Sound to Sense corpora

Annett Jorschick
Faculty for Linguistics and Literary Sciences
University of Bielefeld
Universitätsstraße 25
D-33615 Bielefeld
annett.jorschick@uni-bielefeld.de

Content

1 Introduction

This manual provides an overview of the corpora collected and provided to the S2S community by S2S members. Moreover, the (minimal) requirements on data sharing are discussed and a collection of the experiences of corpus builders is given. Due to its structure, it should also provide a framework and guidance for the construction of new corpora.

Although the overall goal of S2S is the analysis of PD (phonetic detail), the methods vary widely (for example see Annex II of the MTR contract). These different methods make different demands on corpora. Differences can be found in the size of corpora, the selected speech styles, and the tasks to elicit speech, the quality of the recordings, the annotations and far more. This large variation complicates the comparison of corpora along languages and/or speech styles and the integration of databases. However, the existing corpora are valuable recourses for new research, especially because corpus construction is costly and time demanding. Thus, it is important to know which data are already available and how to access them. This manual provides this information for S2S corpora as well as guidelines for the construction of new corpora.

The manual is structured as follows: In chapter two, four styles of speech are defined. These are read speech, elicited speech, semi-spontaneous monologue and conversational speech. The tasks that were used for the construction of S2S corpora are classified along this speech styles and described together with their advantages and disadvantages. The next section gives a summary of the corpora provided by S2S members or constructed within the S2S network. The second part of chapter three summarizes this information in table format to facilitate the comparison between different corpora. Section four lists some thoughts on ethics of sharing data, property rights of the corpora and sharing regularities regarding the provision of metadata, transcription rules, file names, and directory structure. Chapter five provides summaries of selected transcription and database management and query tools.

2 Speech styles and tasks

The speech style of a corpus is strongly interconnected to the task used to collect the data. This chapter provides an overview of the tasks used to elicit a certain style of speech and a short description of the research goals that were investigated using these styles and tasks. In lack of better terms to categorize speech styles, I chose: read, elicited, semi-spontaneous monologue, and conversational. All terms contain a range of tasks and thus a range of speech styles and boundaries between categories often overlap.

All four main sections of this chapter start with a description of a speech style, its advantages and disadvantages, and the research aims it is mainly used for. Subsections contain information on the tasks that result in this speech style, its special advantages and disadvantages, methods that were used to overcome the disadvantages, and the corpora in this manual (see chapter 3) that use the given task and with their specific research aims.

2.1 Read speech

Read speech comprises reading lists of syllables or phones to sophisticated texts, broadcast news, audio books or read dialogue. The boundary between read lists of sentences and read text is quite arbitrary. I will use the term read text if the sentences are connected due to semantics.

The advantage of read speech is the excellent control of the materials even up to prosodic structure in some sense. For example, the investigator can control the number of occurrences of targets, their context and their position in an utterance, which implies at least some control over position in a prosodic unit.

The primary disadvantage is that read speech is carefully pronounced and unnatural. Especially prosodic variables such as intonation might considerably deviate from conversational speech. Moreover, it has to be considered that texts and lists always reflect the wording of the author.

Read speech corpora are often used in ASR research, because large amounts of similar data are needed to train models. Carefully prepared linguistically rich sentences are also used to create materials for the investigation of the acoustic-phonetic realization of segments in various linguistically relevant contexts. The high control of the utterances that will be produced allows the investigation of various linguistic phenomena with reasonable effort and costs. To get enough tokens of a certain kind to study (psycho-) linguistic differences, extremely large corpora of spontaneous speech would be necessary.

To minimize the difference between natural and read speech a number of different methods were used, such as practice, time pressure, and reading dialogue.

2.1.1 Lists

Several corpora contain speech from read lists of syllables, numbers, words, sentences or similar structured material. Lists are easy to prepare and allow an extremely high degree of control of the materials. The wording is available which facilitates orthographic transcriptions and reduces costs. Nevertheless, lists elicit an unnatural speech style and may result in various artefacts, such as effects due to word order on the list and other. Moreover, the use of the materials and recordings is often strongly restricted to the purpose it was constructed.

Many corpora consisting of other speech styles for the main part of the data collection additionally contain read lists of selected utterances or words. This allows a comparison between different speech styles of the same participants later on. This was done for the Ernestus corpus of spontaneous speech (see section 3.1.5), where the participants read lists of monosyllabic words and non-words covering all Dutch monophtongs at the end of the recordings.

Similarly, the Word Segmentation corpus (see section 3.1.17) includes read versions of selected sentences from a conversation task to be able to compare the same utterances in read and conversational speech.

Parts of the CLIPS (see section 3.1.1) corpus consist of lists of linguistically sophisticated sentences that were read from professional and non-professional speakers.

In the Consonant Challenge Corpus, the participants read lists of vowel – consonant- vowel sequences (see section 3.1.2 for details). The data were used to investigate the differences between human and automatic speech recognition (Cooke & Scharenborg, 2008).

The English and French Assimilation (see sections 3.1.4 and 3.1.7) corpus are made up of well-prepared sentences that allow assimilation of alveolar and postalveolar fricatives in both languages.

The Norwegian corpora FonDat 1 to 3 (see section 3.1.6) is composed of large amounts of read sentences that were selected from newspaper due to readability purposes. Here the participants were asked to read in different styles, such as expressive or everyday manner.

GRID (Cooke et al., 2006; see section 3.1.8) is an audiovisual corpus consisting of read sentences like "put red at G9 now" which allows the investigation of the coherence between audio and visual information.

2.1.2 Prepared texts

Similar to lists, read text is often used to create corpora. The advantage of texts over lists is that the coherence of the text may result in higher naturalness and fluency. Furthermore, the influence of larger prosodic units (such as breath group or phrase boundaries) can be investigated. Similar to lists, self-prepared texts allow large control over material (although participants do not always

read the texts in the suggested way) and the availability of the manuscripts facilitate orthographic transcription. To obtain a semantic coherence of the text, filling phrases may be necessary that lengthen the text and complicate its writing. The value of the texts for research might be restricted to the purposes it was constructed for. Moreover, read speech is not natural. Due to practise, time pressure, reading in a dialogue structure and/or the instruction, the type of elicited speech can be manipulated.

In the Emory story corpus (see section 3.1.3) participants read a prose text in which the topic structure was varied allowing investigating the phonetic implementation of those variations.

Prague Phonetic corpus (see section 3.1.12) contains short phonetically sophisticated texts and a read piece of prose. Both are read by the same speakers. However, the prose is more demanding to read as the prepared text, naturally varying the reading style and allowing comparisons of the same speakers.

2.1.3 Broadcast news / audio books

Read speech can also originate from radio or TV recordings or audio books. The advantages of those recordings are that they are easily available and result in large amounts of data at relatively low costs. In some cases (e.g. cooperation with the broadcast station), the manuscripts may be available which reduces effort of orthographic transcription. The speech from those recordings is normally carefully pronounced and originates from professional speakers. The topic of the broadcast restricts the wording and has to be taken into account for investigations of frequencies in language use and similar research topics.

Recordings of broadcasts news are used in the CLIPS corpus (see section 3.1.1) and the Rundkast corpus (see section 3.1.13).

2.2 Elicited experimental speech

Elicited speech covers a wide area of speech styles that do not fit into other categories. One part is experimental elicited speech for which various paradigms were developed. Most of these paradigms are used to obtain spontaneous monologues and dialogues and thus are discussed in sections 2.3 and 2.4. However, there exist also experimental tasks used to get specific utterances or words from a single participant. Experiments are relatively easy to construct and allow full control of materials. The orthographic transcription is fast and cheap since the elicited speech often consists of single word utterances. There is a kind of shared assumption that experimentally elicited speech is slightly more natural than read speech. A disadvantage especially is that useable material is often limited to pictures, which also restrict the usable vocabulary. Longer stretches of experimental controlled speech is hard to elicit.

Experimentally elicited speech is especially useful for research on language acquisition or speech impairments, for example, when children are too young to read.

Furthermore, I assigned stage dialogue and recited text to this category.

2.2.1 Stage Dialogue / Recitation

Stage dialogue may be used as a resource for speech and might be due to its overlearning and repeated practicing more natural than reading. Moreover, material is easy to get (e.g. from TV). The texts are often available simplifying the orthographic transcription and reducing its costs. However, the speech style is not natural and often especially expressive. Moreover, control of materials is limited if real speech on stage is used. Long training of the participants may be necessary if self-constructed materials is used which will result in higher effort.

Stage dialog and recitation are listed here for the sake of completeness. No S2S corpus used this task to collect speech.

2.2.2 Repetition

In repetition tasks, the participants repeat the utterances pronounced by the experimenter. Sequences can range from words to sentences or even longer units. Material is easy to construct and every utterance can be used (even non-words or non-speech sounds) resulting in an extraordinary high control of materials. However, the setting is susceptible to artefacts such as copying the speech style of the experimenter entirely: pronunciation, speech rate, accent, intonation, pitch, dialect and other. Thus, the usefulness for the investigation of normal speech is limited but the task is very useful for the investigation of patients with impaired speech and language acquisition of children.

The Speech Impairment Corpus (see section 3.1.15) uses repetition for the investigation of speech impairments and the development of standards for diagnosis of aphasia.

2.2.3 Picture naming

Picture naming is a common psycholinguistic paradigm to investigate lexical effects on latencies or error rates. It also is useful in research on speech impairment and language acquisition as no reading skills are required. Deviating from repetition, participants cannot copy the speech of the experimenter. A limitation of this task is that only figurative material can be used.

The Speech Impairment Corpus (see section 3.1.15) uses picture naming for the investigation of speech impairments and the development standards for its diagnosis.

2.3 Semi-spontaneous monologue

Semi-spontaneous monologue refers to speech from a single speaker that is elicited by a task. Many of the tasks listed in section 2.4 can be used to elicit semi-spontaneous monologue (e.g. map task). Often this is done with the instruction to talk to another speaker (by pretending he is in the next room or will listen to the recordings later). Monologues are also available from broadcast that is described in 2.4.1.

Semi-spontaneous speech can already be very natural although often less casual than speech from a conversation. Since there is only one person speaking at the time of the recordings, the resulting audio files have a better quality than those of conversational speech, in which speech overlaps and background noise frequently occur. Speech from semi-spontaneous monologue reflects the wording of the speaker. This implements little control over the materials and utterances. Orthographic transcriptions are as complex and expensive as for conversational speech. Another disadvantage of semi-spontaneous speech is the recording situation itself and the knowledge of the participants of being recorded.

This type of speech is frequently used and of interest for various linguistic research especially for the investigation of pronunciation.

2.3.1 Narratives

Telling a story is one way to elicit a semi-spontaneous speech style. To gain some control of the used utterances, participants might get a list of words or phrases they have to use. Alternatively, pictures or cartoon strips may be used. The elicited speech is relatively natural and getting the data needs less effort than conversational speech. However, orthographic transcription of such data is as costly as conversational speech. The specification of the topic restricts the value of those recordings for word frequency measures or similar research topics. Giving the participants some time for preparation and practicing before the recordings may result in higher fluency of the story but also in a more careful wording.

The Prague Phonetic Corpus (see section 3.1.12) uses this task to elicit semi-spontaneous speech. The participants get a cartoon strip (see Appendix A) and some time to prepare a story that goes with the pictures. After preparation, they are recorded.

2.3.2 Interview

Interviews are another way to elicit spontaneous speech. Although at least two people are involved, only the speech of the interviewee can be considered as spontaneous. To get a more natural speech style, the interview should contain little structure and be open to topic shifts. According to Labov (1972, see also Ernestus, 2000), topics which recreate already felt emotions

can result in a casual speech style. Asking for opinions on topics and/or contrasting uttered opinions, might also result in a more casual speech style.

The Buckeye corpus (Pitt et al., 2007, http://vic.psy.ohio-state.edu/) is a large data collection that used interviews to elicit spontaneous speech in American English.

2.4 Conversational speech

Conversational speech is the most natural speech style and of interest for all research areas connected to speech. Indeed, it is difficult to elicit natural spontaneous speech in laboratory surroundings, especially if the participants are aware of being recorded. Some remarks to elicit natural speech are given by Labov (1972) and Ernestus (2000). This includes selecting friends as participants, choosing topics with emotional content and encouraging speakers to talk before and between tasks.

Different tasks were invented to elicit conversational speech. Some of these tasks will be introduced here. As mentioned earlier, several of these tasks can be used to elicit semi-spontaneous speech of a single speaker by pretending to speak to another.

2.4.1 Broadcast: radio and TV interviews and shows

The easiest way to get conversational speech is recordings from radio or TV shows and interviews. For the recordings just a soundcard and a computer with sufficient storage is needed and reduces costs. Even large quantities of data are easily accessible which is especially important for computational speech research. Nevertheless, recordings from broadcast have various disadvantages. First, the recordings need authorisation of the broadcast station before publishing results. Second, the quality of the recordings might be poor due to the recording method or post-processing of the sounds on the part of the broadcast station before sending it. Third, speakers are aware of being recorded certainly influencing their speech style. Fourth, the speakers are generally professional speakers.

