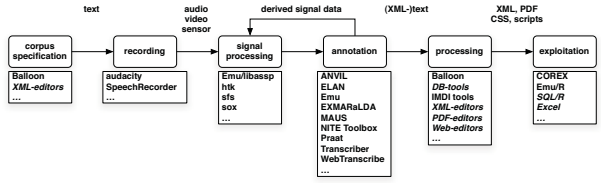


Figure 2: Speech database creation and exploitation workflow and a selection of tools used in the individual tasks.

Note the different types of media involved: specifications, annotations, log file and reports are text data in a variety of formats (XML, plain text, PDF, or DOC), signal data are time-dependent audio, video and sensor data, again in a variety of formats.
Plain text and XML formatted data is generally stored in the database, whereas other types of text data are stored in the file system, with only their local address stored in the database. Media data is always stored in the file system; the database contains merely the links.

3. Case study 1: Formant analysis of Scottish English

Since 2007 BAS has been recording adolescent speakers in the VOYS project via the web in grammar schools in the metropolitan areas of Scotland (Dickie et al., 2009). Until now, 251 speakers and a total of 25466 utterances have been recorded in 10 recording locations. Each recording session yields metadata on the speaker and the recording location, and consists of up to 102 utterances of read and prompted speech. This data is entered into the database immediately via web forms or background data transfer during the recording session.

3.1. Annotation

Once a recording session is complete, the data is made available for a basic annotation using WebTranscribe (Draxler, 2005). For read speech, the prompt text is displayed in the editor so that the annotator simply needs to modify the text according to the annotation guidelines and the actual content of the utterance. For non-scripted speech, the annotator has to enter the entire content of the utterance manually (automatic speech recognition does not yet work very well for non-standard English of adolescent speakers).

3.1.1. Automatic segmentation

Once the orthographic contents of the utterance have been entered, the automatic phonetic segmentation is computed using MAUS (Schiel, 2004). MAUS is an HMM-based forced alignment system that takes into account coarticulation and uses hypothesis graphs to find the most probable segmentation. MAUS may be adapted to other languages than German in two ways: either via a simple translation table from German to the other languages phoneme inventory, or by retraining MAUS with a suitably large training database. For Scottish English, a simple mapping table proved to be sufficient.
MAUS generates either Praat TextGrid or BAS Partitur File formatted files. These are imported into the segments table of the database using a perl script.

3.1.2. F0 and formant computation

In a parallel process, a simple shell script calls Praat to compute the formants of all signal files and writes the results to a text file which is then imported into the database.
At this point, the database contains metadata on the speakers and the recording site, technical data on the recordings, an orthographic annotation of the utterances, an automatically generated phonetic segmentation of the utterances, and time-aligned formant data for all utterances.
It is now ready to be used for phonetic research, e.g., a formant analysis of vowels of Scottish English.

3.2. Querying the database

The researcher logs into the database and extracts a list of segments he or she wishes to analyse. For a typical formant analysis, this list would contain the vowel labels, begin and end time of the segment in the signal file, age, sex and regional accent of the speaker, and possibly the word context in which the vowel is used. This request is formulated as an SQL query (Figure 3).
Clearly, linguists or phoneticians cannot be expected to formulate such queries. Hence, to facilitate access to the database for researchers with little experience in SQL, the database administrator may define so-called views, i.e. virtual tables which are a syntactic shorthand for queries over many relation tables (Figure 4).

3.3. Statistical analysis

For a statistical analysis, the researcher uses a simple spreadsheet software such as Excel or, preferably, a statistics package such as R. Both Excel and R have database interfaces for almost any type of relational database (in general, these interfaces have to be configured by the system

```
select m.label as phoneme, k.label as canonic, o.label as
     word, count(f.f1), avg(f.f1)::int as f1, avg(f.f2)::int as f2

from session ses
join signalfile sig on ses.session = substring(sig.filename, 3, 9)
join segment m on sig.id = m.signal_id and m.tier = 'MAU:'
join segment o on m.signal_id = o.signal_id and
     m.ref_seg = o.ref_seg and o.tier = 'ORT:'
join segment k on k.signal_id = o.signal_id and
     k.ref_seg = m.ref_seg and k.tier = 'KAN:'
join formant f on m.signal_id = f.signal_id and
     f.time between (m.begin_seg + (m.dur_seg * 0.2)) and
     (m.begin_seg + (m.dur_seg * 0.8))
join speaker spk on ses.session = spk.speaker_code and
     ses.project = spk.project and ses.project = 'VOYS'

where m.label = 'E'

group by m.label, k.label, o.label
order by m.label, k.label, o.label;
```

Figure 3: Sample SQL query to retrieve phoneme segments, their count and the average f1 and f2 values, grouped by phoneme segment label, and word

```
select phoneme, avg(f1)::int, avg(f2)::int
from voys_data
where phoneme = 'E'
group by phoneme
```

Figure 4: Using a view (a predefined virtual table) to express the same query as in Figure 3

administrator). The researcher can now access the database directly from the spreadsheet or statistics software he or she is using, perform the statistical analyses and display the results in text format or diagrams (Figure 5).

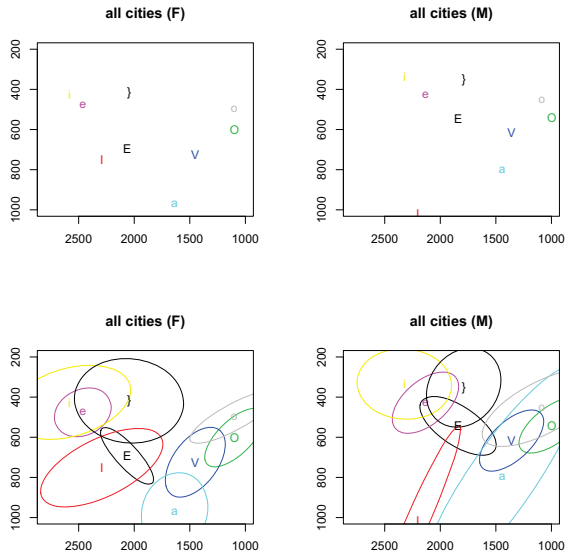


Figure 5: Formant charts for Scottish English vowels from the VOYS speech database

Using this workflow and the global corpus data model, similar analyses or analyses of other corpora may now be performed with little effort and can thus be used in education.

4. Case study 2: A perception experiment on regional variants of sounds

Perception experiments are an essential part of phonetic, linguistic and psychological research. Most such experiments are now performed using a computer, and some speech processing tools, e.g. Praat, directly support performing perception experiments. However, with standalone software, performing an experiment requires that the software is installed on every computer on which the experiment will run. Web-based online experiments overcome this limitation and provide access to potentially large groups of participants (Reips, 2002).

Currently, only a few tools or services exist that allow online experiments with audio. Examples are WebExp (Keller et al., 2009) and Percy (Draxler, 2011). WebExp is a flexible and powerful online experiment software that uses Java applets for media display, and which stores its result data in XML files. Percy is based on HTML5 and stores its data in a relational database on the server.

4.1. Perception experiments in the workflow

In principle, a perception experiment is not very different from an annotation task – speech material is presented and the participant has to enter his or her judgment. Hence, the global corpus data model also covers online perception experiments. Experiment results are simply considered as yet another annotation tier. This allows the same data retrieval mechanism to be used for experiment data, which greatly simplifies further processing.

4.2. Running the experiment

A recent perception experiment on regional variation uses the single digit items recorded in the Ph@ttSessionz project. Participants were asked to judge whether a given phoneme in a digit has certain properties, e.g. whether the initial "s" in the word "sieben" was voiced or voiceless. The question was formulated in colloquial terms so that non-experts could participate in the experiment.

The experiment consists of three steps: the participant registers and provides some personal and context information. During the experiment he or she listens to the recorded audio and enters a judgment, and when all items are done, the experiment displays a map with the geographic locations where the audio files were recorded (Figure 6).

4.3. Statistical analysis

With support from the statistics lab of LMU University, the experiment input data was analysed using mixed-models provided by the R software. Results show that a) sounds effectively differ from one region to the other, and b) that the perception of sound difference depends on the regional background of the listener.

5. Discussion

A global corpus model is a suitable tool to support the workflow in phonetic and linguistic research, development, and education.

However, serious problems remain. The most important are

1. Missing data in the database may lead to broken workflows or inconsistent or unexpected results.

2. New tools and services may not fit into the global corpus model.

a) b) c)

Figure 6: Web experiment on regional variation of speech sounds in German using the Percy software. a) registration web form, b) experiment item screen, c) final display of geographic distribution of recordings.

3. Data exchange with tools with different underlying data models may require manual adaptation.

4. It is debatable whether SQL is a suitable query language for phoneticians or linguists, and whether it is sufficiently powerful to express typical research questions.

5.1. Missing data

Missing data can be controlled by enforcing a minimum standard set of metadata and log data to be collected during the speech database creation. However, this may not be feasible when the speech database comes from an external partner. It is thus necessary to manually check external databases before incorporating them into the WikiSpeech system, and to document the data manipulations applied.

Database systems provide a default null value for unknown data, which at least ensures that queries can be run. If queries return unexpected results, then default null data maybe the reason for this. In a relational database the database schema can be inspected either via queries or via a graphical user interface, and thus it is in general quite straightforward to find out whether missing data is a possible reason for unexpected query results.

