3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

The Translanguage English Database (TED) [†]

L.F. Lamel+, F. Schiel++, A. Fourcin+++, J. Mariani+, H. Tillmann++
+LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
++Institute of Phonetics, Univ. Munich, Theresienstr. 3, 80799 München, GERMANY
+++University College London, Wolfson House, 4 Stephenson Way, London, NW1 2HE, UK
euro@phonetics.ucl.ac.uk

10.21437/ICSLP.1994-451

ABSTRACT

The Translanguage English Database is a corpus of recordings made of oral presentations at *Eurospeech93* in Berlin. The corpus name derives from the high percentage of presentations given in English by non-native speakers of English. 224 oral presentations at the conference were successfully recorded, providing a total of about 75 hours of speech material. These recordings provide a relatively large number of speakers speaking a variant of the same language (English) over a relatively large amount of time (15 min each + 5 min discussion) on a specific topic. A subset of speakers were recorded with a laryngograph in addition to the standard microphone. A set of Polyphone-like recordings were made, for which a subset also had a laryngograph signal recorded. These recordings were made in English and in the speaker's mother language.

In addition to the spoken material, associated text materials are being collected. These include written versions of the proceedings papers and any oral preparations texts which were made available. The text materials will provide vocabulary items and data for language modeling. Speakers were also asked to complete a short questionnaire regarding their mother language, any other languages they speak, as well as their knowledge of English.

INTRODUCTION

In this paper we report on a potentially important aspect of speech corpora work in which the problems associated with differences between speakers are enhanced by their use of a language which is not their own. The Translanguage English Database (TED corpus) is the result of recordings made of oral presentations at *Eurospeech93* in Berlin. In fact, the original idea of TED came about after informal discussions between J. Mariani and J.M. Dolmazon (ICP, France) during the Chaviari COCOSDA meeting following *Eurospeech91*. A formal presentation of the project was made by J. Mariani at the COCOSDA meeting following *ICSLP92* in Banff, after which planning continued informally amongst the partners, becoming finalized with the acceptance of the LRE EUROCOCOSDA project, which supports the European participation in COCOSDA actions.

The corpus name derives from the high percentage of oral presentations given in English by non-native speakers of English. In addition to the recording of the speeches (TEDspeeches), 2 auxilliary sets of recordings were made. The first consists of a subset of speakers who were recorded with a laryngograph (TEDlaryngo) in addition to the standard microphone. The second set are Polyphone-like recordings made for 55 speakers (TEDphone), a subset for whom a laryngograph signal was simultaneously recorded. The TEDphone recordings were made by the speakers in English and in their mother language. All the recordings, with the exception of the telephone channel of the TEDphone recordings (which were directly sampled on a speech server in Munich), were made on digital audio tapes (DAT) cassettes and are in the process of being digitized (by University of Munich) for production on CDROM.

Of the 287 oral presentations, 224 were successfully recorded on two channels. The primary channel contains the signal from the speaker's microphone, and the secondary channel is a mix of the signal from all the other room microphones. In total there is about 75 hours of speech material for each channel.

These recordings provide a relatively large number of speakers speaking a variant of the same language (English) over a relatively large amount of time (15 minutes each + 5 minutes discussion) on a specific topic. The recorded speakers can be subdivided into two sets: native speakers of English with their different dialects (L1-variation) and non-native, speakers of English as a foreign language (L2-variation). The TED corpus is thus a multi-dialectal and multi-accent speech corpus. The recording situation - that of making an oral presentation at a conference - is both "natural" to some degree "stressful". While this data will certainly present a challange to any existing speech recognizers for English, it can also be used for a range of other research activities, including detecting accents in English, automatic accent adaptation (which may entail automatic selection of accent-specific pronunciation models), speech training and test material for language learning (such as for the ESPRIT project SPELL), as well as data for speech coding applications (teleconferences).

In addition to the spoken material, an associated text corpus (TEDtexts) is being collected. TEDtexts will include written versions of the proceedings papers as well as any texts or notes for oral preparations that were supplied by the authors. This text corpus will provide related material for language modeling. By including text-only versions of all the articles published in the proceedings, not only

[†] This work is partially funded by the LRE project 62-057 EUROCOCOSDA.

those for oral presentations, will increase the amount of text data available for language modeling and the known vocabulary for the subject domain. Speakers were also requested to complete a short questionnaire regarding their mother language (L1) and any other languages they speak (second language learning (L2)), as well as their knowledge of English. The texts and questionnaires have been acquired by email.