Broadcasts are used by the CLIPS (see section 3.1.1), Rundkast (see section 3.1.13), and vegtalk (see section 3.1.16) corpora. The Rundkast project directly received the recordings from the broadcast station resulting in a high quality of the sounds.

2.4.2 Role play

In role-play, participants are asked to act according to a particular task related character. The less precise the character is defined the more natural the resulting speech, since it is more likely to reflect the natural speech of the speaker. The role-play topic defines the content of the conversations. Thus, topic related utterances are more likely to occur. Role plays need to be set up

carefully to be taken seriously. To keep the conversation going for a longer time span, the participants may get opposed goals for the outcome of the role-play.

The Ernestus corpus of spontaneous Dutch (Ernestus, 2000; see also section 3.1.5) used a sales conversation task to elicit conversational speech. One speaker acted as a shop owner and the other one as a salesperson. The complexity of the negotiation task was enhanced by giving the speakers different goals and background information. The participants were also instructed that the shop owner and the salesperson were friends and both were only half time jobs. Thus, allowing the participants to talk about their own occupations to elicit a more natural speech style (see Ernestus, 2000 for further details on the setting). The role-play elicited to approximately 20 minutes of conversation. To lengthen the dialogues the role-play was embedded into further tasks.

The CLIPS (see section 3.1.1) telephone sub-corpus also used role-play. There the speakers simulated a hotel reservation.

Role-play was also used in non-S2S corpora. For example, the participants of the AMI corpus (http://corpus.amiproject.org/) simulated a design-team project meeting. The Kiel Corpus of spontaneous speech (http://www.ipds.uni-kiel.de/ipds/pub.htm#SpontaneousSpeech) used an appointment task in which the participants had to arrange meetings.

2.4.3 Map task

Another widely used way to elicit conversational speech is the map task. The participants get a map and their task is to describe the way from one landmark to another to a second speaker. One version includes somewhat deviating maps encouraging the discussion between participants. Another way to make all participants speak is rotating the task. Often only one speaker is present and asked to pretend to speak to another person. This would result in speech that is clearly not conversational but semi-spontaneous (see section 2.3). Due to the design of the map, specific phrases and utterances can be elicited.

The map task was used by the CLIPS *dialogico* subcorpus (see section 3.1.1).

The Word Segmentation Corpus also uses the map task. In this corpus, highly elaborated maps were used to elicit utterances that could create lexical ambiguity, such as "grey tanker" vs. "great anchor" (see section 3.1.17). Before the recordings, participants learn the names of the landmarks in a training session.

2.4.4 Cross-word puzzles solving

Another task to elicit conversational speech of multiple speakers is solving crossword puzzles. Due to the type of task some words (e.g. "up", "down"), numbers and the alphabet will occur frequently in the recordings. Thus, this task is especially adequate if large quantities of those

utterances are needed (for example in ASR research). Solving a crossword lasts approximately 30 minutes depending on the difficulty of the puzzle.

The ShATR corpus (http://www.dcs.shef.ac.uk/spandh/projects/shatrweb/, Crawford et al., 1994, see also 3.1.14) used this task to elicit conversational speech to model sound source segregation and localisation, speaker identification and develop ASR (automatic speech recognition) tools. To elicit speech of multiple talkers at the same time, there were five speakers in the ShATR corpus: Two pairs of speakers solved each one crossword puzzle and one speaker acted as a hint-giver.

2.4.5 Picture drawing

In this task, one speaker is asked to describe a picture to another speaker, who is trying to draw it according to the first speaker's instructions. The task requires active cooperation and interaction of speakers to produce a drawing as close to the original as possible and clarify inconsistencies. The content of the picture can be used to elicit selective words, even phrases with higher probability. This also can be seen as a drawback of the task, as it restricts the vocabulary used in the dialogs. Moreover, the person drawing the picture will be talking far less than the person describing the picture. The task can be used to elicit 30 to 40 minutes of conversational speech.

The picture-drawing task was used in the Kachna corpus (see 2.1.10) to elicit speech for the investigation of supra-segmental characteristics of conversational speech and their relationship in first and second language. In order to reduce the potential unbalance in the dialogue, the "drawing" speaker received an additional task to identify the content of small detail sections and determine their location within the picture. For the pictures and answer, sheets used by the Kachna corpus see Appendix A.

2.4.6 Spot the difference dialogues

This task is similar to the picture-drawing task. Each of the participants gets a picture, but pictures differ in some respects. Participants are not allowed to see the picture of the other(s). Their task is to find the differences among the pictures by discussing the content of it. As with the picture-drawing task, the vocabulary of the elicited speech depends on the content of the picture.

The task was used in the CLIPS *dialogico* sub-corpus (see also section 3.1.1).

2.4.7 Discussion

In this task, the participants get some topics that they have to talk about. There might be a cover story and/or the participants might be assigned different roles for the discussion resulting in a similarity to the role-play task. However, lively discussions are hard to elicit by just presenting the participants a list of topics. Critically, the corpora that contain discussions often used other ways to

elicit speech before using discussions and/or a combination of discussion and role-play. Discussions are fruitful after a warming up phase that prepares a relaxed atmosphere and when the participants are friends. The presented topics will influence the vocabulary used in the discussion.

The Ernestus corpus of spontaneous Dutch (Ernestus, 2000; see also section 3.1.5) used discussions in combination with a role-play. The participants simulated a sales conversation between buddies and discussed some prepared topics during their negotiations.

In the Nijmegen corpus of Spontaneous French (see also section 3.1.10) and the Nijmegen corpus of Spontaneous Spanish (see also section 3.1.11), the participants had a discussion at the end of their recording session. They got a list of topics about political and social issues. They were asked to choose at least five topics from the list and discuss these in order to derive shared conclusions.

The word segmentation corpus (see also section 3.1.17) also contains conversational speech from a discussion like task. The participants were given a list of topics and they can choose what to talk about.

2.4.8 Spontaneous/Free conversation

Real spontaneous conversations provide the most natural speech that is of interest for many research purposes. This type of speech is especially hart to elicit as already the knowledge of being recorded might change the speaking style. Due to ethical reasons, covert recordings of conversations are not allowed. Some corpora tried to solve this conflict using cover stories and informed the participants after the recordings.

One way to elicit spontaneous speech is to record over a long session. Speech between tasks can be seen as spontaneous conversations (see section 3.1.5; Ernestus, 2000).

The Nijmegen corpus of Spontaneous French (section 3.1.10) and the Nijmegen corpus of Spontaneous Spanish (section 3.1.11) used a highly elaborated way to elicit spontaneous conversations. In each session, one participant knew about the recordings and was instructed to bring two friends that were naïve to the purpose. After all participants were in a soundproof chamber, they got headphones and the recordings started while the confederate participants and the experimenter left the room. Participants started talking after some minutes resulting in spontaneous speech of 20 to 30 minutes. Before they became concerned, the confederate went back into the room with the instruction to talk as little as possible but keep the conversation going.

3 Available databases

The followings sections provide summaries of the available corpora as texts in alphabetical order. Section 3.2 displays the details in table form for an easier overview and comparison.

3.1 Summaries of the available databases

Corpus descriptions organized in the following way: First, the person(s) responsible for the corpus and their contact data are provided. The first paragraph specifies the name of the corpus, the language and dialect of the collected speech, the purpose for which the corpus was collected and research which was conducted using the corpus, its size and the dates of its collection. Then details on the setting, materials and tasks and participants are given. The paragraph after this, presents details of the recordings, such as audio format, sampling rate and equipment, followed by information about the annotation of the corpus and their format. The last paragraph informs about the access to the data and how to cite it.

3.1.1 CLIPS

Dr Francesco Cutugno
Dept of Informatics
University "Frederick II" of Naples, Italy
cutugno@unina.it

The '*Corpora e Lessici dell'Italiano Parlato e scritto'* is a large corpus of spoken Italian. It covers speech from 15 Italian regions. The main goal was gaining a corpus representative for the present Italian language for various research aims. It consists of about 100 hours of speech. The data were collected during 1999 to 2003.

The corpus is divided into 5 sub-corpora according to the setting and tasks used for the recordings:

a) *radiotelevisio*: radio and television broadcasts consisting of news, interviews, and talk shows,
b) *dialogico*: conversational speech from map task and spot the difference task,
c) *letto*: read speech from non-professional speakers, consisting of read sentence lists and texts covering medium-high frequency Italian words,
d) *telefonico*: telephone recordings in a conversational speech style using a hotel desk service simulation role-play task,
e) *ortofonico*: read speech from professional speakers, consisting of read sentence lists and texts covering all phonotactic sequences and medium-high frequency Italian words.

Speakers of both genders participated for the recordings. They had different social backgrounds and age ranged from 19 to 40. Data collection was done in 15 Italian cities representing different variants of Italian. In each of the cities, all five sub-corpora were collected.

The technical equipment differed for each sub-corpus (see descriptions at the web page). Audio files were processed with Goldwave and are available in wav format with a sampling rate of 22.05 kHz and a quantization of 16 bit.

Parts of the corpus are annotated in Timit style (Garofolo et al., 1993) using notepad. Transcriptions include turn, phrase, word, pauses and segmental detail. Audio and annotation files are freely available for download: http://www.clips.unina.it/. The web page also provides detailed documentations in Italian.

3.1.2 Consonant Challenge Corpus

Prof Martin Cooke
Dept of Computer Science
University of Sheffield, UK
m.cooke@dcs.shef.ac.uk

Dr Odette Scharenborg
Centre for Language and Speech Technology
Department of Language and Speech
Radboud University Nijmegen, Netherlands
O.Scharenborg@let.ru.nl

The Consonant Challenge Corpus is a British English corpus that was developed to investigate human and automatic speech recognition of consonants (Cooke & Scharenborg, 2008; García Lecumberri et al., 2008; Scharenborg & Cooke, 2008). The data consists of 10368 tokens and were collected in 2007.

The task was to read lists of vowel – consonant – vowel targets that were presented at a computer screen. The material covers all 24 English consonants in nine different vowel contexts. Moreover, the tokens were read with two different stress types. The tokens were read by 24 speakers (12 female and 12 male) aged from 18 to 49.

Recordings were made in a soundproof chamber using a desk microphone (B&K 4190). Audio files are available in wav format. They were digitally recorded and down-sampled to 25 kHz 16 bit quantization using Matlab. Segments were automatically labelled and annotations were manually corrected.

The audio and segmentation files are freely available from: www.odettes.dds.nl/challenge_IS08/downloads.html. Further details regarding materials, recordings, audio processing and results is provided at: www.odettes.dds.nl/challenge_IS08/.

How to cite:

Cooke, M., Scharenborg, O. (2008). The Interspeech 2008 Consonant Challenge, *Proceedings of Interspeech*, Brisbane, Australia, September 2008.

3.1.3 Emory Story corpus

Meg Zellers
Research Centre for English and Applied Linguistics
University of Cambridge, UK
mkz21@cam.ac.uk

The Emory Story corpus consists of read speech of native English with standard southern British accent. It was used to investigate the relationship between prosodic patterns and topic structure and segmental influences on it. The corpus sums up to approximately 4 hours of speech and was collected during 2008.

The speakers read a highly sophisticated prose text in which topic and segmental structure were varied. Before the start of the recordings, they were asked to practice. Appendix A gives an annotated version of the materials. There were 13 female and five male speakers aged from 18 to 32.

Recordings were made in a soundproof chamber using a desk microphone (B Sennheiser MKH 40 P48). Audio files were directly recoded to wav format recorded using a Compact-Flash-Card Recorder (Marantz PMD670) with a sampling frequency of 44.1 and a quantization of 32 bit. Segments were automatically labelled and annotations were manually corrected.

The corpus is orthographically annotated using Praat TextGrid format. The audio and annotation files are available on request.

How to cite:
The corpus has not been published yet, please use the present paper.

3.1.4 English Assimilation

Dr Oliver Niebuhr
Laboratoire Parole et Langage
University of Provence, Aix-en-Provence, France
Oliver.Niebuhr@lpl-aix.fr

Dr Gareth Gaskell
Dept of Psychology
University of York, UK
g.gaskell@psych.york.ac.uk

The English Assimilation corpus covers read standard British English. The corpus was collected to investigate regressive assimilation of place of articulation of alveolar and postalveolar fricatives across languages. It consists of approximately 1.5 hours of speech and was collected in 2008.

Similar to the French Assimilation corpus, the speakers read lists of sentences containing targets that allow such assimilation. The 42 sentences were repeated four times by each speaker during the recordings. The corpus was conducted with four female undergraduate students.

Recordings were conducted in an anechoic chamber using a desk microphone (Sennheiser) directly on CD by a CD writer. The audio files are available in wav format with a sampling rate of 44.1 kHz and a quantization of 16 bit.

The corpus is orthographically annotated, pauses are also marked. Transcriptions are available in Praat TextGrid format. Audio and transcription files are available on request.

How to cite:

Niebuhr, O., Clayards, M., Lancia. L., & Meunier, C. (2008). Place Assimilation in Sibilant Sequences - Comparing French and English. Poster presented at the 4th S2S workshop, Prague, Czech Republic.