5.2. Incorporating new tools

A database schema is intended to remain stable over time so that predefined views, proceduralized queries or scripts continue to work. However, new tools and services may use data items not present in the database. If the new data cannot be represented the given database schema, this schema has to be extended. In general, simply extending the data model, e.g. by adding new attributes in the relation tables, or even adding new tables, is not critical. Critical changes include changing the relationships between data items, or removing attributes or relational tables. Such changes however will be very rare because the data model has reached a very stable state by now.

Any change to the data model can only be performed by the database administrator, and it may entail the modification of existing scripts and queries.

5.3. Manual adaption of data

For tools that use a different underlying data model, any data that is exported to the tool and then reimported must be processed to minimize the loss of information. For example, in a data model that uses symbolic references between elements on different annotation tiers, all items must be given explicit time stamps and unique ids to be used in a purely time-based annotation tool such as e.g. Praat. Upon reimporting the data after processing in Praat, these timestamps have to be removed for non time-aligned annotations. Such modifications are in generally implemented using a parameterized import and export script. Several such scripts may be necessary for the different tools used in the workflow.

5.4. SQL as a query language

SQL is the de facto standard query language for relational databases. However, it is quite verbose and lacks many of the operators needed in phonetic or linguistic research, namely sequence and dominance operators. Both sequence and dominance operators can be expressed by providing explicit position attributes or linking relations between data records, but they make queries even longer and more complex.

The SQL view mechanism is a simple way of formulating simple queries. For example, in the web experiment, where experiment setup and input spread over 8 relational tables, a single view contains all data fields of the experiment and appears to the user as one single wide table, which can be directly imported into R or Excel.

As an alternative to the direct use of SQL, high-level application domain specific query languages can be envisaged, which are then automatically translated to SQL for execution. This separation of high-level query language and the SQL query evaluation is desirable, because it opens the possibility to provide many different query language to the same underlying database. Such query languages can be text-based or graphical, very specific to a particular application domain or quite abstract. In fact, many graphical front-ends to database systems already allow form-like query languages or explorative interactive graphical query languages.

6. Conclusion

The global corpus data model for speech databases is a pragmatic approach to supporting phonetic and linguistic research. It provides a means to exchange data with a multitude of speech processing tools, and allows proceduralizing often-needed research and development tasks. The two case studies show that quite different tasks can be performed on the same data base in parallel.

The global corpus data model will slowly evolve, and it will be slightly different in different research and development labs, because of different requirements and tools used. However, simply the fact that there exists a global corpus data model with visible and formally specified data structures, e.g. in a relational schema description, will lead to a much higher degree of consistency and coverage in speech database creation.

One major challenge is the development of a query language suitable for linguistic and phonetic researchers. This query language must be much closer to the application domain than SQL can ever be, and it must be sufficiently powerful to express the queries that researchers in phonetics or linguistics ask. A promising approach is to provide a query language such as XQuery or a graphical query interface to the database base, and to compile this query into SQL for efficient execution in the database.

Acknowledgements

The author thanks the Statistical Consulting Unit, Department of Statistics, Ludwig-Maximilians-Universitt, Munich for its support of the analysis of the German dialect perception experiment data. Thanks also go the the pupils and schools in Scotland who participated in the Scottish English VOYS recordings, Catherine Dicke and Felix Schaeffler for organizing the recordings in Scotland, and the students at LMU who transcribed and annotated these recordings.

7. References

St. Bird and M. Liberman. 2001. A Formal Framework for Linguistic Annotation. *Speech Communication*, 33(1,2):23–60.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

C. Dickie, F. Schaeffler, Chr. Draxler, and K. Jänsch. 2009. Speech recordings via the internet: An overview of the VOYS project in Scotland. In *Proc. Interspeech*, pages 1807–1810, Brighton.

Chr. Draxler and K. Jänsch. 2004. SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software. In *Proc. of LREC*, pages 559–562, Lisbon.

Chr. Draxler and K. Jänsch. 2008. Wikispeech – a content management system for speech databases. In *Proc. of Interspeech*, pages 1646–1649, Brisbane.

Chr. Draxler. 2005. Webtranscribe – an extensible web-based speech annotation framework. In *Proc. of TSD 2005*, pages 61–68, Karlsbad, Czech Republic.

Chr. Draxler. 2006. Exploring the Unknown – Collecting 1000 speakers over the Internet for the Ph@ttSessionz Database of Adolescent Speakers. In *Proc. of Interspeech*, pages pp. 173–176, Pittsburgh, PA.

Chr. Draxler. 2011. Percy – an HTML5 framework for media rich web experiments on mobile devices. In *Proc. Interspeech*, pages 3339–3340, Florence, Italy.

F. Keller, G. Subahshini, N. Mayo, and M. Corley. 2009. Timing accuracy of web experiments: A case study using the webexp software package. *Behavior Research Methods, Instruments and Computers*, 41(1):1–12.

U. Reips. 2002. Standards for internet-based experimenting. *Experimental Psychology*, 49(4):243–256.

F. Schiel, Chr. Heinrich, S. Barfüßer, and Th. Gilg. 2010. Alcohol Language Corpus – a publicly available large corpus of alcoholized speech. In *IAFPA Annual Conference*, Trier, Germany.

F. Schiel. 2004. MAUS goes iterative. In *Proc. of LREC*, pages 1015–1018, Lisbon, Portugal.

Th. Schmidt and K. Wörner. 2005. Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. *Gesprächsforschung*, Vol. 6:171–195.

T. Schmidt, S. Duncan, O. Ehmer, J. Hoyt, M. Kipp, D. Loehr, M. Magnusson, T. Rose, and H. Sloetjes. 2009. An exchange format for multimodal annotations. In *Multimodal Corpora*, volume 5509 of *Lecture Notes in Computer Science*, pages 207–221. Springer Verlag.

H. Sloetjes, A. Russel, and A. Klassmann. 2007. ELAN: a free and open-source multimedia annotation tool. In *Proc. of Interspeech*, pages 4015–4016, Antwerp.

Best Practices in the TalkBank Framework

Brian MacWhinney, Yvan Rose, Leonid Spektor, and Franklin Chen
Carnegie Mellon University, Psychology
5000 Forbes Ave. Pittsburgh, PA 15213
E-mail: macw@cmu.edu

Abstract

TalkBank is an interdisciplinary research project funded by the National Institutes of Health and the National Science Foundation. The goal of the project is to support data sharing and direct, community-wide access to naturalistic recordings and transcripts of spoken communication. TalkBank has developed consistent practices for data sharing, metadata creation, transcription methods, transcription standards, interoperability, automatic annotation, and dissemination. The database includes corpora from a wide variety of linguistic fields all governed by a comprehensive XML Schema. For each component research subfield, TalkBank must provide special purpose annotations and tools as a subset of the overall system. Together, these various TalkBank standards can serve as guides to further improvements in the use of speech corpora for linguistic research.

Keywords: corpora, speech, conversation analysis, phonology, syntax, transcription, metadata

1. Best Practices

The goal of this workshop is to examine best practices for configuring speech corpora for linguistic research. This would seem to be a fairly well defined goal. Ideally, one could formulate a single set of best practices that would apply across the board. However, when we consider specific corpora, systems, groups, issues, and constraints, the characterization of "best practices" becomes more complicated. Take the CallFriend corpus, as an example. The Linguistic Data Consortium (LDC) created this phone call corpus for the purposes of developing automatic speech recognition (ASR) systems. Thanks to the generosity of LDC, segments of CallFriend have been made available to the TalkBank system for transcription and further linguistic analysis. We have transcribed these calls in the CHAT editor, using Conversation Analysis standards and linked them on the utterance level to the audio media. The best practices in this case depend heavily on the particular shape of the corpus and the uses to which it will be put. These are phone calls with good stereo separation, but there are often noises on the phone line. This seems to violate best practices in speech technology, but it is quite adequate for the purposes of Conversation Analysis. On the other hand, the demographic information associated with each call is inadequate for standard sociolinguistic or sociophonetic analysis. Also, LDC provided no transcriptions for these calls, so the issue of best practices in transcription rests totally outside of the realm of the initial data collection.

When we consider best practices across a wide collection of corpora, the problem becomes further magnified. Ini particular, for each of the 386 corpora in the TalkBank database, collected under a myriad of different conditions with differing goals, we could conduct an analysis of best practices, usually with quite different results. This suggests that we should view best practices not as a single framework, but as a Swiss Army knife that presents the user with a variety of tools, each suited for a given type of linguistic analysis.

The TalkBank system is an atttempt to provide just this type of Swiss Army knife. For researchers studying child phonology, it offers the PhonBank system (Rose & MacWhinney, in press). For morphosyntactic analysis, it provides taggers (MacWhinney, 2008) and parsers (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2010). For Conversation Analysts, it provides Jeffersonian coding (Jefferson, 1984) and formats for gestural analysis (MacWhinney, Fromm, Forbes, & Holland, 2011). Some of the blades of the knife can be used for many purposes; others are more specialized. In this report, we will explain how each blade has been adapted to the task at hand. In some cases, the blades offered by TalkBank are not the best available and we need to then explain how data in the TalkBank format can then be exported to other programs. In other areas, such as metadata coding, TalkBank has essentially off-loaded the issue of best practices to other systems.