There are many considerations that must be dealt with to carry out the recordings, including overall coordination, task subdivision, notifying participants and obtaining their permissions, hardware considerations (recording media, recording center, microphones, transmission channels), and personel to monitor the recordings. In the remainder of this paper we provide more detailed information about making the actual recordings and the processing work that has been carried out since. We also try to provide hints and pointers that can help others who wish to record similar style corpora.

RECORDINGS AT *EUROSPEECH93*, BERLIN

The recordings took place at the *Eurospeech* Conference in Berlin in September 1993. The aim was

- to record as many oral presentations as possible during the conference (TEDspeeches),
- to make simultaneous laryngograph recordings with a small subset of speakers (TEDlaryngo), and
- to do additional polyphone-like recordings for a subset of speakers (TEDphone)

Preparatory Work

The vast majority of the organization was done by the University of Munich in coordination with the *Eurospeech* organizers Berlin, and members of the EUROCOCOSDA consortium.

Material Setup

A DAT recorder was needed for each conference room. 6 DAT recorders of different types were rented so as to avoid confusion, which caused some minor problems for the operators. The recordings were made in a central recording room, where the signal lines from the different session rooms were mixed. All connections were tested the day before the conference started, with each channel individually check for noise and to ensure proper connection. The wireless microphone was recorded on the right channel, on the left channel the mixed signal of all the microphones including the room microphone, microphone of the session chair and the wireless microphone. Unfortunately the wireless microphones supplied from the conference center *Haus am Koelnischen Platz* were not all of the same type. Headphones were available to monitor the recordings.

The recordings were made at 48kHz. 120 minute DAT tapes were used to record entire sessions on the same tape. The tapes were labelled with the session number in advance to minimize confusion. The laryngograph recordings were done on a portable SONY DAT TCD-3. All the equipment was provided by UCL and was battery operated and carried by the speaker.

Coordination/Staffing

The recordings were coordinated by the University of Munich, who arranged the schedules for the operators in the central recording room and in the session rooms, and trained the operators before the conference. At the end of each day, the tapes of the recordings were arranged for safe-keeping.

In the central recording room the tapes were changed and all machines were set up prior to each session. Recording started 5 min before the start of the session. The signals were frequently monitored by the operators (2-3) with headphones to detect problems. 5 min after the session ended the recordings were stopped. The operators kept a log of all events (malfunctions, noise, change of microphones, etc.)

An operator was assigned to each session room. The operator was responsible for checking the functioning of the wireless microphone before each presentation, and to help attach the microphone to the speaker in the same position. It was useful to have a few extra operators available to deal with problems that arise, as well as to transmit messages to the control room. A source of fresh batteries and extra tapes should also be readily available.

Special trained operators were required for the laryngograph recordings These were carried out by UCL under the supervision of A. Fourcin. A few minutes before each presentation the operator had to find the speaker to attach, test and calibrate the equipment. Each speaker was recorded on a separate tape. The operator remained with the speaker during the presentation, and then accompanied the speaker to record the TEDphone data with the same equipment.

Permissions

According to German law, permission of the individual speakers is required to record them. Obtaining permissions from speakers is in fact becoming a general practise in the creation of large corpora, a consideration often not taken into account in early corpora creation efforts. The following permission forms were mailed out several months before the conference.

> *'I hereby agree to the recording of my presentation during the Eurospeech 1993 in Berlin.'*
> *Session:* *Paper:*
> *(signature)*

> *'I am also willing to be recorded with laryngograph electrodes (non-invasive electrodes) during my presentation as above.'*
> *(signature)*

In order to encourage participation EUROCOCOSDA promised every speaker who agreed to be recorded that he can get the whole database at a minimum price (ie. production costs). In response to this letter over 50% of the speakers accepted to be recorded. Missing signatures were obtained during the conference when possible, or after the fact by email or by return of the speaker questionnaire.