3.1.5 Ernestus corpus of spontaneous Dutch

Dr Mirjam Ernestus
Radboud University Nijmegen & Max Planck Institute for Psycholinguistics
Mirjam.Ernestus@mpi.nl

The Ernestus corpus of spontaneous Dutch represents native Dutch of the western parts of the Netherlands. It has been collected to extract a very casual speech style for the study of reduction and coarticulation processes and its implication for speech recognition (e.g. Ernestus, 2000; Plug, 2005). It also has been used to develop and improve automatic transcription systems for casual speech (Schuppler et al., 2008). The corpus covers approximately 15 hours of speech and was collected in 1995 and 1996.

To obtain a natural speech style, participants were asked to bring their friends for the recordings. Sessions were composed of two parts. First, free conversations about different topics for approximately 40 minutes per session. Second, a sales conversation role-play (see also section 2.4.2) of nearly 40 minutes per session. Furthermore, the corpus includes spontaneous conversations between tasks of roughly 5 minutes per session and read lists of monosyllabic words and non-words covering all Dutch monophtongs from each speaker. The speakers all had an academic degree, were male and between 21 and 55 years old.

The recordings were made in a soundproof booth using desk microphones in front of the speakers (Sennheiser MD527) and taped by DAT-recorder (Denon DTR 2000). The audio files are available in wav-format with a sampling frequency of 44.1 kHz.

The corpus is manually orthographically annotated and time-aligned at an utterance level. An automatically generated broad phonemic transcription is also available. The transcriptions are in Praat TextGrid format and include repetitions, hesitations, false starts and contrastive accents. Audio and transcription files are available on request.

How to cite:

M. Ernestus (2000). *Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface*. Utrecht: LOT.

3.1.6 FonDat 1 to 3

Dr Torbjørn Svendsen
Dept of Telecommunications and Electronics
University of Science and Technology, Trondheim, Norway
torbjorn@iet.ntnu.no

FonDat is a series of corpora of read Norwegian using South-East Norwegian dialect. They were collected to gain high quality materials for text to speech synthesis (Amdal & Svendsen, 2005, 2006; Amdal et al., 2006; Bjørkan et al., 2005; Natvig & Heggtveit, 2004; Meen et al., 2005; Svendsen et al. 2005). They consist of approximately 4000 (FonDat 1), 6000 (FonDat 2), and 12000 (FonDat 3) sentences and were collected during 2004 to 2008.

For all corpora, lists of sentences originating from newspaper texts were read. The chosen sentences were selected due to readability criteria and to cover all possible Norwegian diphones. All manuscripts were written in Norwegian Bokmål. The sentences were read one at a time with the instruction to read in an expressive (FonDat 1) or everyday (FonDat 2 & 3) manner. Professional speakers (actors and radio voices) were chosen on voice pleasantness and the ability to read aloud consistently and accurately. In total there are 13 different speakers (see Table 2 for details) with gender equally distributed in all corpora.

Recordings were made in a soundproof booth in isolation using desk microphones in front of the speakers (Milab LSR 1000). Besides the sound, EGG was recorded for FonDat1. For further details see Table 1. The audio files are available in wav-format with a sampling frequency of 16 kHz (downsampled from 48 kHz) and 16 bit quantization.

For all three corpora the distribution includes the orthographic manuscript updated with corrections during recording and post-processing. Automatic phonemic annotation based on the manuscript is available in TextGrid format. F0 estimation was performed for all data and can be added to the distribution on request. Approximately 10% of FonDat1 was manually annotated phonemically and prosodically in the Praat TextGrid format. For this subset XML format Meta data files are available containing e.g. automatic prosodic analysis. Further details are provided by in Amdal and Svendsen (2006) and the web page of the project: http://www.iet.ntnu.no/projects/fonema/index_eng.php.

How to cite:

Amdal, I. & Svendsen, T. (2006). FonDat1: A Speech Synthesis Corpus for Norwegian. *Proceedings of LREC-2006*. Genova, Italy.

Table 1: Details on recordings of the FonDat corpora, separated for each corpus.

	FonDat 1	*FonDat 2*	*FonDat 3*
Speakers	*2: 1f, 1 m*	*12: 6 f, 6 m*	*2: 1f, 1 m*
Size of speech per speaker	approximately 2000 sentences, 3 hours	approximately 500 sentences, 40 minutes	approximately 6000 sentences, 10 hours
Audio format	wav	wav	wav
Channels	Ch1: microphone Ch 2: EGG: Laryngograph Ltd	Ch1: microphone	Ch1: microphone
Microphone	desk microphone: Milab LSR 1000	desk microphone: Milab LSR 1000	desk microphone: Milab LSR 1000
Recorder / Sound Card	DAT: Fostex D10 Digital Master Recorder Creative Studios Sound Blaster Live 5.1 Platinum	EDIROL UA-25 SpeechRecorder	EDIROL UA-25 SpeechRecorder
Sampling Rate	16 kHz (resampled from originally 48 kHz)	16 kHz (resampled from originally 48 kHz)	16 kHz (resampled from originally 48 kHz)
Quantization	16 bit	16 bit	16 bit
Audio-processing Software	ESPS, EST and Sox	ESPS, EST and Sox	ESPS, EST and Sox

Note: Both FonDat2 and FonDat3 were recorded using a head-mounted miniature microphone DPA 4060-FM in addition to the desk microphone. These recordings were not processed further and are not a part of the distribution.

Table 2: Details on speakers

Speaker ID	*Database content*			*Gender*	*Age group*
	FonDat1	*FonDat2*	*FonDat3*		
tjf	2060 sent, 3.1 hours	-	-	male	50-60
t01	-	502 sent, 37 min	-	male	40-50
tjk,t02	2063 sent, 2.7 hours	520 sent, 43 min	-	female	30-40
t03	-	518 sent, 45 min	-	male	50-60
t04	-	519 sent, 41 min	-	female	40-50
t11	-	519 sent, 42 min	-	male	40-50
t12	-	516 sent, 42 min	-	female	20-30
t13	-	519 sent, 38 min	-	male	20-30
t14	-	520 sent, 42 min	-	female	40-50
t15	-	520 sent, 41 min	6391 sent, 9.8 hours	male	30-40
t16	-	519 sent, 42 min	6391 sent, 9.8 hours	female	50-60
t17	-	513 sent, 40 min	-	male	30-40
t18	-	517 sent, 39 min	-	female	30-40

Note: tjk and t02 is the same person, the speaker ID system of FonDat1 and FonDat2 differs.

3.1.7 French Assimilation Corpus

Dr Oliver Niebuhr
Laboratoire Parole et Langage
University of Provence, Aix-en-Provence, France
Oliver.Niebuhr@lpl-aix.fr

The French Assimilation Corpus covers read standard French. The corpus was collected to investigate regressive assimilation of place of articulation of alveolar and postalveolar fricatives. It sums up to approximately 1.5 hours of speech and was collected in 2007.

Similar to the English assimilation corpus, the speakers read lists of sentences containing targets that allow such assimilation. The 72 sentences were repeated four times by each speaker during the recordings. The corpus was conducted with four female speakers aged 23 to 54 in an anechoic chamber.

The corpus is orthographically annotated, pauses are also marked. Transcriptions are available in Praat TextGrid format and audio files in wav format with a sampling rate of 44.1 kHz and 16 bit resolution.

Audio and transcription files are available on request.

How to cite:

Niebuhr, O., C. Meunier, L. Lancia (2008). On place assimilation in French sibilant sequences. *Proceedings of the 8th international seminar on speech production*, Strasbourg, France.

3.1.8 GRID

Prof Martin Cooke
Dept of Computer Science
University of Sheffield, UK
m.cooke@dcs.shef.ac.uk

GRID is an audio-visual corpus of read English of different English accents. It was collected 'to support joint computational-behavioural studies in speech perception' (Cooke et al., 2006) and used to model intelligibility in noise (Barker & Cooke, 2007) and to support the Speech Separation Challenge (2006). It consists of 34000 sentences and was collected in 2005.

The subjects read lists of sentences of the form "put red at G9 now" displayed on a computer screen. There were 34 speakers (16 female, 18 male), aged from 18 to 49.

Recordings were made in a soundproof booth with a desk microphone (B&K 4190). Audio files are available in wav format with a sampling rate of 50 kHz and 25 kHz. Orthographic annotation of words was done automatically using forced-alignment techniques using TIMIT style. All audio, video and annotation files can be downloaded from http://www.dcs.shef.ac.uk/spandh/gridcorpus/. Further details regarding materials, recordings, and audio processing is provided in the paper of Cooke et al., 2006.

How to cite:

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am*, 120 (5), 2421–2424. Online-version: http://www.dcs.shef.ac.uk/~martin/jasagrid.pdf.

3.1.9 Kachna corpus

Helena Spilkova
Norwegian University of Science and Technology, Trondheim, Norway
hele.spilkova@email.cz

Prof Wim van Dommelen
Dept of Linguistics
Norwegian University of Science and Technology, Trondheim, Norway
wim.van.dommelen@hf.ntnu.no

Dr Jacques Koreman
Institute for Language and Communication Studies
Norwegian University of Science and Technology, Trondheim, Norway
jacques.koreman@hf.ntnu.no

The Kachna corpus consists of conversational speech of native Norwegian, native Czech and second language English from native Norwegians and native Czechs. It was collected to investigate suprasegmental phenomena in conversational interaction (in particular turn taking, backchannel responses) of first and second language speakers. It consists of approximately 12 hours of speech and was collected in 2008.

The speakers fulfilled a picture description task. One speaker described a picture to the other one who was trying to draw it. In order to reduce the potential unbalance in the dialogue, the 'drawing' speaker received additional tasks. An example of the picture is attached in appendix A, see also section 3.4.5 for further details on the task. In each recording session, the task was first performed in English as L2 and then repeated (with switched roles and a different picture) in the native language of the speakers. There were five pairs of speakers each for native Czech and native Norwegian language. The speaker pairs were either friends or classmates. There were five female and five male native Czech speakers and four female and six male native Norwegian speakers. Speakers of both countries had various accents. The age ranged between 19 and 35 years.

Recordings were made in a soundproof booth (in Trondheim and Prague). Audio files are in wav format with a sampling rate of 44.1 kHz. Table 2 gives further details concerning the equipment and audio. The corpus will be orthographically annotated at sentence level during the next year. Audio files will be provided on request.

Table 2: Details on speakers and recordings of the Kachna corpus, divided by sub-corpora.

	Czech / English	*Norwegian /English*
Speakers	5 f, 5 m	4 f, 6 m
Size of speech per speaker	Each pair 30 to 40 min	Each pair 30 to 40 min
Audio format	wav	wav
Channels	Ch1: desk microphone speaker 1 Ch2: desk microphone speaker 2	Ch1: desk microphone speaker 1 Ch2: desk microphone speaker 2
Microphone	AKG C 4500 B-BC	MILAB LSR-1000
Recorder / Sound card	Sound Blaster Audigy 4	Creative SB Live
Sampling Rate	44.1	44.1
Quantization	16	16
Audio-processing software	Sound Audio Studio 8.0	

How to cite:

The corpus has not been published yet, please use the present paper.

3.1.10 Nijmegen corpus of casual French

Dr Mirjam Ernestus
Radboud University Nijmegen & Max Planck Institute for Psycholinguistics
Mirjam.Ernestus@mpi.nl

The Nijmegen corpus of spontaneous French is an audio-visual collection of conversational speech of native French. The aim of the corpus was to collect large amounts of conversational speech for phonetic and linguistic analysis. It covers about 34.5 hours of speech and was collected in 2007.

To obtain a natural speech style, participants were asked to bring two friends for the recordings. These two friends were naïve to the task and were the only ones being recorded. One session was split into three parts. The first part consisted of spontaneous conversational speech after the experimenter and their initiated friend left the room. The second part also consisted of conversational speech with the initiated friend joining his/her friends in the room again with the task to speak as little as possible but keep the conversation going. In the third part, several politics and social issues were discussed. There were 46 speakers, 22 female and 24 male with age ranging from 18 to 51 years. All speakers had completed the secondary education level in France.

The recordings were made in a soundproof booth using head-mounted microphones (Samson QV), an Edirol R-09 solid-state recorder and a stereo microphone preamplifier. The audio files are in wav format with a sampling frequency of 48 kHz and 32 bit quantization. Video files were conducted with a Canon XM2 Mini-DV video camera.

The corpus will be orthographically annotated. Audio, video and transcription data can be provided on request.

How to cite:

Torreira, F. & Ernestus, E. (subm). The Nijmegen Corpus of Casual French.

3.1.11 Nijmegen corpus of spontaneous Spanish

Dr Mirjam Ernestus
Radboud University Nijmegen & Max Planck Institute for Psycholinguistics
Mirjam.Ernestus@mpi.nl

The Nijmegen corpus of spontaneous Spanish is an audio-visual collection of conversational speech of native Spanish. The aim of the corpus was to collect large amounts of conversational speech for phonetic and linguistic analysis. It covers approximately 30 hours of speech and was collected in 2008.