2. Background

TalkBank (http://talkbank.org) is an interdisciplinary research project funded by the National Institutes of Health and the National Science Foundation. The goal of the project is to support data sharing and direct, community-wide access to naturalistic recordings and transcripts of spoken communication. TalkBank extends the model for data sharing and analysis first developed in the context of the CHILDES project (MacWhinney, 2000). Although CHILDES is the most established of these datasets, other systems, such as PhonBank, AphasiaBank, CABank, BilingBank, and SLABank have also achieved general recognition and acceptance within the relevant research communities.

CHILDES contains 68 million words of child-adult

conversation across 26 languages; the other segments of TalkBank include 63 million words of adult-adult conversation with the bulk in English. Although many earlier child language corpora were not contributed along with their media, the current default format for both CHILDES and TalkBank assumes that transcripts will be linked to either audio or video on the level of the utterance. This means that all new TalkBank corpora are, in effect, speech corpora. To the degree that the methods of speech technology can be applied to naturalistic conversational data of the type collected in TalkBank, the merger of speech technology with linguistic analysis envisioned in this workshop has already taken place in the TalkBank framework.

This workshop has specified a set of 12 themes for analysis of best practices. These are:

1. speech corpus designs and corpus stratification schemes
2. metadata descriptions of speakers and communications
3. legal issues in creating, using and publishing speech corpora for linguistic research
4. transcription and annotation tools for authentic speech data
5. use of automatic methods for tagging, annotating authentic speech data
6. transcription conventions in conversation analysis, dialectology, sociolinguistics, pragmatics and discourse analysis
7. corpus management systems for speech corpora
8. workflows and processing chains for speech corpora in linguistic research
9. data models and data formats for transcription and annotation data
10. standardization issues for speech corpora in linguistic research
11. dissemination platforms for speech corpora
12. integration of speech corpora from linguistic research into digital infrastructures

TalkBank addresses issues 2, 3, 4, 5, 6, 7, 9, and 11. Issues 1, 8, 10, and 12 lie outside the scope of TalkBank and are left either to individual researchers or the wider scientific community. In the next sections, we will outline TalkBank approaches to the eight best practices issues it has addressed.

3. Metadata

TalkBank has addressed the Metadata issue by subscribing to both the OLAC and IMDI formats. For each of the 386 corpora in TalkBank, we create a single text file in a consistent format that provides information relevant to all files in the corpus. The OLAC program, which is built into the CLAN programs, compiles this information across the database into a single file for harvesting by OLAC. For IMDI, we also include headers in the individual files that provide further file-specific metadata. Rather than using the ARBIL program, we use the IMDI program in CLAN to combine this information into files that can be included in IMDI. In addition, all of the transcripts and media of the complete CHILDES and TalkBank databases are included in IMDI and freely available through that system.

We are working to specify further detailed best practice specifications for metadata in the area of sociolinguistics. Toward that end, we have contributed to recent workshops organized by Chris Cieri and Malcah Yaeger-Dror at NWAV and LSA, designed to improve best practices in the coding of sociolinguistic metadata and to stimulate data-sharing in that field. To facilitate the declaration of sociolinguistic and other corpora, we have provided a page at http://talkbank.org/metamaker that allows researchers to describe the shape and availability of corpora that are not yet included in any major database. This information is then transmitted to OLAC.

4. Legal Issues and Data Sharing

The 386 corpora in TalkBank have all cleared IRB review, and nearly all are available for open access and downloading. In the process of establishing this level of open access, we have acquired decades of experience with IRB and legal issues. The results of this experience are encoded in a set of principles for data-sharing, IRB guidelines, suggested informed consent forms, alternative levels of access or password protection, and methods for anonymizing data, all available from http://talkbank.org/share

In practice, the only corpora that require password access are those from participants with clinical disabilities. For the other corpora, we are careful to replace last names with the capitalized English word “Lastname” and addresses with the word “Address”. For some corpora, such as the Danish SamtaleBank corpus, we have also replaced the last names and addresses in the audio files with silence.

Often researchers claim that their data cannot be shared, because access has not been approved by their IRB. In practice, we have found that this is seldom the case. IRBs will nearly always approve data sharing with anonymization and password protection. In reality, researchers use IRB restrictions as a way of avoiding opening their data to other investigators, because they believe that other researchers can achieve a competitive advantage. In this sense, the reference to legal and IRB issues is frequently used to divert discussion of the underlying problem of competitive advantage in academics. We believe that best solution to this problem is for granting agencies to require data sharing as a condition for further funding.

5. Transcription Tools

Apart from its scope, coverage, multilinguality, and size, there is another core feature that characterizes TalkBank. This is the fact that all of the data in the system are formatted in accord with a single consistent standard called CHAT that is bidirectionally convertible to

TalkBank XML (http://talkbank.org/xsddoc). Over the years, CHAT has been crafted as a superset of its component transcription standards. For example, it supports at the same time standard Jeffersonian Conversation Analysis (CA) coding, the linguistically-oriented transcription methods of child language, phonological coding methods through IPA, disfluency analysis methods for speech errors and stuttering, and new methods for gesture coding in nested dependent files. Each of these transcription standards is implemented as a subcomponent of the overall TalkBank CHAT standard and individual transcripts can declare to which set of conventions they adhere. This approach allows us to provide all the codes that are needed for each subdiscipline without requiring any of them to make use of all the codes for their own special corpora.

The benefit of this approach is that the analysis programs can operate on all corpora in a consistent way and users only need to learn the CLAN program (http://talkbank.org/software) to analyze everything in TalkBank. In this regard, the TalkBank framework differs fundamentally from that of other systems such as LDC or Lacito. These other archiving system accept corpora in a wide variety of formats and users must learn different tools and methods to process each of the alternative corpora, even within a particular topic area.

The imposition of consistent coding standards comes at a cost. Transcription in CHAT can be rigorous and demanding. For the beginner, it takes several days to learn to transcribe smoothly. In some other cases, researchers are unwilling to use CHAT at all and prefer to create corpora in their own formats. When those corpora are contributed to the database, we then write special purpose programs to reformat them. However, we can automatically convert corpora formatted in SALT, ELAN, EXMARaLDA, Transcriber, Praat, or ANVIL.

To facilitate the mechanics of transcription, the CLAN editor supports several methods of linking to the media during and after transcription. These methods include Transcriber Mode, Sound Walker Mode, Sonic Mode, and Hand Editing Mode. Transcriber Mode uses the space-bar method of the Transcriber program (http://trans.sourceforge.net). Sound Walker mode operates like the old dictation machine with an optional foot pedal. Sonic Mode relies on display of the waveform for both audio and video files. We are interested in further improvements of CHAT transcription based on presegmentation of the audio using HTK routines.

6. Automatic Annotation

TalkBank has developed systems for automatic tagging of morphology (MOR), dependency syntax (GRASP), and phonology (Phon). Based on the morphosyntactic codes produced by MOR and GRASP, the CLAN programs can automatically compute syntactic profiles for the DSS (Lee, 1974) and IPSyn (Sagae, Lavie, & MacWhinney, 2005). MOR part-of-speech taggers have been developed for 11 languages and GRASP dependency grammars for 3 languages. These systems are described in detail in another LREC paper in this volume. In the area of phonology, the Phon program requires manual IPA transcription of non-standard child forms. However, the IPA representation for standard adult forms can be inserted automatically from the orthographic transcription. In addition, Phon provides automatic segmentation of phonological forms into syllables and syllable positions.

Apart from automatic tagging, CLAN provides methods for automatic transcript analysis. For example, the MORTABLE program provides complete counts of all grammatical morphemes in a set of transcripts, based on codes in the %mor line. The EVAL program provides package analyses of overlaps, pauses, morpheme counts and so on. We are now working to supplement these methods for automatic tagging and analysis with methods that automatically align transcripts to media at the word level and then compute a variety of fluency measures. For more careful, special purpose analyses, CLAN provides 14 analytic measures such as VOCD (Malvern, Richards, Chipere, & Purán, 2004), MLU, FREQ, and many others.

7. Transcription Conventions

To provide detailed coding methods for specific subfields, the TalkBank XML format strives to integrate best practices from each of the relevant subfields into a single unified annotation format. Unlike Partitur systems such as Anvil, EXMaRALDA, or ELAN that use time marks as the fundamental encoding framework, TalkBank XML takes the spoken word as the fundamental encoding framework. This provides results that are easy to scan across the page. Overlap alignment is also well supported through special Unicode characters that mark overlap begin and end. However, the display of overlap is not as graphic and intuitive as in the Partitur format. Because CHAT can be quickly transformed into ELAN and EXMARaLDA formats, users who need to study overlap in this way can have both views available. The only problem with this solution is that editing work done in the other systems may not be importable back to CHAT, unless the user is careful to only use CHAT conventions in the other system.

Here, we will summarize the major dimensions of CHAT transcription, coding, and annotation. The basic format involves a main line that is then supplemented by a series of dependent tiers.