TED LARYNGOGRAPH RECORDINGS

There are substantial advantages for speech signal analysis, noise resistant processing and subsequent labelling and annotation, when the acoustic pressure signal is acquired simultaneously with the larynx closure waveform obtained from an electro-laryngograph. In consequence, in addition to providing the option of making individual speaker acoustic recordings during the oral presentations, special wearable equipment was made available for two sensor speech data acquisition. Each volunteer speaker for this mode of data gathering wore a lapel-mounted, electret omni-directional pressure sensitive microphone (Sennheisser MKE-10) giving the speech pressure signal, Sp, and an elastic neckband which supported the two gold plated electrodes of a miniature electro-laryngograph which gave the vocal fold closure waveform, Lx. These two signals were respectively recorded on the left, channel 1, and right, channel 2, tracks of a 16 bit digital analogue tape, running at a sampling frequency of 48 kHz. The microphone had a 3 dB frequency response from 60Hz to 12kHz and the Lx signal had a 3dB frequency response range extending from 20 Hz to 12 kHz. Time calibration, and phase correction for Lx, were provided by a Laryngograph Ltd "artificial neck" crystal controlled calibrator. The DAT recorder (Sony TCD-D3), microphone pre-amplifier, and the miniature field laryngograph (Laryngograph Ltd) were all battery powered and carried in small strap-worn pouches across the speaker's shoulder so as not to restrict either movement or gesture.

PROCESSING OF DATS

The recordings are organized into three subcorpora, TEDspeeches, TEDlaryngo and TEDphone. All speeches for which a written permission was available by the end of Feb 1994 and were technically usable, were subsequently processed. Technically usable means that the right channel was without permanent noise and had no interruptions longer than 1s. The processing involved splitting the 2 channels on the DAT, digital filtering of each channel to 8kHz, and downsampling to 16kHz. During the transfer from DAT to PC, the speech was monitored and relevant events marked, including demonstrations. Speeches are processed from beginning of speech ('Good morning Ladies and ...') to end of speech ('Thank you'). Discussions are processed from end of speech (Chair: 'Any questions') to end of discussion (Chair: 'Thank you again...'). TEDspeeches currently has talks from 188 speakers, corresponding to 47h of spoken material. TEDlaryngo currently has 11 speakers.

The signal files were then compressed using SHORTEN[1], and NIST SPHERE header files and SAM label files were created. This was a compromise decision taken to provide compatability with the two widely used file formats. A simple batch routine was written to uncompress the file (removing the header and copying it and the SAM label file to the current directory on the host PC), thus providing both (pseudo) SAM and NIST compatability.

The data were backed-up on ExaByte tape, and in-house master CDROMs (ISO-9660 format) made of the primary channel.[2]

For the TEDphone subcorpus there are two different types of recording: the microphone and laryngograph recordings on DAT, as for TEDlaryngo, and the telephone data, recorded digitially via a connection to a server in Munich, with each utterance in a separate file. A total of 55 speakers participated in TEDphone, 35 of whom also using the the laryngograph equipment. Each session consists of 31 prompts, which the speaker first said in English, and then said in their mother tongue. Since the final formats of the TEDphone corpus have not been decided, a preliminary version will contain the microphone and laryngograph data processed in the same manner as for TEDlaryngo, ie. one file per session, and the telephone channel will remain as individual files.

The filenames for TEDspeeches and TEDlaryngo includes the speaker's initials, the session and paper number, the type of signal (speech or laryngograph), the speech style (presentation), language (English), as well as an indication that the file is compressed. For TEDphone a different naming format was used, where the session and paper numbers were replaced with information about the country code and the prompt text, and additional options are available for the speech style and language.

QUESTIONNAIRES

The TED questionnaire was designed by the EUROCOCOSDA partners in collaboration with Joachim Llisterri (U. Barcelona) and Valerie Hazan (UCL). The questionnaire is primarily concerned with the speakers primary and secondary languages, as well as their knowledge of English. The questionnaire is given in Figure 1.

Often in large speech corpora the speaker identifier is such that the speaker's identity remains unknown. Since in the TED corpus, the paper number, title and authors are known from the written paper, we have not attempted to insure the anonymity of the speaker. Given this decision, the questionnaires will be distributed exactly as they were completed by the speakers. Some of the more relevant information will be summarised in speaker information files. To date, we have received questionnaires from 135 of the

[1] SHORTEN is a losses compression algorithm developed by Tony Robinson at Cambridge University and is freely available by ftp for non-commercial use. The signal compression obtained on the TED data is about 50%. About 40 single channel recordings fit on one CD in compressed format.