To obtain a natural speech style, participants were asked to bring two friends to the recordings. One session was split into three parts. The first part consisted of spontaneous conversational speech after the experimenter and their initiated friend left the room. The second part also consist of conversational speech with the initiated friend joining his/her friends in the room again with the task to speak as little as possible but keep the conversation going. In the third part, several politics and social issues were discussed. There were 40 speakers, 20 female and 20 male. All speakers were students of the University of Madrid and aged between 19 and 25.

The recordings were made in a sound attenuated booth using head-mounted microphones (Samson QV) and an Edirol R-09 solid-state recorder. The audio files are in wav format with a sampling frequency of 44.1 kHz and 16 bit resolution. Video files were conducted with a Sony HDR-SR7 video camera.

The corpus will be orthographically annotated. Audio, video and transcription data can be provided on request.

How to cite:

Torreira, F. & Ernestus, E. (in prep). The Nijmegen Corpus of Spontaneous Spanish.

3.1.12 Prague Phonetic Corpus

Dr Jan Volín
Dept of Phonetics
Charles University, Prague, Czech Republic
jan.volin@ff.cuni.cz

The Prague Phonetic Corpus is a collection of native Czech speech. It is collected for linguistic-phonetic research and was used in studies addressing manual and automatic segmentation, acoustic properties of segments and intonation (Machač et al., 2007; Volín et al., 2008). Data collection covers 250 speakers for each of them read and spontaneous speech is available. The corpus was initialized in 1998 and is still growing.

Participants fulfil four tasks: self-introduction, reading phonetically balanced sentences and a poem and telling a narrative. The data of the first task will not be provided. The phonetic balanced sentences contain all Czech phones, different syntactic structures and are easy to read. The prose is far more demanding to read. The narrative was based on a carton strip and the participants were asked to tell a story that goes with the pictures (see appendix A for the materials). Each of the tasks lasts approximately 3 minutes. Before the recordings, participants get some time for preparation. 250 students of Czech language and phonetics participated for the recordings (183 female and 67 male). All had a central Bohemian dialect with age ranging from 18 to 40 years.

Recordings were made in a soundproof chamber with a desk microphone (AKG C 4500 B-BC) in front of the speakers. The speech was digitally recorded using the Sound Audio Studio 8.0 software and a Sound Blaster Audigy 4 sound card. Audio files are in wav format with sampling rates of 22.05 or 32 kHz and a quantization of 16 bit.

A small part of the corpus was manually annotated using Praat software and TextGrid format. The transcriptions provide orthographical, segmental and linguistic information. The audio and transcription files are available on request.

How to cite:

Volín, J., Skarnitzl, R., Machač, P., Janoušková, J. & Veroňková, J. (2008). Reliabilita a validita popisných kategorií v Pražském fonetickém korpusu. In M. Kopřivová & M. Waclawičová (Eds.), *Čeština v mluveném korpusu* (249–254). Praha: Nakladatelství LN / ÚČNK.

3.1.13 RUNDKAST

Dr Torbjørn Svendsen
Dept of Telecommunications and Electronics
University of Science and Technology, Trondheim, Norway
torbjorn@iet.ntnu.no

RUNDKAST is a broadcast corpus of standard Norwegian. It was collected to gain training materials for automatic speech recognition. It covers various speech styles such as read and conversational speech from news, interviews and debates. It consists of approximately 77 hours of speech of various Norwegian dialects. Recordings were collected during 1995 to 2006.

Material was obtained from the Norwegian Broadcasting Corporation (NRK), which is Norway's major broadcasting institution. It consists of 115 episodes of eight different radio programs ranging from 15 to 60 minutes.

The quality of the recordings varies from studio to telephone. The originally audio files had a sampling frequency of 48 kHz compressed to 384 kbs mpeg2 format. After decompressing, they were down-sampled to 16 kHz and a quantization of 16 bit using *sox* tool under linux and stored in wav format.

The corpus is manually annotated. Transcriptions give details about hierarchical levels such as *sections*, *turns*, and *segments* (stretches of 2 to 5 min of speech) and *background* sounds. Transcriber was used for the annotations and the transcriptions are in transcriber-XML format (trs files). A subset also contains a broad phonetic annotation that was done manually using Praat and will be provided as TextGrid files. The Rundkast manual provides further details on the transcription guidelines. The audio and transcription files are available on request.

How to cite:

Strand, O. M., Svendsen, T., Amdal, I. (in prep). *RUNDKAST: A Norwegian Broadcast News Speech Corpus.*

3.1.14 ShATR

Dr Guy Brown
Dept of Computer Science
University of Sheffield, UK
g.brown@dcs.shef.ac.uk

The Sheffield-ATR multiple-simultaneous speaker database consists of conversational British and American English. It was collected as a corpus of overlapping speech for the investigation of computational auditory scene analysis and sound source segregation and localisation (Crawford et al. 1994, Karlsen et al., 1998). It sums up to approximately three hours and was collected in 1994.

Four of the five participants worked in pairs to solve two crosswords. The fifth participant was a hint-giver. The participant who acted as hint-giver walked around in one session to create background noise. Besides the crossword solving task each speaker read lists consisting of the alphabet, numbers, crossword specific words and Timit shibboleths. A passage of a newspaper text was also read. All five speakers were male. Four of them were native Britain and the fifth native American with age ranging from 19 to 40 years.

Recordings were made in a reverberation chamber in Kyoto, Japan. In front of each speaker there was a desk microphone (B&K type 4134). Furthermore, there were two manikins with microphones for each ear and an omnidirectional microphone in the middle of the table. The speech was digitally recorded by a TASCAM DA-88 recorder with a sampling rate of 48 kHz and a 16 bit quantization. Audio files have been down-sampled to 16 kHz and split into one-minute segments for each of the eight channels. Audio files are in wav format.

The corpus was manually and automatically annotated. Manual transcription provides structural, noise and orthographic information. Separation into words and phones was automatically conducted. Annotations are in Timit style. The audio and transcription files are available for download at: http://www.dcs.shef.ac.uk/spandh/projects/shatrweb/. A further detail regarding setting and equipment provides the paper of Crawford et al. (1994).

How to cite:

Crawford, M. D., Brown, G. J, Cooke, M. P., & Green, P. D. (1994). Design, collection and analysis of a multi-simultaneous- speaker corpus, *Proc. Inst. Acoustics*, 16(5), 183-190.

3.1.15 Speech Impairment Corpus

Dr Sara Howard
Department of Human Communication Sciences
University of Sheffield, UK
s.howard@sheffield.ac.uk

The Speech Impairment Corpus is an audio-visual corpus of spoken native Britain English. The aim of the corpus was to investigate speech impairments (e.g. Howard, 2001, 2004a, 2004b, 2007) and to improve diagnosis tools and therapy. The corpus sums up to approximately 8 hours and was collected during 2000 to 2008.

The corpus consists of read speech, elicited speech by repetition of sounds, syllables and words, and semi-spontaneous speech from picture descriptions and conversations in an interview style. The participants were 20 children with speech disorders. They were in mainstream education and aged from 6 to 16.

Recordings were made in a soundproof clinic room using a standard EPG microphone. Audio files vary in format and quality. Later EPG recordings have a sampling rate of 100 frames per second.

The corpus is manually annotated. Transcriptions include orthographical, phonetic, morphological, semantic, and feature information.

The access to the data is strongly restricted due to ethical reasons.

How to cite:

Howard, S. J. (2001). The realisation of affricates in a group of individuals with atypical speech production: a perceptual and instrumental study. Clinical Linguistics & Phonetics, 14, 133–138.

3.1.16 Vegtalk

Dr Richard Ogden
Dept of Language and Linguistic Science
University of York, UK
rao1@york.ac.uk

Vegtalk is a broadcast corpus of mainly conversational British and American English. It was collected to gather assessments in spontaneous talk and was used to study the relationship between conveying agreement and phonetic format (Ogden, 2006). The corpus consists of about 1.5 hours of speech and was collected in 2001.

The recordings come from a radio phone-in show called Vegtalk in which the two presenters, each one guest and a number of callers discuss food in an interview style. The data were collected from three programmes, one with a male and two with female guests. The two presenters are Britain (London), both are male and in their 40s. Guests and callers are mainly Britain of both genders with dialects from all around Britain.

The audio recordings are in .wav and .aiff format and of good quality. Parts of the corpus are annotated in a conversational analysis style, particularly overlapping talk and some instances of assessments and vocatives. Transcriptions are in .doc format. Audio and transcription files are available on request. Further details on the radio show provides this page: http://www.bbc.co.uk/food/tv_and_radio/vegtalk_index.shtml.

How to cite:
The corpus has not been published yet, please use the present paper.

3.1.17 Word Segmentation Corpus

Dr Sven Mattys
Dept of Psychology
University of Bristol, UK
sven.mattys@bristol.ac.uk

The Word Segmentation Corpus is a collection of read and conversational speech of standard southern British English. It was/is collected to study word segmentation in lexical ambiguous conditions. Furthermore, different speech styles, such as read lists, read texts and conversational speech will be available for the same speakers. The final corpus will be about 70 to 80 hours of speech. The collection was initiated in 2008 and scheduled end is 2010.

All speakers will fulfil three tasks: First, they participate in a variant of the Map Task (Anderson et al., 1991) in which they explain the way to certain landmarks on the map to another speaker. The maps are designed in a way that allows lexical ambiguous utterances such as 'great anchor' vs. 'grey tanker'. Second, a selection of words and phrases of the map task will be carefully read by the speaker in a subsequent recording session, resulting in a collection of the same utterances in conversational and read speech. Third, spontaneous conversational speech between pairs of participants will be collected. There will be about 30 speakers with age ranging from 19 to 40.

Recordings were made using head-mounted microphones (Shure WH20) straight to hard disc using CoolEdit as recording software. Audio files are in wav format with a sampling frequency of 32 kHz and a quantization of 16 bit. The corpus will be manually annotated using Praat. Transcriptions provide detail regarding orthography, prosody, pauses, disfluencies and word boundaries in TextGrid format. The audio and transcription files will be available on request at the end of the data collection period.

How to cite:
The corpus has not been published yet, please use the present paper.

3.1.18 York Lab Data

Dr Richard Ogden
Dept of Language and Linguistic Science
University of York, UK
rao1@york.ac.uk

The York Lab Data is a collection of conversational speech of British and American English. It was collected to gain conversational speech for conversation analysis purposes. It was used to study the relationship between conveying agreement and phonetic format (Ogden, 2006). The corpus includes approximately 4 hours of speech and was collected from 2000 to 2003.

To elicit a very natural speech style, pairs of friends were recruited as participants. The participants were not given any specific task, but were told that they were being recorded for the purposes of general linguistic research. Each pair of friends was chatting for about 45 to 50 minutes. Overall, there were four pairs of speakers speaking dialects from all over Britain, one pair speaking US English and one Singaporean pair. Overall, there were nine female and three male speakers with age ranging from 19 to 40.

Recordings were conducted in a soundproof room using a desk microphone (B&K) and a DAT recorder. The audio files were digitized to wav format with a sampling frequency of 44.1 kHz and a quantization of 16 bit. A part of the corpus is manually annotated in conversational analysis style using Microsoft word as text editor. Audio and annotation files are available on request.

How to cite:

Ogden, R. (2006) Phonetics and social action in agreements and disagreements. *Journal of Pragmatics*, 38, 1752–1775.

3.2 Table overviews

Some abbreviations were used for the construction of the tables.