1. **The main line.** This line uses a combination of eye-dialect and conventional orthography to indicate the basic spoken text. A full explication of the entire CHAT coding scheme would be outside of the scope of the current chapter. The manual of conventions is available at http://childes.psy.cmu.edu/manuals. These conventions include a wide variety of CA

codes marked through special Unicode characters entered through combinations of the F1 and F2 function keys with other characters. This system is described at http://talkbank.org/CABank/codes.html and in MacWhinney and Wagner (2010)

2. **Morphological and syntactic lines.** The MOR and GRASP programs compute these two annotation lines automatically. The forms on these lines stand in a one-to-one relation with main line forms, excluding retraces and nonwords. This alignment, which is maintained in the XML, permits a wide variety of detailed morphosyntactic analyses. We also hope to use this alignment to provide methods for writing from the XML to a formatted display of interlinear aligned morphological analysis.
3. **Phonological line**. The %pho line stands in a one-to-one relation with all words on the main line, including retraces and nonwords. This line uses standard IPA coding to represent the phonological forms of words on the main line. To represent elision processes, main line forms may be grouped for correspondence to the %pho line. The Phon program developed by Yvan Rose and colleagues (Rose, Hedlund, Byrne, Wareham, & MacWhinney, 2007; Rose & MacWhinney, in press) is able to directly import and export valid TalkBank XML.
4. **Error analysis**. In earlier versions of the system, errors were coded on a separate line. However, we have found that it is more effective to word-level code errors directly on the main line, using a system specifically elaborated for aphasic speech at http://talkbank.org/AphasiaBank/errors.doc.
5. **Gesture coding**. Although programs such as ELAN and Anvil provide powerful methods for gesture coding, we have found that it is often difficult to use these programs to obtain an intuitive understanding of gesture sequences. Simply linking a series of gesture codes to the main line in TalkBank XML is similarly inadequate. To address this need, we have developed a new method of coding through nested coding files linked to particular stretches of the main line. These coding files can be nested indefinitely, but we have found that two levels of embedding are enough for current analysis needs. Examples of these gesture coding methods can be found at http://talkbank.org/CABank/gesture.zip.
6. **Special coding lines**. CLAN and TalkBank XML also support a wide variety of additional coding lines for speech act coding, analysis of written texts, situational background, and commentary. These coding tiers are not aligned only to utterances and not to individual words.

8. Dissemination Platforms

The fundamental idea underlying the construction of TalkBank is the notion of data sharing. By pooling their hard-won data together, researchers can generate increasingly accurate and powerful answers to fundamental research questions. The CHILDES and TalkBank web sites are designed to maximize the dissemination of the data, programs, and related methods. Transcript data can be downloaded in .zip format. Media can be downloaded or played back over the web through QuickTime reference movie files. The TalkBank browser allows users to view any TalkBank transcript in the browser and listen to the corresponding audio or see the corresponding video in continuous playback mode, linked on the utterance level. We also provide methods for running CLAN analyses over the web, which we are now supplementing with analyses that use the XML database as served through the Mark Logic interface.To teach the use of the system, we have produced manuals, instructional videos and powerpoint demonstrations which we use in a wide variety of workshops internationally

9. Conclusion

Together these various TalkBank facilities provide a comprehensive, interoperable set of best practices for the coding of spoken language corpora for research in linguistics, psycholinguistics, speech technology, and related disciplines. New methods and improvements to these practices are continually in development, as we expand the database to include a fuller representation of the many forms of spoken communication.

10. References

Jefferson, G. (1984). Transcript notation. In J. Atkinson & J. Heritage (Eds.), *Structures of social interaction: Studies in conversation analysis* (pp. 134-162). Cambridge: Cambridge University Press.

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data* (pp. 165-198). Amsterdam: John Benjamins.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25*, 1286-1307.

MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprächsforschung, 2*, 1-20.

Malvern, D. D., Richards, B. J., Chipere, N., & Purán, P. (2004). *Lexical diversity and language development*. New York: Palgrave Macmillan.

Rose, Y., & MacWhinney, B. (in press). The Phon and PhonBank initiatives.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language, 37*, 705-729.

Sagae, K., Lavie, A., & MacWhinney, B. (2005). Automatic measurement of syntactic development in child language *Proceedings of the 43rd Meeting of the Association for Computational Linguistics* (pp. 197-204). Ann Arbor: ACL.

Toward the Harmonization of Metadata Practice for Spoken Languages Resources

Christopher Cieri[*], Malcah Yaeger-Dror[*●]
[*]Linguistic Data Consortium, University of Pennsylvania, [●]University of Arizona
[*]3600 Market Street, Suite 810, Philadelphia, PA 19104, USA
E-mail: ccieri@ldc.upenn.edu, malcah@email.arizona.edu

Abstract

This paper addresses issues related to the elicitation and encoding of demographic, situational and attitudinal metadata for sociolinguistic research with an eye toward standardization to facilitate data sharing. The discussion results from a series of workshops that have recently taken place at the NWAV and LSA conferences. These discussions have focused principally on the granularity of the metadata and the subset of categories that could be considered required for sociolinguistic fieldwork generally. Although a great deal of research on quantitative sociolinguists has taken place in the Unites Stated, the workshops participants actually represent research conducted in North and South America, Europe, Asian, the Middle East, Africa and Oceania. Although the paper does not attempt to consider the metadata necessary to characterize every possible speaker population, we present evidence that the methodological issues and findings apply generally to speech collections concerned with the demographics and attitudes or the speaker pools and the situations under which speech is elicited.

Keywords: metadata, sociolinguistics, standards

1. Introduction

The brief history of building digital, shareable language resources (LRs) to support language related education research and technology development is marked by numerous attempts to create and enforce standards. The motivations behind the standards are numerous. For example, standards offer the possibility of making explicit the process by which LRs are created, establishing minimum quality levels and facilitating sharing. Nevertheless, there have been instances in which the pre-mature or inappropriate promulgation or adoption of standards has lead to its own set of problems (Osborn 2010, p. 74ff, Mah, et. al. 1997) as researchers struggle to apply to their use cases standard that were not truly representative and perhaps not intended to be. To reduce the potential effort expended in developing, promoting and using proposed standards that may subsequently be found difficult to sustain, we propose that standardization is a late step in a multipart process that begins with understanding, progresses to documentation that may itself encourage consistency in practice within small groups at which point the question of standardization begins to ripen.

2. Background

The present workshop seeks to survey current initiatives in speech corpus creation with an eye toward standardization across sub-disciplines. Such standardization could permit resource sharing among researchers working in conversation and discourse analysis, sociolinguistics and dialectology among others and between those fields and others who depend upon similar kinds of data including language engineers (Popescu-Belis, Zufferey 2007). Coincidentally, the authors have been involved in a number of workshops on related themes including a series taking place at the annual NWAV (New Ways of Analyzing Variation) meetings on speech data collection, annotation and distribution including documentation and metadata description. More recently they lead a workshop funded by the U.S. National Science Foundation at the 2012 winter meeting of the Linguistics Society of America[1]. The principal topics of the latter were metadata description and related legal issues in the creation of spoken language corpora for sociolinguistics. This paper constitutes a summary of efforts within that community to begin understanding metadata encoding practice as a first step toward consistency, sharing and standardization.

3. Towards Standardization

Before metadata practice can be standardized, individual researchers must first understand their practices, the variations among them, the causes for variation, the tradeoffs of different approaches and their potential uses. In particular, researchers need to know if they can apply their metadata categories consistently, a question that is not frequently asked but must be if the goal is to adopt a standard that will be used by many independent groups with the intent of sharing corpora. Once the practice is understood it must be documented so that potential users can evaluate it and competing practices can be harmonized to permit appropriate comparisons. With adequate documentation independent researchers can decide if they want to adopt consistent practices.

4. Metadata

Within sociolinguistics, some researchers' position is that each study requires its own set of demographics. However, the ultimate consensus at the workshops was that cross community comparative corpus-based studies are only possible if there is a shared set of specific coding choices. Some of the demographic information is generally accepted within the larger sociolinguistic community: sex, birth year, years of education, and some designation of job description are fairly common

[1]http://projects.ldc.upenn.edu/NSF_Coding_Workshop_LSA/index.html

demographic fields, as are designations for where the speaker grew up, where the speaker lives at the time of the interaction, along with what years a given speaker has spent in specific regions.

5. Ethnicity

Within the American linguistics community 'ethnicity' frequently conflates three quite distinct demographic features: race, region and religion. Each of these will be discussed in turn.

5.1 Race[2]

While recent US based studies generally distinguish between black[3] or African American, Hispanic and other ethnic categories, sometimes referred to as "dominant dialect"; this is now understood to be insufficient: "black" speakers may be of Haitian, Jamaican, Dominican, or African provenance, and may not consider their primary identity as African American [henceforth AA]. Within the US, African Americans whose parents grew up in the North, can generally be distinguished from those whose parents grew up in the South. So if ethnicity choices are limited to the above three, there may be no confusion in a community where all "black" speakers are in fact African American, but in large cities much confusion could result from the failure of the coarse term to capture the three-way distinction [Blake and Shousterman 2010]. Speakers of mixed race [e.g., Purnell 2009 2010] have also been shown to differ consistently from both "white" and "black" linguistic groups within their communities.

While both the Pew Trust[4], and the Mumford Center[5] have treated Asian as a viable group, it is clear that speakers whose parents emigrated from India and Pakistan have very little in common [ethnically, regionally or religiously] with those whose parents hailed from Japan or Korea or China. It has been shown that even different Chinese groups can be distinguished from each other [Hall-Lew/Wong 2012]. It has also been shown that coding subjects for when their forebears left their country of origin reveals correlation with linguistic choice, a connection that in retrospect should not be surprising since the settlement patterns and trajectory of integration into the larger community differed for speakers arriving at different times [Sharma 2011; Wong/Hall-Lew 2012].