[2] Being unsure of the interest in the second channel data with the mix of microphones, the decision was taken to first produce in-house CDs of the primary channel, storing the secondary channel on Exabyte tape for the time being.

Paper No.: Name:
Address: Email:
Tel: Fax:
Sex:
Date of birth: Place of birth (city and country):
At present, what do you consider to be your primary language?
Has this always been the case?
Which language do you consider to be your second language, if any?
Do you consider yourself a bilingual?
If yes, in which languages?
Which language do you use more frequently at home?
Which language do you use more frequently at work?
Do you speak any other languages?
At which age did you start to learn English?
Where and how have you learnt English?
Mother's native language:
Father's native language:
Highest education level:
Language (or languages) of education (number of years studied):
List any countries that you have lived in for periods longer than 1 year:

Figure 1: Questionnaire for *Eurospeech93* TED speakers.

speakers, and are still attempting to obtain those that are missing. To facilitate the gathering of such information, we strongly recommend that the questionnaires be combined with the permissions and gathered before the conference or immediately after.

DOCUMENTATION

Online documentation files have been written for the TED corpus. These consist of files documenting how the recordings were made, as well as how the data are organized on the CDROMs. Additional documentation files will be available on an associated diskette, which will also provide the newest release of the compression software and the batch to convert the compressed signal to a SAM compatible format. The speaker questionnaires and information files will also be distributed on the same diskette. All of the TED documentation will be available by ftp through ELSNET or RELATOR.

TED TEXT CORPUS

A text corpus containing ascii versions of the written proceedings papers for *Eurospeech93* will accompany the spoken corpus. These text materials can be used to provide data for language modeling, and lists of vocabulary items (even references provide lists of proper names) that are likely to show up in presentations that are in the same or in a related domain. For about 50 of the oral presentations, the authors were able to supply a notes or written text which they used during the presentation. These texts should correspond more closely with the recordings, than will the texts of the written papers.

The most commonly used language models are statistically based, and require large text corpora in order to estimate the frequencies of words and word sequences. At the time of this writing we have obtained the texts of 348 presentations, and are still actively trying to obtain others. The average written paper has on the order of 3000 words, which is not enough to provide a good estimate of word frequencies. The total amount of text received thus far is about 1 million words.

While chasing after authors to obtain the written version of their paper and the completed questionnaires, has been quite time-consuming, the success rate after insistence has been high. Only a few papers (about 20) were received in response to the letter sent out be the General Manager of *Eurospeech93*, K. Fellbaum. Follow-up general email messages brought in about 100 more texts, but by far the most effective method means of obtaining the texts are by personalized email messages or personalized letters.

Should additional finances be obtained, and text processing tools located, the text materials will be normalized. In the interim, preliminary versions will be accessible by ftp.

SITUATION AND FUTURE ACTIVITIES

At the time of this writing internal-use copies of the 5 CDROMs for TEDspeeches (188 speeches) and the TEDlaryngo (11 speeches) CDROM have been made and are being verified by the consortium members. The TEDphone data is just starting to be processed.

Questionnaires have been completed by 112 of the speakers. For text materials, we have obtained 360 of the possible 537 texts published in the proceedings. We are still actively the searching the missing papers.

As within the EUROCOCOSDA project there are no funds available to press the CDROMs for wider dissemination (or to actually handle the dissemination activities), the partners are exploring a variety of possible dissmenination options. Preliminary requests for over 50 copies of the corpus have been received.

Should additional permissions be obtained for the remaining usuable speeches, we will attempt produce another CDROM with this data. Should there be sufficient interest in the secondary channel containing the mix of microphones, we will try to find funding to produce this data too.

While transcription of this corpus is beyond the scope of this project, we will provide recommendations for transcribing the data, and we hope that eventually orthographic transcriptions will be available. One suggestion was to ask each speaker to orthographically transcribe their own speech. In order to entice the speakers to participate, they would receive a free copy of the CD containing their speech.

Possible extensions to TED are the presentations recorded at the *ESCA Workshop Automatic Speaker Recognition, Identificaton and Verification* (Martigny, April 1997), as well as the talks to be given at *Eurospeech95* in Madrid. One proposal is to hold a competition for real-time recognition of speeches at a future *Eurospeech*, maybe in 1995 or 1997.