General information:

Language:

L1	native language of the speaker
L2	(any) non native language of the speaker

Speech style:

C	conversational speech
S	semi-spontaneous monologue
E	(experimental) elicited speech
R	read speech

Speakers

N	number of speakers

Gender

f	female
m	male

Dialect

US

UK

SSBE	standard Southern Britain English
SBE	standard Britain English

Audio

Channels

ch	channel
mike	microphone

3.2.1 General information

Name	*Language*		*Size*	*Speech style* C	S	E	R	*Used tasks to elicit speech*	*Aim of corpus construction*	*Dates*	*Sharing*	*Cite*
CLIPS	Italian	L1	≈ 100 hours	✓	✓		✓	radio and TV broadcasts, map task and spot the difference task conversations, read lists and texts, role-play telephone conversation	representative corpus for present Italian language for various research goals	1999-2003	public free	
Consonant Challenge Corpus	English	L1	10 368 tokens				✓	read lists of vowel - consonant - vowel sequences	consonant recognition in humans and machines	2008	public free	Cooke & Scharenborg 2008
Emory Story corpus	English	L1	≈ 4 hours				✓	read sophisticated prose text with varying topic structure	investigation of the relationship between prosodic pattern, topic and segmental structure	2008	on request	
English assimilation	English	L1	672 sentences				✓	read lists of sentences that were prepared to elicit assimilation	investigate alveolar and postalveolar assimilation in fricatives across languages (English & French)	2008	on request	
Ernestus corpus of spontaneous speech	Dutch	L1	≈ 15 hours	✓				spontaneous conversation, role-play, read lists of words and non-words	representative corpus of spontaneous Dutch, main research on reduction and coarticulation	1995-1996	on request	Ernestus, 2000
FonDat 1	Norwe-gian	L1	≈ 4 000 sentences				✓	read lists of sentences compiled from newspaper texts	collecting high quality materials for text to speech synthesis	2004	on request	Amdal & Svendsen, 2006
FonDat 2	Norwe-gian	L1	≈ 6 000 sentences				✓	read lists of sentences compiled from newspaper texts	collecting high quality materials for text to speech synthesis	2006	on request	
FonDat 3	Norwe-gian	L1	≈ 12 000 sentences				✓	read lists of sentences compiled from newspaper texts	collecting high quality materials for text to speech synthesis	2007-2008	on request	
French assimilation	French	L1	11904 sentences				✓	read lists of sentences that were prepared to elicit assimilation	investigate alveolar and postalveolar assimilation in fricatives across languages (English & French)	2007	on request	Niebuhr et al., 2008
GRID	English	L1	34 000 sentences				✓	read lists of sentences (e.g. "put red at G9 now")	collection of audio-visual materials for computational-behavioural studies	2005	public free	Cooke et al., 2006
Kachna	Norwe-gian, Czech, English	L1, L2	≈ 12 hours	✓				conversional speech from a picture drawing task	collection of conversational speech in a Germanic (Norwegian) and a Slavic (Czech) language as native languages of the speakers and their second language English; used for the investigation of turn-taking	2008	on request	

Table: General information (continued)

Name	*Language*		*Size*	*Speech style* C	S	E	R	*Used tasks to elicit speech*	*Aim of corpus construction*	*Dates*	*Sharing*	*Cite*
Nijmegen corpus of spontaneous French	French	L1	≈ 34.5 hours	✓				conversational speech from free conversations and a discussion	collection of conversational speech for various purposes	2007	on request	Torreira & Ernestus, in prep
Nijmegen corpus of spontaneous Spanish	Spanish	L1	≈ 39 hours	✓				conversational speech from free conversations and a discussion	collection of conversational speech for various purposes	2008	on request	
Prague Phonetic Corpus	Czech	L1	≈ 12.5 hours		✓		✓	read texts, spontaneous speech from a narrative	manual and automatic segmentation, acoustic properties of segments and intonation, acoustic realization of Czech	1995-growing	on request	Volín et al. 2008
Rundkast	Norwe-gian	L1	≈ 77 hours	✓	✓		✓	radio and TV broadcasts containing read, spontaneous and conversational speech	large pool of materials for research on automatic speech recognition	1995-2006	on request	Strand et al., in prep
ShaTR	English	L1	37 minutes	✓				conversational speech from a cross-word solving task and read list of selected words	collection of overlapping speech in noisy conditions, used for research on sound segregation	1994	public free	Crawford et al., 1994
Speech Impairment Corpus	English	L1	≈ 8 hours		✓	✓	✓	repetition, picture description, interview	investigation of speech impairments, development of diagnosis tools	2000-2008	restricted	Howards, 2001
Vegtalk	English	L1	≈ 1.5 hours	✓			✓	conversational and read speech from broadcasts	collection of mainly conversational speech for conversational analysis	2001	on request	Ogden, 2006
Word Segmentation Corpus	English	L1		✓			✓	conversational speech from map task and free dialogues, read lists of selected utterances from the map task	collection of conversational and read speech from the same speakers for research on word segmentation and lexical ambiguity	2008-2010	public free	
York Lab data	English	L1	≈ 4 hours	✓				conversational speech	the corpus was collected and used for conversational analysis	2000-2003	on request	Ogden, 2006

3.2.2 Speakers

Name	*Language*		*N*	*Gender*	*Age*	*Education*	*Dialect*	*Other*
CLIPS	Italian	L1		f, m	19-40	various	15 Italian regions	professional and non-professional speakers
Consonant Challenge Corpus	English	L1	24	12 f, 12 m	18-49	academic	UK	
Emory Story corpus	English	L1	20	5 f, 15 m	18-32	mainly students	SSBE	
English assimilation	English	L1	4	4 f	19-40	students	SBE	
Ernestus corpus of spontaneous speech	Dutch	L1	16	m	21-55	academics	Western Dutch	speakers were friends or colleagues
FonDat 1	Norwegian	L1	2	1 f, 1 m	19-40		South-East Norwegian	prof. speakers, selected by reading skills
FonDat 2	Norwegian	L1	12	6 f, 6 m	19-40		South-East Norwegian	prof. speakers, selected by reading skills
FonDat 3	Norwegian	L1	2	1 f, 1, m	19-40		South-East Norwegian	prof. speakers, selected by reading skills
French assimilation	French	L1	4	4 f	19-40		Standard French	
GRID	English	L1	34	16 f, 18 m	19-40	academics	various British	
Kachna	Norwegian, Czech, English	L1, L2	20	9 f, 11 m	19-35	students		speakers were friends or classmates
Nijmegen corpus of spontaneous French	French	L1	46	22 f, 24 m	18-51	min. secondary education	Central/Northern French	speakers were friends or colleagues
Nijmegen corpus of spontaneous Spanish	Spanish	L1	52	f, m	19-25		Madrid region	speakers were friends or colleagues
Prague Phonetic Corpus	Czech	L1	250	f, m	18 -	students	Standard Bohemian	
Rundkast	Norwegian	L1	≈ 30	f, m	19-40			
ShaTR	English	L1	5	5 m	19-40	academics	UK, US	
Speech Impairment corpus	English	L1	20	f	6-16	pupils	UK	patients with speech disorders
vegtalk	English	L1	3	m	19-40		UK, US	callers
Word Segmentation Corpus	English	L1	≈ 30	f, m	19-40	students	SSBE	
York Lab data	English	L1	10	f, m	19-40		UK, US, Singapore	speakers were friends

3.2.3 Audio

Name	*Format*	*Channels*	*Sampling frequency in kHz*	*Quantization in bit*	*Microphones*	*Sound card / Recorder*	*Processing / Recording Software*
CLIPS	wav	depending on task	22.05	16			
Consonant Challenge Corpus	wav	ch1: desk mike	25	16	B&K 4190		Matlab
Emory Story corpus	wav			16			
English assimilation	wav	ch1: desk mike	44.1	16			
Ernestus corpus of spontaneous speech	wav	ch1: desk mike speaker 1; ch2: desk mike speaker 2	44.1	16	Sennheiser MD527	DAT: Denon DTR 2000	
FonDat 1 to 3	wav	ch1: desk mike, ch2: EGG: Laryngograph Ltd	48/16	16	Milab LSR 1000	DAT: Fostex D10 Digital Master Recorder	Creative Studios Sound Blaster Live 5.1 Platinum
French assimilation	wav	ch1: desk mike	44.1	16			
GRID	wav	ch1: desk mike; video	50/25	16	B&K 4190		Matlab
Kachna	wav	ch1: desk mike speaker 1; ch2: desk mike speaker2	44.1	16	AKG C 4500 B-BC (Cz-En); MILAB LSR-1000 (No-En)	Sound Blaster Audigy 4 (Cz-En); Creative SB Live (No-En)	Sound Audio Studio 8.0 (Cz-En)
Nijmegen corpus of spontaneous French	wav	ch1: head mounted mike speaker 1; ch2: head mounted mike speaker 2; video	48	32	Samson QV	Edirol R-09 solid-state recorder, Canon XM2 Mini-DV video camera	
Nijmegen corpus of spontaneous Spanish	wav	ch1: head mounted mike speaker 1; ch2: head mounted mike speaker 2; video	44.1	32	Samson QV	Edirol R-09 solid-state recorder, Sony HDR-SR7 video camera	
Prague Phonetic Corpus	wav	ch1: desk mike	22.05/32	16	AKG C 4500 B-BC	Sound Blaster Audigy 4	Sound Audio Studio 8.0
Rundkast	wav		48/16	16			sox
ShaTR	wav	ch 1-5: head mounted mikes speakers 1 to 5; ch 6: omnidirectional mike; ch 7 & 8: manikins, video	48	16	speakers: RAMSA WM-S10, omnidirectional: CROWN PZM30, mannikin: B&K 4134	Yamaha HA-8 preamplifier, TASCAM DA-88 recorder	
Speech Impairments corpus	wav	ch1: mike, ch2: EPG; video	differs				
vegtalk	wav		radio	16			
Word Segmentation Corpus	wav	ch1: head mounted mike speaker 1; ch2: head mounted mike speaker 2	32	16	Shure WH20		CoolEdit
York Lab data	wav		studio				

4 S2S data sharing regulations outline

Various ethical aspects need to be considered when sharing data, for example property rights, rights of the participant, access regularities, or the secure storage of the data. In order to facilitate the usage of the data by other persons and thus allow the sharing of corpora, formal criteria regarding the data are also necessary. This concerns questions about which data should be provided, how the data should be structured and named, which secondary information such as transcribers guidelines are necessary.

There already exist distribution centres for language resources that work on quality standards of language/speech data. For European languages, ELRA (European Language Resources Association, http://www.elra.info/) is the largest according association, and ELDA (Evaluations and Language resources Distribution Agency) is a company incorporated in ELRA dealing with the all the commercial and business-oriented tasks of the association.

At the 5th S2S meeting in Aix-en-Provence the members of project 1 agreed that for documentation and data sharing the ELRA standards should apply. The according criteria are given by van den Heuvel and colleagues (2000; 2008) manual on the validation of language recourses. The present chapter therefore refers to this paper for an overview of the requirements. Additionally, the book of Schiel and Draxler (2003) provides a detailed introduction to the production and validation of speech corpora.

Only some points will be addressed in this chapter, especially those relevant and specific for the corpora collected within S2S. On the other hand, some additionally considerations regarding the special requirements of spontaneous / conversational speech corpora are provided that do no claim to be complete.

4.1 Ethics and property rights

The following section summarizes some relevant details of part C of annex II of the S2S contract. Additionally to the S2S contract country specific laws apply. Thus, parties sharing data or developing corpora together should make sure that their agreements are in accordance with national law.

4.1.1 Proprietary rights

Most of the corpora listed in his manual were not collected within the S2S network. The details of these corpora were kindly provided by the researchers in charge. There may be different property rights and conditions for these corpora than for the ones collected within

S2S. A third group are corpora that were collected in cooperation with third parties. Besides the clauses of the S2S contract country specific laws apply to the property rights.

Corpora collected (entirely) within S2S are: The Emory Story Corpus (3.1.3), the English Assimilation Corpus (3.1.4), the French Assimilation Corpus (3.1.7), and the Kachna corpus (3.1.9). These will be property of the contractor that is the corresponding University (core contract, Article 1.2) of the researcher in charge. If more than one contractor contributed to the corpus (which might be case with the English Assimilation corpus) they have joint ownership on it.

There are plans for future S2S and third party cooperation's for building corpora. Rein Owe Sikveland plans to build a corpus of conversational Norwegian in cooperation with KTH in Sweden. The Word segmentation corpus and two corpora by Carlos Gussenhoven shall get support for the orthographical transcription of the corpora by a therefore appointed S2S fellows. The status of these corpora is unclear to me.

4.1.2 Access

The researcher in charge is given in the corpus description section (see 3.1) and should be contacted for further information on the sharing of the corpus.

All contractors have granted access rights upon written request (II.32). This should be royalty-free unless otherwise agreed before signature of the contract. Access of third parties is up to the contractor, as long as ethical principles and the interests of the commission and the other contractors are preserved.

4.1.3 Citation

Researchers who use these corpora for their research must include citations of articles by the contributors the corpus in their publications. The adequate citations can be found in the particular corpus description section in 3.1.

4.1.4 Data storage

Annex II.30 states that the owner (contractor) should provide for adequate and effective protection of knowledge that also applies for speech corpora. Especially sensitive are Meta data of the speakers (see 4.2.2). For the purpose of sharing the corpus, it should therefore be guaranteed by the new party that these data will be treated carefully. The owner is responsible for an appropriate contract.

4.1.5 Protection of participants/speakers identities

Participant data are very sensitive and should be handled carefully. All speakers are entitled to decide whether they allow the storage and distribution of their speech. Participant identities need protection using pseudonyms in documentation and file names. If this is not entirely possible, for example in conversational speech, participants have to be informed that their names will emerge in the corpus and to agree on that. Another possibility to solve the problem of anonymity in conversional speech is the substitution of participant names by a sound. Some data (e.g. patient corpora) and / or country legislation demand stronger restrictions and the researcher in charge is responsible for their adherence.

The contract between the researcher in charge and the speaker is often referred to as *consent form*. This consent form should at least contain information about the aim of the recordings and how the recordings and the participant data will be used in future and the anonymization of participant data. The participants sign that they are aware of this purposes and forgo their rights on the recordings. An example of a consent form is provided at the AMI corpus Web Site (http://corpus.amiproject.org/documentations/pdf/ami_consent_form_081004.pdf/). The consent form provides an opportunity to collect the necessary participant information (see 4.2.2).

4.2 Data structure and documentation

When sharing corpora it is critical that the ‘new’ users are able to work with the data. Thus, an informative and exhaustive documentation is very important. Moreover, all data within the corpus should have a consistent and matching structure.