5.2 Regional/Linguistic Heritage

Given that Hispanic ancestry speakers are racially quite diverse within the Americas, the discussion of Hispanic heritage speakers of various racial and regional provenance is even more complicated. While the English syntax of Hispanic ancestry speakers seems to be convergent [Bonnici/Bayley 2010], the English phonology differs for even the most similar regional groups, for example Cuban, Puerto Rican and Colombian-Costeños [Bayley/Bonnici 2009] or Mexican, Texan, Californian and New Mexican Chicanos. As with the 'Asian' speakers discussed earlier, the interaction of settlement conditions with date of arrival has a strong influence on speaker variation. [Bayley/Bonnici 2009].

5.3 Religion

There have now been many studies which demonstrate that specific racial and regional heritage groups should also be divided by religion: For example, it has long been known that even in Ireland, [Milroy, 1980], Wales [Bourhis/Giles 1978] Belgium [BOURHIS, et al 1979) and the Middle East [Miller 2007, Walters 2011] different religious groups, which share the same racial and regional heritage, speak quite differently from each other, even to the extent of using different languages. For example Sunni, Shia, Copt, Maronite all speak quite differently, despite the fact that they are ethnically 'Egyptian' [Miller 2005]. Conversely, the 'New York Jews' referred to in Tannen's early work [Tannen 1981] – not to mention 'Muslims' [Miller 2007] or, for that matter 'Christians' [Wagner to appear] can belong to quite different racial and regional heritage groups, and are often linguistically quite distinct. As a result, conflating 'racial heritage' 'regional heritage' and 'religion' threatens to obscure distinctions that have been shown to be significant in numerous community studies.

Within individual studies, it is necessary that field sociolinguists determine which racial, regional and religious heritage speakers are likely to be included in their sample and prepare to control effectively for these distinctions. Unfortunately, such information is generally not coded for easy access. In fact, among the corpora currently available, even in the few cases that include protocols for eliciting speaker metadata, the protocols generally do not suggest asking these questions of speakers. Even sociolinguist interviewers, who are 'primed' by their protocol to elicit appropriate demographic information fail to probe in order to distinguish among relevant subgroups. Moreover, researchers often assume that if subjects have answered demographic questions, these answers are somehow available, despite the fact that the information may be buried in the often untranscribed interview audio. Furthermore, Lieberson (1992) shows that interviewees are not always honest or accurate in their representation of the regional, racial and religious background they belonged to during their formative years.

5.4 The Melting Pot and Multiple Identities

While in some societies, there may be little mixing among demographic or religious groups, in the US large numbers of those born since the 1970's actually belong to, and identify with, multiple demographic groups (Blake and Shousterman 2010). Coding practice needs to

[2] Any discussion of the validity of the concept or label 'race' is well beyond the scope of this paper. When we use the term here, we are merely referring to the traditional use of the term as a very broad categorization.
[3] We use the term occasionally to highlight the lack of further analysis.
[4] http://www.pewtrusts.org
[5] http://mumford1.dyndns.org/cen2000

permit the association of multiple values even for a single speaker and a single variable. A researcher may decide to give priority to the first-named 'identity', but the schema should allow for multiple listings. Mature metadata schema should also acknowledge the possibility of changing affiliation over time.

6 Encoding Demographics

Sociolinguists, historically, have assumed that the best way to do so is to incorporate relevant questions about 'ethnicity' and attitudes toward 'ethnicity' into a questionnaire executed during an interview. However, unless the interviewer has been sensitized to the fact that finer distinctions are needed, they may feel no obligation to spend time on the relevant questions. Furthermore there are no generally accepted instructions for encoding subjects' free form answers into regularized form so that future researchers can access it without having to listen to the interview in its entirety. In short a protocol for eliciting information about demographic and attitudes must be accompanies by a protocol for encoding this information into a form searchable by future scholars even if future scholars is ultimately only the same researcher returning to the data after some hiatus. Recent work has made clear that an accurate assessment of dialect change requires returning to a community 20 or more years later [Wagner 2012, in press], by which point even the original research team may no longer recall the details of an original interview. Even someone returning to a group of speakers previously studied will be under-served by a coding protocol that assumes that demographic information is adequately encapsulated in the interview itself and need not be formally coded.

7 Socioeconomic Information

Although many corpora include metadata for 'years of education', years spent in a technical school are not distinguished from those spent in what is commonly referred to as 'higher education', a fact that some research communities has already noted (Graff, pc). Moreover, multiple studies have demonstrated the usefulness of community specific scales for the importance of the 'dominant dialect' among speakers with different job descriptions, the so-called *linguistic marketplace* [Sankoff, Laberge 1978]. Even where a scale has not been devised in a given community, each speaker's occupation could be listed as well, which will permit subsequent scaling of socioeconomic and linguistic marketplace variation within a given community.

8 Politics

While it is not always possible to ask speakers about their political opinions, there have been recent articles showing that since speakers' politics strongly influence their attitudes toward their own and other groups, and their attitude toward the 'dominant dialect' of their region [Abrams et al 2011, Bourhis et al 2009, Hall-Lew et al 2010]. Some awareness of speakers' politics should be coded if possible.

9 Social Situation

Labov's early work clearly demonstrated the importance of the social situation. (See Labov 2001 for an overview.) However, the presumptions on the part of sociolinguists that every speaker is equally aware of the current social situation, that those speakers present an accurate view of the situation to interviewers and that the knowledge the community researcher has come internalize is equally obvious to outside readers are all likely to mislead. A transparent means for encoding and preserving descriptions of social situations would improve the usefulness of data sets and the ability to compare one to the other.

9.1 Interlocutor Dynamics

It has been shown that even in a straightforward interaction, the actual interlocutor is not necessarily the principal 'audience' [Bell, 1984]. At the same time, even in an interview situation, the interlocutor [interviewer] effect is pervasive [Hay/Drager 2010; Llamas et al 2009]. That said, very few corpora provide adequate descriptions of the interlocutors, including interviewers, despite the fact that this is significant in the analysis of the subject's speech.

9.2 Social Attitudes

The recent workshop at LSA as well as 4 decade's evidence from social psychological studies documented the importance of speakers' attitudes toward their own and other groups for the analysis of their speech [Giles 1973, Giles et al 1977]. In fact, the earliest studies in the social psychology of language demonstrated the variability of social attitudes even within one interaction [Giles 1973]. These factors could also be coded for, particularly if a post interaction questionnaire could be provided. While social psychologists have proposed elaborate and extensive questionnaires (Abrams et al 2009, Bourhis et al 2009, Noels 2012.) Recent work by Labov et al (2011) and by Llamas and her coworkers (Llamas 2012) have shown that the critical information can be determined with fewer questions, and with those questions presented online.

10 Broader Methodological Issues

Although our focus has thus far centered on studies conducted by sociolinguists, frequently within the United States, a number of tensions have emerged for which we have no solutions yet but which must figure into any discussion of metadata for speech corpora. We have seen that conflating 'racial heritage' 'regional heritage' and 'religion' may obscure distinctions we wish to preserve. Taken to its logical extreme, the desire for completeness and fine-granularity in elicited speaker metadata must necessarily be constrained by the limited time available for any single speaker given the other requirements of a representative speaker sample. We also see tensions between the communities with which a speaker may identify and those with which an outsider may associate the speaker. A third tension exists among the actual

methods for eliciting metadata. Checklists and multiple choice questionnaires offer the promise, perhaps misleading, of clean distinctions between metadata categories and values while ethnographic style interviews tend to recognize the inherent ambiguity of categories but exact a cost later in the analytic process of rendering textual descriptions into categories of comparison.

10 Conclusion

To reduce the effort expended in developing, promoting and using proposed standards that may subsequently be found difficult to sustain, standardization should be a late step in a process that begins with understanding, progresses to documentation that hopefully leads to consistent practice and the ultimately to standardization. The research community focusing on quantitative analysis of language variation has begun to examine its own processes and identifies a number of challenges even in the assignment of metadata for speakers and interview sessions. Among them we have noted too the use of metadata categories that are too coarse to reveal correlation already shown to exist in the literature, the conflation of multiple dimensions into a single super-category that, again, fails to capture distinctions expected to be significant. In addition we have noted a generally absence of explicit descriptions of the complete elicitation and encoding practices and, presumably as a result, a tendency to avoid entire metadata categories that other scholars have found to be revealing. By carefully enumerating the opportunities for improving metadata elicitation and providing infrastructure to support new efforts, such as template questions and coding schemata, it is the authors' hope that the community will begin to move toward consistent practice that facilitates greater data sharing and the benefits that naturally result from it.

11 Acknowledgements

We are grateful for the funding supplied by NSF BCS Grant #1144480, which made much of this work possible.

12 References

Abrams, Jessica, Valerie Berker & Howard Giles. 2009. An examination of the validity of the Subjective Vitality Questionnaire Journal of Multilingual and Multicultural Development. 30:59-72.

Bayley, R. & Lisa Bonnici. 2009. Recent research on Latinos in the United States and Canada, part 1: language maintenance and shift and English varieties. Language and Linguistics Compass 3:1300 –1313.

Baker, W and D. Bowie 2009. Religious affiliation as a correlate of linguistic behavior. PWPL 15 (Article 2) URL: repository.upenn.edu/pwpl.

Bell, A. 1984. Language Style as audience design. Language in Society 13: 145-204.

Benor, Sarah (ed) 2011. Special issue of Language & Communication 31

Blake and Shousterman 2010. Diachrony and AAE: St. Louis, Hip-Hop, and Sound Change outside of the Mainstream Journal of English Linguistics 38: 230-247

Bonnici, Lisa & R. Bayley 2010 Recent research on Latinos in the USA. Part 2: Spanish Varieties. Language and Linguistic Compass 4: 121-134.

Bourhis, Richard, G. Barrette, S.El-Geledi, R. Schmidt 2009. Acculturation Orientations and Social Relations between Immigrants and Host Community Members in California. Journal of Cross-Cultural Psychology 40: 443-467.

BOURHIS, R. Y. & GILES, H. 1977. The Language of Intergroup Distinctiveness. In H. Giles (Ed.), Language, Ethnicity and Intergroup Relations. London: Academic Press. Pp 119-135.

BOURHIS, R.Y., GILES, H., LEYENS, J.P. & TAJFEL, H. 1979. Psycholinguistic distinctiveness: Language divergence in Belgium. In H. Giles & R. St-Clair (Eds.), Language and Social Psychology, Oxford: Blackwell. Pp. 158-185.

Bowie, David 2012. Religion: elicitation and metadata. Presented at the LSA in Portland, to appear. (http://projects.ldc.upenn.edu/NSF_Coding_Workshop_LSA/index.html)

Giles, H. 1973. Accent mobility: A model and some data. Anthropological Linguistics, 15, 87–105.

Hall-Lew, Lauren 2010. Ethnicity and sociolinguistic variation in San Francisco Language and Linguistic Compass 4(7):458-72.

Hall-Lew, Lauren, Elizabeth Coppock and Rebecca L. Starr. 2010. Indexing Political Persuasion: Variation in the Iraq Vowels. American Speech, 85(1):91-102.

Hay, Jennifer and Katie Drager 2010. Stuffed toys and speech perception. Linguistics 48(4):865-892.

Hay, Jennifer, Paul Warren and Katie Drager 2010. Short-term exposure to one dialect affects processing of another. Language and Speech 53(4):447-471.

Hay, Jennifer, Katie Drager and Paul Warren 2009. Careful who you talk to: An effect of experimenter identity on the production of the NEAR/SQUARE merger in New Zealand English. Australian Journal of Linguistics 29(2):269-285.

Labov, William 2001. Principles of Linguistic Change Vol. II: Social Factors. Blackwell: Oxford.

Labov, William, Sharon Ash, Maya Ravindranath, Tracey Weldon, Maciej Baranowski and Naomi Nagy 2011. Properties of the sociolinguistic monitor. Journal of Sociolinguistics 15: 431–463

Lieberson, Stanley 1992. The enumeration of ethnic and racial groups in the census: Some devilish principles. In: J. Charest & R.Brown (eds) Challenges of measuring an ethnic world. US Govt Printing Office: Washington.

Llamas, Carmen 2012, to appear. Paper presented at the LSA Workshop. (http://projects.ldc.upenn.edu/NSF_Coding_Workshop_LSA/index.html)

Llamas, C., D. Watt, & Daniel Ezra Johnson. 2009. Linguistic accommodation and the salience of national identity markers in a border town. Journal of Language and Social Psychology 28(4). 381-407.

Mah, Carole, Julia Flander, John Lavagnino. 1997, Some Problems of TEI Markup and Early Printed Books, Computers and the Humanities 31:31–46.

CAROLE MAH, JULIA FLANDERS and JOHN LAVAGNINO

Miller, Catherine 2005. Between accommodation and resistance: Upper Egyptian migrants in Cairo. Linguistics 43(5): 903–956.
Miller, Catherine 2007. Arabic in the city: issues in dialect contact and language variation. Routledge.
Milroy, L. 1980. Language and Social Networks. Oxford: Blackwell Publishers.
Noels, Kim. 2012, to appear. Paper presented at the LSA Workshop. (http://projects.ldc.upenn.edu/NSF_Coding_Workshop_LSA/index.html)
Osborn, Don, 2010, African Languages in a Digital Age: Challenges and Opportunities for Indigenous Language Computing, Ottawa, International Development Research Center.
Popescu-Belis, Andrea, Sandrine Zufferer, 2007, Contrasting the Automatic Identification of Two Discourse Markers in Multiparty Dialogues, in Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 10–17, Antwerp, Association for Computational Linguistics.
Purnell T. 2009. Convergence and contact in Milwaukee: Evidence from select African American and white vowel space features. Journal of Language and Social Psychology 28(4): 408-427
Purnell, Thomas C. 2010. The Vowel Phonology of Urban Southeastern Wisconsin. In Yaeger-Dror and Thomas, eds, AAE speakers and their participation in local sound changes: A comparative study. Publications of American Dialect Society #94. Raleigh: Duke University Press. 191-217.
Sankoff, D. and Suzanne Laberge 1978. The linguistic market and the statistical explanation of variability. In D. Sankoff (ed.), Linguistic Variation: Models & Methods. NY: Academic Press. 239-50.
Sharma, Devyani 2011. Style repertoire and social change in British Asian English. Journal of Sociolinguistics 15: 464-492.
Suarez, Eva Maria 2010. Dominican identity and lg choice in the Puerto Rican diaspora. Nwav39.
Tannen, D. 1981. New York Jewish conversational style. IJSL 30:133-149.
Wagner, Suzanne Evans. 2012, in press. Age grading in sociolinguistic theory. Language and Linguistics Compass.
Wagner, Suzanne (to appear) Linguistic correlates of Irish-American and Italian-American ethnicity in high school and beyond. In Yaeger-Dror & Guy. PADS #97. Duke University Press: Raleigh.
Walters, Keith 2011. Gendering French in Tunisia: language ideologies and nationalism
IJSL 211: 83-111.

Best practices in the design, creation and dissemination of speech corpora at The Language Archive

Sebastian Drude, Daan Broeder, Peter Wittenburg, Han Sloetjes
Max Planck Institute for Psycholinguistics,
The Language Archive,
P.O. Box 310, 6500 AH Nijmegen, The Netherlands
E-mail: {Sebastian.Drude, Daan.Broeder, Peter.Wittenburg, Han.Sloetjes}@mpi.nl

Abstract

In the last 15 years, the Technical Group (now: "The Language Archive", TLA) at the Max Planck Institute for Psycholinguistics (MPI) has been engaged in building corpora of natural speech and making them available for further research. The MPI has set standards with respect to archiving such resources, and has developed tools that are now widely used, or serve as a reference for good practice. We cover here core aspects of corpus design, annotation, metadata and data dissemination of the corpora hosted at TLA.

Keywords: annotation software, language documentation, speech corpora

1. Introduction

This paper summarizes the central facts concerning speech corpora at the Max Planck Institute for Psycholinguistics, now under the responsibility of a new unit called "The Language Archive" (TLA[1]). This unit, besides maintaining the archive proper, also develops software relevant for creating, archiving and using language resources, and is involved in larger infrastructure projects aiming at integrating resources and making them reliably available. The TLA team is, however, not responsible for designing, collecting and creating the corpora, which is done by researchers. Therefore this paper covers mostly technical aspects or reports on other aspects from an indirect and technical perspective. Most facts reported here are the (preliminary) result of on-going and long-term investments and developments. As such, they are mostly not new unpublished results, but still give a good overview over many of the solutions applied in TLA for relevant questions about speech corpora.

The speech corpora at The Language Archive at the Max Planck Institute for Psycholinguistics (MPI-PL) have so far mainly come from two disciplines not mentioned in the call for papers: language acquisition and linguistic fieldwork on small languages worldwide.

Due to their provenience, these corpora differ considerably from usual corpora applied in speech technology or other areas traditionally concerned with linguistic corpora.

The language acquisition corpora at the MPI have mostly been annotated using a particular annotation format, CHAT (MacWhinney 2000), developed and used since the early 1980ies in the CHILDES project and database. Although applicable to other areas of research, CHAT is tailored to reveal emergent grammatical properties and the adaptive solution of communicative needs by children or second language learners. CHAT can be considered an excellent standard for annotating acquisition corpora, and there are powerful statistical tools for this corpus available.

However, there are other areas of research that deal with audio and video data that are to be annotated, as for instance corpora of natural or elicited speech from the many native languages around the world. As at other centres, also at the MPI-PL such corpora have first been collected in field-research for purposes of description and comparison. Since the 1990ies, however, when the threat of extinction of the overwhelming part of linguistic diversity, it became obvious that the documentation of endangered and other understudied languages is an important scientific goal in its own right, and research programs such as DOBES were established. This even gave rise to a new sub-discipline of linguistics which is primarily concerned with the building of multimedia corpora of speech, viz. "Language Documentation" (sometimes "documentary linguistics").

Besides tools and web-services for archiving language data, the technical group at the MPI-PL (now TLA) is engaged in developing a multi-purpose annotation tool for speech data, ELAN (Wittenburg et.al. 2006). This tool was first applied in the documentation of endangered languages and other linguistic field research, but then proved to be useful in the annotation of sign language data and generally in the area of multimodal research. The data available in the ELAN annotation format (EAF) as generated by the ELAN tool is suited for machine processing (XML Schema based), and thus it is now at the core of most developments in TLA and well supported in the TLA archive software (e.g. TROVA, ANNEX).