Keeping these points in mind already during the recordings and data collections might save time in the final stages of corpus construction. The book of Schiel and Draxler (2003) is very useful for structuring the workflow throughout those early stages. After finishing a corpus before its distribution there should be a stage of quality assessment of the data structure and documentation, which is called validation. There will be no external validation by S2S but every researcher should try to validate the content of the corpus before distributing it. For the validation of speech corpora, please see the ELRA standards (van den Heuvel et al., 2000; 2008).

4.2.1 Documentation

Corpora should come with documentation in English. Section 3 of the present paper provides summaries of selected information that are not sufficient for distributional purposes.

According to the ELRA standards (van den Heuvel et al. 2008; see also van den Heuvel et al. 2000; Schiel & Draxler, 2003) the documentation should contain:

- Owner and contact point
- Database layout and media
- Application potential for the SLR
- Directory structure and file names
- Recording equipment
- Design and contents of the recordings
- Coding and format of the speech files
- Contents and format of the annotation files and speech files
- Speaker demographic information
- Recording environments distinguished
- Transcription conventions and procedure
- Lexicon: format and transcription conventions included.

The criteria were developed for corpus resources focusing on speech technology research. There are differences to corpora used for phonetic research and conversational analysis. For the latter, a precise description of the recruitment strategies, the setting of the recordings, and the used materials and more demographic speaker details are of stronger interest (see also 4.2.2). On the other hand, a lexicon might not be available.

4.2.2 Meta Data

In addition to the recordings and annotations, details about these and their collection are very important. The term *Meta data* refers to those kinds of data. Especially for linguistic and psycholinguistic research these data are very important, for example because the realization of speech depends on dialect, social status, education, and other factors of the speakers. Importantly, these details need to be collected with the recording sessions because speakers might not be available afterwards any more. Schiel and Draxler (2003) provide nice checklists for the preparation of a recording protocol.

If the speech was manually annotated, similar protocols should be provided for transcriber details including:

- transcriber ID
- first language
- growing up region
- second language (if relevant for the transcriptions)
- years of practise in second language
- (years) of training in annotation

Meta data for speakers and transcribers are best provided in table form. Importantly, the relationship between speaker and the according sound files, as well as transcriber and the according annotations should be clearly identifiable.

4.2.3 File formats and directory structure

To enhance the usage of the corpus a clearly arranged directory structure and the usage of common file formats (e.g. for the recordings) are necessary for distributional purposes. The structure of the medium containing all files (e.g. CD, DVD, external drive) should also be outlined in the documentation. Table 3 displays an example taken from the ELRA validation manual (van den Heuvel et al., 2000).

Table 3: Example of directory structure for speech data (Van den Heuvel et al., 2000)

\ (root)	The readme file, the copyright file
\...\<DOC>	Documentation files
\...\<INDEX>	Index files, e.g. contents file, corpus contents files, corpus list files, ...
\...\<TABLE>	Speaker, session, recording condition and lexicon tables
\...\<SOURCE>	Any source code supplied
\...\<PROMPT>	Prompt sheet if present (with appropriate sub-directory structure if needed);

4.2.4 File naming

File names should be as informative as possible to facilitate user orientation. Moreover, they should be unique for a given corpus if possible in that the file names cannot mixed up with files of a different corpus. Audio files and annotation files should have the same name. If data files and annotation files are different, there should be a table file providing the information on the relationship of sound and annotation files. The documentation should contain a description of the meaning of file names and their relation to Meta data.

4.2.5 Transcriptions

The conventions, describing the annotation process of a corpus should be distributed together with the corpus. These should contain the following information (see also Van den Heuvel et al., 2000):

- the used tools and file formats
- the procedure of annotation
- the guidelines and instructions for the transcribers
- details on the transcribers: selection, dialect, first language (others if relevant for the transcriptions), practise in phonetics or transcription training, experience in the used transcription tools

- validation/cross-checking/quality assurance procedures of the annotations
- the symbols used in the annotation files together with their explanation.

For corpora built within S2S, it was agreed to transcribe them orthographically with a separate layer for each speaker and one layer for noise. The orthographic transcriptions will largely correspond to the orthography of the language, including punctuation and uppercase nouns. Filled pauses (e.g. eh, hmmm) and clearly audible speech noises will be transcribed. Overlapping talk, broken words, laughter, imitated speech, and unintelligible speech will be marked. The annotations will be split into chunks up to 3 seconds driven by the requirements of automatic broad phonemic transcriptions of spontaneous speech (Schuppler et al., 2008). The noise layer will contain any non-speech noise also when it was produced by a speaker (e.g. clapping hands). Transcription guidelines will be produced on the fly while working on the annotations. The estimated time to annotate one hour spontaneous speech this way was 40 hours.

5 Software and tools

This section provides an overview of selected software for the annotation of speech data and the management and query of corpora. The first section introduces criteria that were used to select the evaluated software. The second section provides an overview of tools to annotate speech. The third section introduces programs to build, manage and query databases and their advantages and disadvantages.

Unfortunately, there is not one software that can be used for all the required tasks.

5.1 Criteria of software selection

To annotate speech data an editor such as notepad and a program playing sounds could be sufficient. However, various tools were developed to facilitate the annotation of speech data. Those tools help to align sounds and text in time, offer visualization of the sounds as oscillogram or spectrograms. During the last years many programs were developed that can be used to annotate speech, analyse speech or create, manage, and query speech databases. This section provides an overview of a selection of tools that might be helpful while working with speech.

There are some criteria that successful tools need to be listed for evaluation:

- The software should be free of charge.
- The software should support different computing platforms: Linux, Mac OS, and Windows. This criterion is important for people working on different computers and computer platform since it allows the use of the same program independent of the operating system.
- The software should be well known and distributed. Tools that are widely distributed are more likely to be known by other S2S fellows that can provide guidance and help for the program.
- The software should be supported and under development. Support is important if a special problem occurs that cannot be solved without help of the developers. The software should also be under development because it is more likely that such software will be improved and can be used in future as well.

Other criteria are preferable but will not be fulfilled by every tool listed below, because no software fulfils all of them.

- The software can cope with audio and video data. At present very little tools to annotate speech or manage speech databases support video recordings. However,

especially for the investigation of the interaction of speaker, non-verbal indicators such as gazes, nods, and other are also important.

- The software should be based on XML. XML format is one of the most recent standards that is very useful for the representation of large complex data structures. Its inherent structure allows the handling of large amounts of data at high speed and offers simple ways to change and expand properties of the data (e.g. named entities).
- The software should be well documented. Manuals are very important for the usability of a software.
- The software should be easy to use and intuitive to handle. Both points are important to get acquainted with new software. For new users and users with little computational experiences graphical user interfaces (GUI) are very important. Handling refers to the (intuitive) understanding of the layout and the names of the commands, which is knowledge due to the use of other software.
- The software should support the automation of processes that will be repeatedly executed, due to either batch processing or scripts.

5.2 Transcription tools

5.2.1 PRAAT

Praat (Boersma, 2001; http://www.fon.hum.uva.nl/praat/) is a very powerful tool for labelling, speech analysis, synthesis and manipulation, and even stochastic learning. It supports multiple layers that can be interval or point tiers. Sounds can be displayed as oscillograms and spectrograms with nearly unlimited zoom into the time domain. Various phonetic-acoustical parameters of the sound can be displayed simultaneously in the spectrogram window (e.g. formants, intensity and pitch contour). This kind of sound representation makes it especially useful for the transcription of small linguistic units, such as broad phonetic transcriptions.

Praat is a well-known and widely distributed annotation tool. It is free software and runs on various computing platforms: Linux, Mac OS, Solaris, and Windows. It is under support (forum: http://groups.yahoo.com/group/praat-users) and will be further developed.

Praat does not support XML but uses its own specific annotation format *TextGrid*. As it is one of the most common tools for speech analysis, many programmes can handle this format.

It comes with an extensive manual available in the help function of the programme or external downloadable html files. The focus of the manual introduction to Praat is on sound analysis and thus the manual is not the ideal getting started guide for transcribers new to the programme. There exist many good Praat tutorials on the web.

The handling of Praat is not very intuitive or user friendly. Although, annotating in Praat is easy, the uncommon design of the program requires some introduction. Since it is widely used, many S2S members can help beginners with Praat. Once familiar with Praat, its handling is easy.

Praat comes with its own very simple scripting language. Moreover, scripts can be "written" by clicking the according buttons and pasting the history. This very useful function facilitates the scripting of automatic processes for persons with no scripting experience. On the internet, there exist many resources of all kinds of Praat scripts that make working with Praat much easier. A collection of online Praat tutorials and script recourses can be found at http://liceu.uab.es/~joaquim/phonetics/fon_anal_acus/herram_anal_acus.html#Praat.

5.2.2 Transcriber

Transcriber (Barras et al., 1998, 2000; http://trans.sourceforge.net/en/presentation.php) is a tool for annotating speech signals. It was designed to transcribe broadcast news. It supports the labelling of several layers such as speech turns of multiple speakers, topic changes or acoustic conditions. Moreover, it allows easy handling of the properties of named entities. The main interface consists of a text editor tool and the display of the waveform. The structure of the program provides easy access to numerous different features and tools facilitating annotation, and the organization and management of metadata, such as speakers, topics or overlapping speech. This design makes it especially useful for orthographic transcriptions of long speech recordings of single or multiple speakers.

Transcriber is widely used, especially for annotating broadcast news. It is free software and works on different computing platforms: e.g. Linux, Windows, Mac OS. It is under support and will be further developed.

The annotation output format is *trs* that is based on XML and can be exported to various other formats.

It comes with manuals in English and French useful for self-studying.

The graphical user interface facilitates intuitive handling of the program and in combination with the manual self-learning is possible. Since it is widely used, there are S2S members familiar with the programme from which beginners can profit.

I have not found anything regarding automatic processing of routines (e.g. changing the properties of one speaker in all files). I am not sure whether this is necessary for the purposes the program is made for.

5.2.3 Wavesurfer

Wavesurfer (http://www.speech.kth.se/wavesurfer/index.html/) is another widely distributed speech analysis and manipulation tool. It can be used to annotate speech data. There exist several plug-ins to configure the tool for transcriptions as well as for video files. It displays oscillogram and spectrogram and allows multiple layers for labels. Similar to Praat it is more useful for broad phonetic annotations than for orthographic annotations of long stretches of speech.

Wavesurfer is well known and widely distributed. It is freely available and runs on various computer platforms: e.g. Linux, Mac OS, Solaris, and Windows. As far as I know, it is not supported anymore and is listed here because it is still widely used.

The different plug-ins support various label formats, e.g. TIMIT, ESPS, HTK, and Phondat. It does not support XML.

The wavesurfer manual is an easy to use getting started guide.

The graphical user interface also facilitates the handling of the program. Since it is widely used, there are S2S members familiar with the programme from which beginners can profit.

It supports automatic processing on the basis of self-written scripts in TCL/TK scripting language.

5.2.4 ELAN

ELAN (EUDICO Linguistic Annotator; http://www.lat-mpi.eu/tools/elan/) is an annotation tool especially developed for multimedia recordings. It supports multiple annotation layers that can be hierarchically interconnected. Up to four videos can be associated with a single annotation document. There are several different modes to display video, waveform and annotations. Moreover, ELAN provides the option to query one or more annotations using regular expressions. The software is most useful for orthographic transcriptions of long sound and video files.

ELAN is widely used. It is free software and runs on various computer platforms: Linux, Mac OS, and Windows. It is under support forum http://www.lat-mpi.eu/tools/elan/elanforum/) and will be further developed.

The annotations are stored in eaf format that is based on XML. The use of XML standard facilitates the creation and modification of named entities (here *linguistic types*). ELAN can handle formats from other various other speech related applications, such as

Shoebox/Toolbox, CHAT, Praat, Transcriber (import only), or csv/tab-delimited text files. It supports various video formats: e.g. Windows Media Player, QuickTime or JMF (Java Media Framework).

The ELAN manual is very detailed. On the web, simpler getting started guides for ELAN can be found, for example: http://www.lat-mpi.eu/tools/elan/thirdparty.

ELAN comes with a graphical interface, which facilitates its use. However, ELAN is a very complex and powerful tool with an incredible number of different options that might be exhausting at first glance. At present, only few people within S2S are using the tool.

I have not found anything regarding automatic processing of routines (e.g. changing the properties of one speaker in all files) or scripting / using script to facilitate routines. Moreover, I am not sure whether this is necessary for the purposes the program is made for.

ELAN allows querying one or more annotation files in eaf format. This function is quite powerful supporting various options and regular expressions. However, ELAN is not listed in the next section since it is not a database management tool. Every annotation file that is not in eaf format (for example TextGrid) has to be imported to ELAN manually. To query more than one eaf file, all files need to be manually selected and only one utterance can be queried in a serial way.