The current contribution focuses on speech corpora as archived at TLA, in particular corpora as the result of Language Documentation. We try to address as many topics relevant for speech corpora as possible from an archive's and software development group point of view.

2. Corpus Design and Curation

Considering the design, but also the management and curation of (speech) corpora, there is an important

[1] All underlined terms refer to entries in the references.

difference between corpora that can be seen as 'finished', i.e. static, and others that are 'growing', i.e. where there is a more or less continuous stream of material being added, often over or after several years. A good example of the first kind is the Corpus Spoken Dutch (CGN) for which the exploitation and management environment, COREX, was developed by the TLA group. In such a corpus project, where the data collection can be carefully planned and executed in a relatively short time, a large degree of coherence and consistency can be achieved with respect to the data and metadata formats allowing for efficient exploitation procedures and tool development. The corpus was compiled so as to have a representative sample of the spoken Dutch with many different monological and dialogical text types (CGN Design).

Another example housed at TLA is the Dutch Bilingualism Database (DBD), which is a curation project combining several older Second Language Acquisition (SLA) corpora, integrating them in a new overall corpus structure using coherent metadata descriptions. Obviously, the design of this corpus follows a rather narrow focus on SLA research. Unfortunately, no such coherence was possible also at the annotation level, and hence its usability may be limited.

At the other side of the spectrum there are the language documentation projects (such as in the DOBES program funded by the Volkswagen Foundation) where data collection takes place over longer periods, according to different procedures and without any agreed coding of linguistic phenomena. In the case of such corpora, it is a much bigger challenge to achieve any kind of (semantic) interoperability, e.g. for searching specific events over all corpora. Still, the design with respect to text types and genres is driven by similar criteria as that for the Corpus of Spoken Dutch – the design of these corpora is essentially multi-purpose, because many of the languages will most probably not be spoken in a few generations, so that the documentation is the major (or even only) source of data for future studies, also for neighbouring fields such as ethnomusicology, -botany, -history, and anthropology in general. As a result, again ideally as many as different types of communication events are recorded, from traditional texts (at the core of importance in particular for the speech communities, as this is the most valuable knowledge that often threatens to be lost first with the aging and death of the older generation) via explanations, descriptions, spontaneous stories to natural conversation. Differently from the Corpus of Spoken Dutch, however, the content and its relevance for future studies in other than linguistic research are an important criterion for the selection of recordings. Still, it is one of the major objectives for the data to provide the basis for an extensive description (grammar), and for typological studies. The crucial point about fieldwork corpora is that the linguistic system of the languages being documented is often not well understood, so that details of the analysis underlying the annotation may change and improve over time.

In all cases it is important to notice that the design and creation of the corpora, including the creation of the content of metadata, is done by scientists, not by members of TLA. TLA is responsible for the technical aspects of these corpora, such as data (including metadata) formats and proper archiving.

3. Annotation (including Transcription)

One emergent technical data format for annotated speech is EAF (see above), and this is by now the default for most annotation in corpora hosted at TLA. ELAN does not only allow annotating audio and video recordings, but it also allows to use as many "tiers" (an annotation container without predefined disposition, representing a layer or type of annotation) as necessary for any given speaker, and to relate the data between these tiers in technically practical ways, allowing to organize annotations in hierarchical tier structures.

Thus, ELAN is a generic annotation tool that is applied in various types of research. There is no built-in tier type for speech, phonetic transcription, gesture or whatever other type of aspects of communicative events could be annotated. This renders a flexible transcription environment to which easily and at will tiers can be added.

Being one of the first tools of its kind, ELAN enabled the deeper analysis of sign language. It also fostered the study of "paralinguistic" phenomena such as gestures, which increasingly turn out to be intrinsically related with spoken language, making their study indispensable for the understanding of the latter. Today, ELAN is widely adopted even outside these domains and even outside linguistics.

Under (linguistic) *annotation of* data we understand any symbolic representation of properties of the (speech) event represented in the primary data.[2] By this definition, a transcription of the original utterances (depending on the purpose of the corpus in phonetic, phonological or, most often, orthographical form) in the original language is also annotation, and it is indeed the most basic and most frequent type of annotation in the corpora at TLA.

In the case of the language documentation corpora, the material is usually only interpretable and thus useful to users (other than members of the speech community) if also a translation into a major language is provided, constituting a second important type of annotation (representing semantic properties of the underlying speech events). Together, a transcription and at least one translation, possibly with one further layer of notes or comments, represent what in language documentation can be defined as *basic annotation* – this is indeed the minimum required annotation in the case of the DOBES program.

There are many other possible levels of linguistic, paralinguistic and non-linguistic annotation. One attempt at systematizing the linguistic levels of annotation has

[2] In linguistics, *primary data* are direct representations or results of a *speech event*, for instance a written text or, in particular, an audio/video recording of a speech event.

been made in the pilot phase of DOBES (Lieb and Drude 2001). But only recently the need for standards in categorizing and referring to different levels of annotation has come to the attention of documentary linguists and technicians. The DOBES corpora do not have any agreed format or naming for the types of linguistic levels to be represented in the various annotation layers, and this presents now a major challenge for efforts to make the different corpora comparable and interoperable.

True, a very popular format of annotation of the original recordings in the language documentation corpora is *basic glossing*;[3] in particular the interlinear glosses as formalized by the "Leipzig Glossing Rules", by now an established (but still developing) standard in language description and typology. These interlinear glosses are often created using the Toolbox program. But even when in principle aiming at following the Leipzig Glossing Rules, details of basic glossing in language documentation corpora usually vary quite a bit from corpus to corpus, in particular with respect to the abbreviations applied for abstract functional units. The ISOcat data category repository (Kemp-Snijders et.al., 2008) provides a means to clarify the nature of tiers and the meaning of individual glosses. The ISOcat Data Category Registry (ISO 12620) defines linguistic concepts in a way compliant with the ISO/IEC 11179 family of standards. It is hosted at and developed by the TLA. Thus, one can refer to a certain concept independently from its concrete label or abbreviation – "noun", "N", "Subs(antive)" etc. can all refer to the same data category "Noun" – or to different categories which are connected by a relation of "is-roughly-equivalent-to". Such relationships can be established between different ISOcat data categories with the new RELcat registry (Windhouwer 2012). RELcat is likewise developed by the TLA and currently in the alpha phase.

In ELAN both, tiers and annotations, can refer to a data category. On the tier level this reference indicates the more general type of annotations of that tier, e.g. "part-of-speech", on the annotation level it is the more specific category, e.g. "verb". The goal is to achieve interoperability between different annotations in different corpora despite the broad variation in annotation tiers, conventions and labels we observe in fieldwork and descriptive linguistics.

In addition to basic glossing, some (sub)corpora may have some advanced glossing – which covers one or several of the other linguistic levels, from phonetic via phonological, morphological, syntactic, semantic to pragmatic, or even paralinguistic and non-linguistic levels. Advanced glossings can include, for instance, a phonetic transcription and annotation of the intonation contour, or of the syntactic structure, of grammatical relations, etc. For instance, an emergent standard in the DOBES program is the GRAID annotation (Haig & Schnell 2011). Any kind of manual annotation, from segmenting to coding different linguistic and other information, is the most time expensive step of many workflows. TLA has recently begun the development of new annotation functionalities of ELAN that comprise automatic audio & video recognition and semi-automatic annotation, so that modules developed in an NLP context or at the MPI can be "plugged" into ELAN. Such modules include morpheme-splitters, Toolbox/FLEX-like annotation support and traditional POS-taggers etc.

[3] Under *basic glossing* we understand annotation that, in addition to basic annotation, also provides information on individual units (usually morphs, sometimes words), such as typically an individual gloss (indication of meaning or function) for each morph / word, and perhaps also categorical information such as a part-of-speech tag (or its equivalents on the morphological level).

4. Metadata & Data Management

With respect to long-term availability ("archiving") of speech data, TLA has also had a pioneering role. The solutions developed at the MPI are now one important basis for the construction of large infrastructures for digital language research data (for instance, in the CLARIN project).

In the early 2000s, the technical group at the MPI started developing IMDI, a XML-based metadata standard which is geared to observational multimedia data such as language acquisition and field work recordings. This standard was developed in close cooperation with researchers active in the early years of the DOBES research programme. Metadata are then stored in separate XML files side by side with the bundle of resources (multimedia files with recordings etc., annotation files in different formats such as Shoebox/Toolbox-text files, Transcriber and EAF XML-based files, and a few other) they describe. These resources are linked to the metadata file by pointers in the metadata files – today, persistent identifiers (handles) are used in order to guarantee reliable access even if the location of files should change. The bundle of a metadata file together with the resources that are referenced in it and described by it are called a "session" – it may contain just one video or audio or text file, but also possibly dozens of closely related files, and, for technical reasons, different versions / encodings etc. of the 'same' file.

The IMDI metadata schema contains several dedicated data fields for describing speakers and communicative events. This is the major point in which IMDI and, say, OLAC metadata diverge. The IMDI metadata schema has specializations for general speech corpora as CGN and other TLA corpora such as DBD.