5.3 Database construction and query tools

5.3.1 Emu

The Emu Speech Database System (Bombien et al., 2006; http://emu.sourceforge.net/) is "a collection of software tools for the creation, manipulation and analysis of speech databases." Emu databases provide a hierarchical structure of annotation layers. These can be queried in serial and hierarchical order or combinations of both. The Output of the query is a table of search results, file and times. The results can be saved as a table. Various implemented tools allow the direct and easy installation of databases, display of spectrograms and waveforms, analysis of the signal, the implementation of EPG data and segmentation. Due to Emu-R, processing and query can be conducted from the R programme.

Emu is free software and releases are available for: Linux, Mac OS, and Windows. The software is still supported and under development.

Emu is not based on XML, which is a drawback of the software. However, it is included in this manual because at present it is one of the most elaborated corpus query tools. Emu can handle Praat TextGrid transcriptions as well as wavesurfer EPSP output.

The official Emu manual is for version 1.9 and somewhat outdated. However, at the webpage are various video tutorials to get started with the programme. Furthermore, in Harrington (in prep) there is a good introductory chapter. The official manual is of interest for advanced knowledge such as writing autobuild scripts or the query syntax.

Since version 2.0 Emu has a graphical user interface that facilitates its handling. Still it is a complex tool and creating new corpora often results new in errors.

Emu supports automatic processing in TCL/TK scripting language. Knowledge of TCL/TK is also very useful for automating database construction and the query definition.

A very interesting feature of Emu is the interface to R. Due to Emu-R Emu-databases can be directly queried in R where further processing of the output is possible. Wavesurfer is also one of the implemented tools of Emu allowing signal processing directly from Emu or R.

Construction of new databases. The construction of new databases is relatively easy with the *Convert Labels* tool that can process Praat TextGrid files. Afterward the template file of the new corpus has to be adapted as described in detail in the manual. The last step involves the creation of the hierarchy files and takes some time. Errors can occur due to all of these steps. Some of these might be fixed by the Perl scripts available at the S2S wiki.

Query. The query can be conducted using the graphical query GUI or directly via the command line using TCL/TK syntax for experienced users. Queries can be serial (e.g. all phones a) or hierarchical (e.g. all words with phones a) order. Combinations and long serial searches are possible, such as all phones a before phones b after phone c in words starting with phone d and so on. Further specification of which databases or corpora are to be queried is possible. The outcome of a query is a table consisting providing details of where to find the utterance and starting and ending time of the requested information (in the example above only the times of the phone a will be provided). The results can be saved in txt format. Further selection of utterances within the query is not possible. The results can be combined with the acoustical analysis tools the program additionally provides. All of this can be done from R.

5.3.2 AG-SpIT

The Annotation Graph for Spoken Italian Tool (Savy et al. 2006; http://www.parlaritaliano.it) is a database generation, management and query tool. The data are structured as Annotation Graph that facilitates the implementation of time-aligned and text-aligned linguistic data. Utterances can be displayed as waveform together with their annotation at various levels. Databases can be queried in hierarchical or serial order but not in a combination of both. The Output of the query is a sortable table of the query results, file and

times that can be stored as csv file. A drawback of the software is the difficult creation of new databases. An advantage is the speed of queries due to the use of XML standards. AG-SpIT is most useful for queries in existing databases (such as CLIPS) or large databases consisting of various time-aligned and text-aligned annotations.

AG-SpIT is free software and runs on Linux platforms and Windows XP. The software is supported and under development.

AG-SpIT supports annotations in TIMIT style and file names of the corpus according to TIMIT style are required. These annotations are transformed into a database in XML format that supports queries of one or more large databases at a high speed. Parts of the CLIPS corpus can be downloaded already in the AG-SpIT database format.

There is no English manual for AG-SpIT but at the start of the programme there occur instructions on the screen. The lack of a manual is especially critical during the construction of new corpora and the missing specification of the demands to annotations and file names.

The GUI is very user friendly and self-explaining. The same applies for the query window that is clearly represented and facilitating queries in existing corpora.

At present, the software does not support the use of regular expressions in the query or automatic enrichment of the corpora due to scripting. I did not understand this quite well: Users knowing TCL/TK scripting language might be able to write scripts to change the XML files.

Construction of new databases. Very demanding is the construction of a new database. Only TIMIT style annotations are supported. At the S2S wiki are scripts to transfer Praat annotations to TIMIT annotations. For all of the corpus files the naming rules of the TIMIT file name conventions must strictly be adhered. These are structured as follows: corpusname_i1#i2, in that i1 and i2 are indices of letters or numbers of any length. The software displays brief instruction for the database construction, but errors are not very well documented and the construction of new corpora can be very frustrating.

Query. The GUI of the query is easy understandable. Queries can be conducted in serial (e.g. all phones a) or hierarchical (e.g. all words b with phones a) order. There are further options to concretize the query. At present it s not possible to query for more than one serial utterance, such as all phones a before phone b. Further specification of which databases or corpora are to be queried is possible. The outcome of a query is a sortable table consisting of details of where to find the utterance and starting and ending time of all utterances involved in the query (in the example above details would be provided for the phone a and the words b). The results can be saved in csv format. Further selection of utterances within the query is possible.

6 References

Amdal, I. & Svendsen, T. (2006). FonDat1: A Speech Synthesis Corpus for Norwegian. *Proceedings of LREC-2006*. Genova, Italy.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC map task corpus. *Language and Speech,* 34, 351–366.

Barker, J., Cooke, M. (2007). Modelling speaker intelligibility in noise. *Speech Communication*, 49(5), 402–417.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (1998). Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech, *First International Conference on Language Resources and Evaluation (LREC)*, 1373–1376.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2000). Transcriber: development and use of a tool for assisting speech corpora production, *Speech Communication special issue on Speech Annotation and Corpus Tools*, 33(1–2).

Boersma, P. (2001). *Praat, a system for doing phonetics by computer*. Amsterdam: University of Amsterdam. Web Site: http://www.fon.hum.uva.nl/praat/.

Bombien, L., Cassidy, S., Harrington, J., John, T., & Palethorpe, S. (2006). Recent Developments in the Emu Speech Database System. *Proceedings of the Australian Speech Science and Technology Conference*, Auckland, December 2006.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am*, 120(5), 2421–2424. Online version: http://www.dcs.shef.ac.uk/~martin/jasagrid.pdf.

Cooke, M., Scharenborg, O. (2008). The Interspeech 2008 Consonant Challenge. *Proceedings of Interspeech*, Brisbane, Australia, September 2008.

Crawford, M. D., Brown, G. J., Cooke, M. P., & Green, P. D. (1994). Design, collection and analysis of a multi-simultaneous-speaker corpus, *Proc. Inst. Acoustics*, 16(5), 183–190.

ELRA. Web Site: http://www.elra.info/.

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface*. Utrecht: LOT.

FonDat. Web Site: http://www.iet.ntnu.no/projects/fonema/index_eng.php/.

García Lecumberri, M. L., Cooke, M., Cutugno, F., Giurgiu, M., Meyer, B. T., Scharenborg, O., van Dommelen, W., & Volín, J. (2008). The non-native consonant challenge for European languages. *Proceedings of Interspeech*, Brisbane, Australia.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus.* Linguistic Data Consortium, Philadelphia.

Goldwave. Web Site: http://www.goldwave.com/.

Howard, S. J. (2001). The realisation of affricates in a group of individuals with atypical speech production: a perceptual and instrumental study. *Clinical Linguistics & Phonetics*, 14, 133–138.

Howard, S. J. (2004a). Compensatory articulatory behaviours in adolescents with cleft palate: Comparing the perceptual and instrumental evidence, *Clinical Linguistics & Phonetics*, 18(5), 313–340.

Howard, S. J. (2004b). Connected Speech Processes in Developmental Speech Impairment: Observations from an Electropalatographic Perspective, *Clinical Linguistics & Phonetics*, 18(6-8), 407–417.

Howard, S. J. (2007). The interplay between articulation and prosody in children with impaired speech: observations from electropalatography and perceptual analysis. *International Journal of Speech-Language Pathology*, 9(1), 20–35.

Karlsen, B. L., Brown, G. J., Cooke, M. P., Crawford, M. D., Green, P. D., & Renals, S. J. (1998). Analysis of a multi-simultaneous-speaker corpus, In D. F. Rosenthal & H. G. Okuno (Eds.), *Computational Auditory Scene Analysis* (pp. 321–333). Mahwah, NJ: Lawrence Erlbaum.

Kiel-Corpus. Web Site: http://www.ipds.uni-kiel.de/ipds/pub.htm/.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Machač P., Skarnitzl R. & Volín J. (2007): Interlabeller agreement in segmental boundary placement. In R. Vích (Ed.), *17th Czech-German Workshop - Speech Processing* (pp. 57–61), Prague.

Niebuhr, O., Clayards, M., Lancia. L., & Meunier, C. (2008). *Place Assimilation in Sibilant Sequences - Comparing French and English*. Poster presented at the 4th S2S workshop, Prague, Czech Republic.

Niebuhr, O., Meunier, C., Lancia, L. (2008). On place assimilation in French sibilant sequences. *Proceedings of the 8th International Seminar on Speech Production*, Strasbourg, France.

Ogden, R. (2006). Phonetics and social action in agreements and disagreements. *Journal of Pragmatics*, 38, 1752–1775.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech (2nd release)*, Retrieved from

Columbus, OH: Department of Psychology, Ohio State University (Distributor). Web Site: http://www.buckeyecorpus.osu.edu/.

Plug, L. (2005). From Words to Actions: The Phonetics of Eigenlijk in Two Communicative Contexts. *Phonetica*, 62, 131–145.

Savy, R. Cutugno, F., & Crocco, C. (2006). Multilevel corpus analysis: generating and querying AGset of spoken Italian (SpIt-MDb). *Proceedings of 5th International Conference LREC, Paris, ELRA*, 1654–1659.

Scharenborg, O., & Cooke, M. (2008). Comparing human and machine recognition performance on a VCV corpus", *Proceedings of the workshop on Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark.

Schiel, F., & Draxler, C. (2003). *The production and validation of speech corpora*. Bavarian Archive for Speech Signals. München: Bastard Verlag. (URL: http://www.phonetik.uni-muenchen.de/Forschung/BITS/TP2/Cookbook/, http://www.phonetik.uni-muenchen.de/Forschung/BITS/TP1/Cookbook/)

Schuppler, B., Ernestus, M. , Scharenborg, O. & Boves, L. (2008). Preparing a corpus of Dutch spontaneous dialogues for automatic phonetic analysis, *Proceedings of Interspeech, Brisbane, Australia*, 1638–1641.

ShATR. Web Site: http://www.dcs.shef.ac.uk/spandh/projects/shatrweb/.

Strand, O. M., Svendsen, T., & Amdal, I. (in prep). RUNDKAST: A Norwegian Broadcast News Speech Corpus.

The EMU Speech Database System. Web Site: http://emu.sourceforge.net/.

Torreira, F. & Ernestus, E. (subm). The Nijmegen Corpus of Casual French.

Torreira, F. & Ernestus, E. (in prep). The Nijmegen Corpus of Spontaneous Spanish.

Transcriber. Web Site: http://trans.sourceforge.net/en/presentation.php/.

Van den Heuvel, H., Boves, L., & Sanders, E. (2000). Validation of content and quality of existing SLR: Overview and methodology. ELRA Technical report D1.1.

Van den Heuvel, H., Iskra, D., Sanders, E., & De Vriend, F. (2008). Validation of spoken language resources: an overview of basic aspects. *Language Resources and Evaluation*

Volín, J., Skarnitzl, R., Machač, P., Janoušková, J., & Veroňková, J. (2008). Reliabilita a validita popisných kategorií v Pražském fonetickém korpusu. In M. Kopřivová & M. Waclawičová (Eds.), *Čeština v mluveném korpusu* (pp. 249–254). Praha: Nakladatelství LN / ÚČNK.

Appendix A: Selected Materials

The Emory story corpus

This is the tale of William Emory.
He was a brave and loyal man
who sought adventure and knowledge.
He made great contributions to science,
yet history has forgotten him.

12020
1. Topic – Emory was born in eighteen-ten.
Addition – It was a year that would live in memory.
Elaboration – The Napoleonic wars continued in Europe,
Continuation – annually seeming to gain momentum.
Addition – Meanwhile in England George III was declared insane.

12302
2. Topic – Emory had a conventional youth.
Addition – His mother died when he was young
Continuation – and his father raised him and six siblings.
Addition – They grew up in a cottage near Winchester.
Elaboration – Annual rent was due on the cottage.

12023
3. Topic – The family was very poor,
Continuation – and so Emory had to work.
Elaboration – But his father kept it to a minimum.
Addition – Emory still had little time to play.
Elaboration – He worked constantly to support his family.

10223
4. Topic – Manual labor bored him terribly,
Continuation – and he longed for a more fulfilling life.
Addition – He played mental games while he worked,
Continuation – to try to test himself on his memory.
Addition – He also dreamed of exploring the Amazon.

10202
6. Topic – He knew as an explorer he should also be a scientist.
Elaboration – In the Amazon he would investigate many new things.
Addition – He became curious about everything he saw,
Continuation – loyally devoting his energy to science.
Addition - His discoveries were applauded by his family.

12202
7. Topic – At an annual competition he would prove his worth.
Addition – At 10 he won a contest to invent a liniment.
Addition – At 12 he discovered a new local mineral.
Addition – And an animal-taming project won another prize.
Addition – After that he was asked to stop entering.

12222
44. Locally he was now well-known.
His healing liniment became popular, especially among area farmers.
They used it to soothe aches and pains both in their cattle and on themselves.

12020
5. Topic – University was always a desire for Emory.
Elaboration – He was sure that without an education,
Continuation – minimal opportunities would be in his reach.
Addition – To this end, he studied constantly.
Elaboration – At a minimum he read a book a week.

12023
45. Topic – He made friends with the parish priest.
Addition – The father traded Emory lessons
Continuation – for some annual help in his garden.
Addition – He also promised to aid Emory
Continuation – by paying his university fees.

12020
8. Topic – At 17 the young man left for Cambridge.
Elaboration – Emory was finally following his dream.
Addition – He later fondly recalled the journey.
Elaboration – Because he was poor he had to walk,
Continuation – and he memorized every step on the way.

12230
9. Topic – Emory was in awe when he reached Cambridge. Elaboration – To him it seemed to be royally appointed, Continuation – and it would be like a beacon in his memory.
Addition – The King's College chapel inspired him,
Continuation – minimal though his chances to visit were.

12020
10. Topic – But Emory was disappointed by university.
Elaboration – Annual examinations were dull,
Continuation – and the minimal effort required was vexing.
Addition – He had enemies among the students as well.
Elaboration – His impoverished past was a basis for jokes.

12023
11. Topic – Still the young man remained mannerly.
Elaboration – He tried to be kind to everyone.
Addition – When an animal escaped from a laboratory,
Continuation – he spent hours helping to chase it down.
Addition – He was annually named "Most Helpful Student."

12202
12. Topic – His life changed drastically one April day.
Addition – A visitor came to speak at his college.
Elaboration – Mr. Rinnering was an explorer and voyager.
Elaboration – In the Amazon he had discovered many things.
Addition – Emory was determined to join his team.

12342
13. Topic – Rinnering wasn't looking for more help.
Addition – His journeys always had special teams.
Elaboration – If a man or a youth wanted to join,
Continuation – he needed very particular qualifications.
Elaboration – At the moment he had everyone he needed.

12304
14. Topic –Emory was determined to go on the journey, Continuation – so he made a list of his skills for Rinnering.
Elaboration – Mineral research was a unique ability.
Elaboration – Emory was well-informed in that field.
Elaboration – He was sure it must be useful in the Amazon.

10234
15. Topic – He needed an introduction to Rinnering.
Elaboration – If the explorer were to take him seriously,
Continuation – he required a good recommendation.
Elaboration – A few kind words were a bare minimum.
Addition – Rinnering would not be easily impressed.

12032
16. Topic – An important professor organized the meeting
Continuation – so that Rinnering was sure to come.
Addition – Emory would be there as though by chance.
Addition – With his mineral research in hand
Continuation – Emory waited with anticipation.

10220
42. Topic – When Rinnering arrived that day,
Continuation – animals and minerals of the Amazon
Continuation – were already under discussion.
Addition – Emory got to demonstrate his knowledge
Continuation – as well as his good work ethic.

10020
17. Topic – Mr. Rinnering was delighted with him.
Addition – In a minimum of time he arranged to hire him.
Elaboration – Emory's skills could then serve his next mission.
Elaboration – In the Amazon there were mysteries to be solved,
Continuation – and there would be plenty of need for Emory.
12340
43. On a sunny day they departed from Bristol.
Their ship, the Zinnia, was newly made.
She was not very big, but she was fast.
They hoped to arrive in thirty days,
or thirty-five if the winds were bad.

12320
18. Topic – On the voyage there was time for practical study.
Addition – Rinnering pored over maps and charts.
Addition – Other men prepared for capturing animals.
Addition – Emory reinvented his prizewinning liniment.
Elaboration – Minimal volume with maximum effect was his goal.

12223
46. Topic – He felt quite lonely on the journey.
Elaboration – Mr. Emory was ignored by the others,
Continuation – who knew each other already.
Elaboration – They mostly left him to himself,
Continuation – and assumed that he was ignorant.

12030
19. Topic – The ocean was calm until the last week.
Addition – Then a storm troubled the voyagers.
Addition – Every day was a battle to survive.
Elaboration – The storm became a terrible enemy,
Continuation – in the manner of a hungry monster.

12230
20. Topic –The explorers finally reached land safely.
Addition – Emory was elated to be in the Amazon.
Elaboration – The storm they had survived left his memory.
Addition – Animals and strange plants surrounded him.
Elaboration – It was a whole new world to Emory.

12323
21. Topic – They needed to set up a campsite for safety.
Elaboration – Rinnering knew the forest was their enemy.
Addition – At a minimum they needed to build a fire
Continuation – so that animals would stay away.
Addition – Emory and the others started to work.

12202
22. Topic – Suddenly there was a huge commotion.
Elaboration – Their companions raced back towards them.
Elaboration – Animals also flooded the camp.
Addition – Rinnering gave orders to stay together.
Elaboration – Enemy creatures might be afoot.

12323
23. Topic – Soon they saw the source of the trouble
Continuation – as an animal the size of a ship appeared.
Elaboration – It was twenty feet high at a minimum.
Addition – Leaves and branches filled its fur.
Addition – There were men in a panic all round it.

10222
24. Topic – Emory's heroic behavior saved the day.
Elaboration – He yelled to distract the animal
Continuation – which gave the others time to escape.
Elaboration – Mr. Rinnering and the other men hid
Continuation – while Emory faced down the creature.

12020
25. Topic – With great shouts and bellows,
Continuation – Emory forced the creature away.
Addition – It seemed to be very timid
Continuation – for an enemy of such a great size.
Elaboration – Even the snapping tree branches scared it.

10203
26. Topic – Such bravery could not be ignored.
Addition – Mr. Emory gained everyone's respect
Continuation – despite his youth and inexperience.
Elaboration – Their behavior was more mannerly
Continuation – and they welcomed his presence among them.

12030
27. Topic – Emory was soon seen as an expert.
Elaboration – Mr. Rinnering himself sought his advice.
Addition – As the men's respect for his knowledge grew,
Continuation –liniment became quite popular.
Addition – In exchange they helped him research his minerals.

12202
28. Topic – Meanwhile they continued their explorations.
Addition – In the Amazon there was much to discover,
Continuation – and they had very limited time.
Addition – Experimental tasks fell to Emory,
Continuation – as he was asked to analyze new findings.

12020
29. Topic – Soon the explorers needed to depart.
Elaboration – It was a sad goodbye for Emory
Continuation – who felt he'd only just begun many tasks.
Elaboration – Amazon research clearly required more time.
Addition – Mr. Rinnering promised him a return trip.

12032
30. Topic – Tragedy struck on their return journey.
Elaboration – A storm wrecked the Zinnia,
Continuation – and many men were lost.
Elaboration – Mr. Emory was fortunately saved,
Continuation – but others, like Rinnering, were less lucky.

12030
31. Topic – When the survivors returned to England,
Continuation – the other men told of Emory's bravery.
Elaboration – With the story of the massive beast,
Continuation – Emory became an instant hero.
Elaboration – He even eclipsed the late Rinnering.

10203
32. Topic – Mr. Emory was commended royally.
Elaboration – The King himself thanked him with a speech
Continuation – and an invitation to dinner at the palace.
Addition – A holiday was proclaimed in his honor
Continuation – which children hoped would become annual.

12020
33. Topic – Emory's journeys were all the rage,
Continuation – winning him fame in many places.
Elaboration – Soon his name was a household word
Continuation – for a win or a success in any venture.
Addition – It was a bit overwhelming for Emory.

10202
34. Topic – He was asked to write a book
Continuation – that would contain his every memory.
Elaboration – The book would be called the Adventurer's Manual.
Addition – It would include his maps of the Amazon
Continuation – and a mineral guide to the region.

10220
35. Topic – He also gained other financial benefit
Continuation – when his healing liniment was sold.
Addition – Manual production was too slow for demand.
Elaboration – Liniment took days to prepare
Continuation – due to the careful mixing of minerals.

10230
36. Topic – He planned to return to the Amazon,
Continuation – to continue his travels and studies.
Elaboration – Mr. Emory's acclaimed experiences
Continuation – could continue to grow annually.
Addition – Animals and land remained to be studied.

10202
37. Topic – He advertized for a company of men.
Addition – Amazon experts raced to join him,
Continuation – just like they had followed Rinnering.
Addition – He could pick and choose from among them.
Elaboration – Mr. Emory's team was the cream of the crop.

12023
38. Topic – However, financing was a problem.
Elaboration – World exploration was now so popular
Continuation – that everyone was planning journeys.
Addition – Though Emory wrote to many financiers,
Continuation – Amazon work was no longer funded.

12020

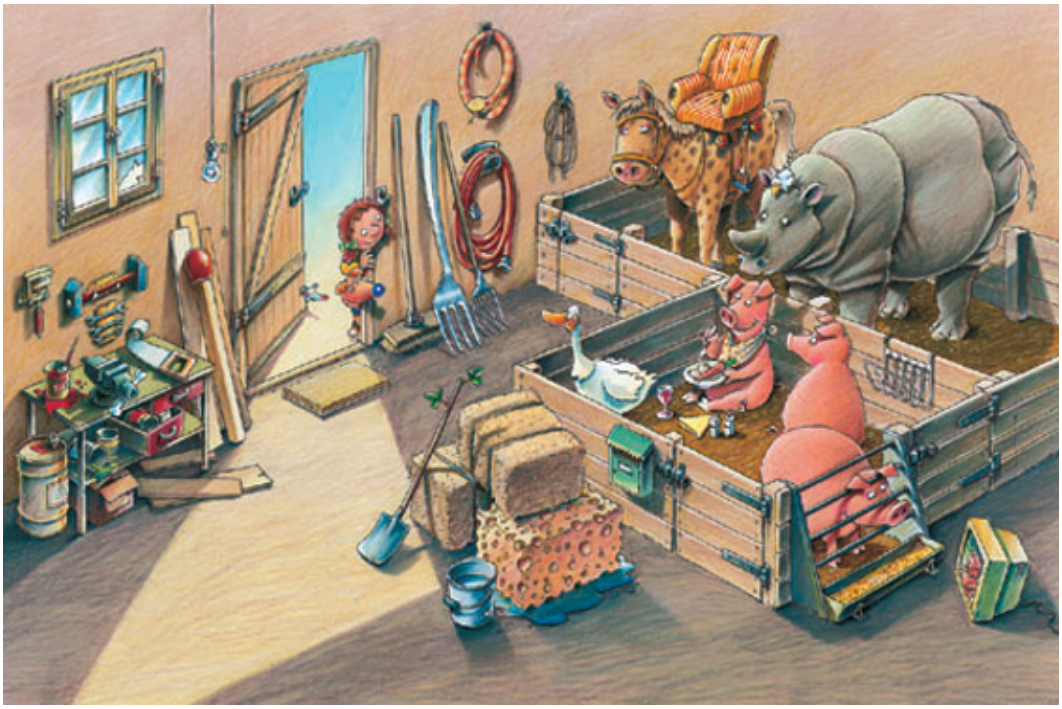
39. Topic – Emory was very disheartened
Continuation – as the way closed to the Amazon.
Elaboration – He tried to remain cheerful,
Continuation – but his happiness was minimal.
Addition – He felt as though his dreams were slipping away.

10202
40. Topic – He decided to take action
Continuation – to help other men like him.
Addition – He created a Travellers' Trust
Continuation – to support those who wished to explore.
Elaboration – It had annual funds available.

10203
41. Alas today Emory has been forgotten.
When the money for his Trust ran out,
interest turned to other people,
and so history cheated him of his rightful dues.
It was a tragic fate for poor Emory.

12002

Kachna corpus



Prague Phonetic Corpus

Phonetically balanced sentences (contain every Czech phoneme at least ones) and are easy to read and pronounce (Janota & Palkova, 1991).

Babička se zeptala Petra:

"Petříku, máš už napsanou úlohu?

Co máte psát?"

Petr odpověděl:

"Já musím napsat větu, že maminka má nové červené boty.

Až budu hotov, dojdu ti do lékárny pro ten neuralgen.

A potom bych byl na fotbale.

Včera jsem dal tři góly.

Neboj se, dám pozor na auta."

Prose text of Ivan Olbracht "Podivné přátelství herce Jesenia" (1919), complicated pronounciation (lexically and syntactically sophisticated).

Hluboko pod ním ležela Praha.

Svítila jediným velkým světlem a uvnitř něho tisíci drobnými.

Velkým světlem v sloupu vyzařovaným vstříc obloze a malými, která bíle planula podél Vltavy a na pásech jejích mostů, vlnivě se obrážela v řece, zářila dlouhými dvojstupy příč městem a mihotala se červenavými tečkami na rozhraní noci.

Světla uprostřed světla, jiskry utkvělé v plameni.

Na hlavách Prahy ležel sníh.

Narrative on cartoon strip of Josef Lada (used since 2002).