A virtual hierarchy or grouping of sessions in a tree-like structure is achieved by a second type of IMDI metadata files, each representing a node in the tree and pointing to other corpus node and / or session IMDI files. In this way, the same set of sessions can be organized by different criteria in parallel.

The advantage of such a system is that all resources, including metadata, are stored as separate files in the file system, without being stored inside some database or other encapsulated file. For quick access and administration a database is used at TLA, too, but this

database can be reconstructed at any time by crawling the metadata tree. The IMDI metadata tree can be accessed by a local standalone viewer and by an online tool, the IMDI browser. Additional online tools allow to integrate new sessions and to manage the data in the online archive (LAMUS, AMS, Broeder et.al. 2006).

In the last years, due to and based on TLA's experience in organizing the Archive at the MPI-PL, the technical group and now TLA is prominently involved in European infrastructure projects, in particular in CLARIN, with links of cooperation to DARIAH. These infrastructures have a much wider range than TLA's focus on speech corpora. TLA is currently implementing a transfer to CMDI, a metadata schema based on a component structure, in order to integrate its resources into the wider CLARIN context and to cope with the appropriate description of a growing amount of experimental and other data, such as eye-tracking or even brain images and in the near future genetic data. The ARBIL tool is being developed for creating and editing CMDI metadata, and is evolving into the basis for a general virtual research environment for data management in the context of CLARIN and beyond.

5. Data Preservation and Dissemination and Legal and Ethical Issues

Integrating different data centres has as one aspect that data can be replicated from one centre to the other, improving the safety of the data. For DOBES data, currently six copies are created automatically at three locations. Also, selected data collections are being returned to the regions where they were recorded. The Max-Planck-Gesellschaft gave a guarantee for preserving the bit-stream for 50 years.

The goal of TLA, however, is not just "archiving" in a traditional sense where the ideal is to preserve the archived material faithfully but to touch it as rarely as possible. In the case of digital archives, or rather data centres, the opposite is the case: the more often the material is accessed and used, the better. That implies that providing not just reliable but also easy and useful access is an important goal of any data centre. Making research data interoperable and integrating it in networks that allow cross-corpora searches and complex analysis is only one aspect of it, which can be done by the data centres. Applying standards and making the data more appealing and useful, for instance by providing complete metadata, in turn, can only be done by the researchers. In fact, most aspects of enhancing resources and data centres can only be done in close cooperation between both partners. An important aspect of a fruitful relation between researcher and data centre is trust. The centres, and this holds for TLA, must not have their own agenda with the data, and they must respect necessary restrictions to the ideal to free access to the resources.

The legal questions are always intricate when human subjects are recorded. For language corpora conditions vary: there are corpora where requirements for access are clear and the research community is well served, e.g. the language acquisition data of the CHILDES system. For others, also at TLA, the situation is much less clear and access can only be permitted on an individual basis. Generally, in the case of linguistic observational data the privacy of the human subjects who are recorded needs to be taken into account. The corpora from fieldwork gain even more intricate legal and ethical dimensions due to the extreme inequality in access to resources and information between the researchers and the speakers, the impossibility of anonymizing the speakers which often live in very small communities, and the international setting of the projects.

Simple answers cannot be given, but attempts at creating clear and fair conditions of use, and hence, ultimately trust between the different stakeholders, can be made. In the DOBES program a code of conduct was created (Max Planck Institute for Psycholinguistics 2005). It excludes commercial use and other uses that are disrespectful to the culture of the respective speech communities.

Handling legal and ethical issues at a responsible level is a serious challenge. For instance, for culture-specific or other reasons, members of the speech communities may withdraw access permissions to certain material even though it was granted at a previous time. On the other hand, after years, necessary restrictions can be withdrawn by the depositor or by representatives of the speaker community. Opening the data as far as legally and ethically possible is generally a requirement, especially when the research was financed with public money. However, scientists and funding agencies or different community members may have different positions. To cope with all kinds of unexpected events a Linguistic Advisory Board consisting of highly respected field researchers was established that can be called upon by the archive to help solve potential difficult questions.

Over the years, four levels of access privileges were agreed upon. These can be set with the AMS tool, using the standard hierarchical organization of the sessions – for instance, below a certain node in the 'tree', free access can be granted to all audio, but not to the video material, or access to annotation can be limited to a certain user group while primary data are freely accessible. The four levels are:

Level 1: Material under this level is directly accessible via the internet;

Level 2: Material at this level requires that users register and accept the Code of Conduct;

Level 3: At this level, access is only granted to users who apply to the responsible researcher (or persons specified by them) and who make their usage intentions explicit;

Level 4: Material at this level will be completely closed, except for the researcher and (some or all) members of the speech communities.

Access level specifications for archived resources may change over time for various reasons, e.g. resources could be opened up a certain number of years after a speaker has passed away, or access restrictions might be loosened after a PhD candidate in a documentation project has

finished their thesis.

The number of external people who requested access to 'level 3' resources over recent years was not that high. We need to see in the future whether the regulations that are currently in place can and should be maintained as explained. Access regulations remain a highly sensitive area, where the technical possibilities opened up by using web-based technologies need to be carefully balanced against the ethical and legal responsibilities which archivists and depositors have towards the speech communities. Despite almost 10 years of on-going discussions and debate, no simple solution to this problem has yet been found.

6. Conclusion

Speech corpora involve many intricate questions on various levels, from corpus design via annotation, metadata and organization, to data preservation and dissemination, and include legal and ethical issues. This paper addressed some of them from the technical point of view of an archive and software development team also engaged in building a federated infrastructure for language resources.

7. Acknowledgements

The Language Archive is a unit of the Max Planck Institute for Psycholinguistics (MPI-PL), funded by the Max Planck Society (MPG), the Berlin-Brandenburg Academy of Sciences (BBAW) and the Royal Netherlands Academy of Sciences (KNAW). The DOBES project is funded by the Volkswagen Foundation. The CLARIN pilot phase was financed by the European Commission, the national CLARIN projects are funded by the individual member states.

8. References

AMS: Archive Management System. Online at tla.mpi.nl/tools/tla-tools/ams.

ARBIL: Metada Editor for IMDI and CMDI. Online at: tla.mpi.nl/tools/tla-tools/arbil.

Broeder, D., Claus, A., Offenga, F., Skiba, R., Trilsbeek, P., & Wittenburg, P. (2006). LAMUS: The Language Archive Management and Upload System. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006) (pp. 2291-2294).

CGN Design: Page "Corpusopbouw" on the pages of the Corpus Gesproken Nederlands (CGN). (2006). Online: tst.inl.nl/cgndocs/doc_Dutch/topics/design/index.htm. Last visited 2.4.2012.

BBAW: Berlin-Brandenburgische Akademie der Wissenschaften. www.bbaw.de. Last visited 23.3.2012.

CLARIN: Common Language Resources and Technology Infrastructure. www.clarin.eu. Last visited 23.3.2012.

DARIAH: Digital Research Infrastructure for the Arts and Humanities. Online at: www.dariah.eu.

DOBES: Dokumentation Bedrohter Sprachen. www.mpi.nl/dobes. Last visited 23.3.2012.

Haug, G. and Schnell, S. (2011). Annotations using GRAID (Grammatical Relations and Animacy in Discourse). Introduction and guidelines for annotators. Version 6.0. Online at www.linguistik.uni-kiel.de/GRAID_manual6.0_08sept.pdf. Last visited 4.4.2012.

ISOcat: ISO Data Category Registry: www.isocat.org

Kemps-Snijders, M., Windhouwer, M. A., Wittenburg, P., Wright, S.E. (2008). ISOcat: Corralling Data Categories in the Wild. In European Language Resources Association (ELRA) (ed), *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Morocco, May 28-30, 2008.

KNAW: Koninklijke Nederlandse Akademie van Wetenschappen. www.knaw.nl. Last visited 23.3.2012.

LAMUS: Language Archive Management and Upload System. Online at: tla.mpi.nl/tools/tla-tools/lamus.

Lieb, H, Drude, S. (2001). Advanced Glossing – A Language Documentation Format. *DOBES Working Papers* 1. Online at www.mpi.nl/DOBES/documents/Advanced-Glossing1.pdf. Last visited 2012-04-02.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates. Newest version online at: childes.psy.cmu.edu/manuals/chat.pdf. Last visited 2012-04-01. (CHAT manual)

MPG: Max-Planck-Gesellschaft. www.mpg.de. Last visited 23.3.2012.

Max Planck Institute for Psycholinguistics (2005). DOBES Code of Conduct. Compiled by Peter Wittenburg with the assistance of several experts. www.mpi.nl/DOBES/ethical_legal_aspects/DOBES-coc-v2.pdf, updated on 6/01/2006. Last visited 2012-04-01.

MPI-PL: Max Planck Institute for Psycholinguistics. www.mpi.nl/. Last visited 23.3.2012.

RELcat: a Relation Registry for linguistics concepts: http://lux13.mpi.nl/relcat/site/index.html

TLA: The Language Archive at the Max-Planck Institute for Psycholinguistics. tla.mpi.nl. Last visited 23.3.2012.

Windhouwer, M.A. (2012). RELcat: a Relation Registry for ISOcat data categories. Accepted for a poster and demonstration at the *Eigth International Conference on Language Resources and Evaluation LREC 2012*, Istambul, May 2012.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In: *Proceedings of LREC 2006*, Fifth International Conference on Language Resources and Evaluation.