SpeechDat(E) - Eastern European Telephone Speech Databases.

Petr Pollák [(1)], Jan Černocký, [(2)] Jérome Boudy [(3)], Khalid Choukri [(4)], Henk van den Heuvel [(5)], Klara Vicsi [(6)], Attila Virag [(6)], Rainer Siemund [(7)], Wojciech Majewski [(8)], Jerzy Sadowski [(8)], Piotr Staroniewicz [(8)], Herbert Tropf [(9)], Julia Kochanina [(10)], Alexander Ostroukhov [(10)], Milan Rusko [(11)], Marian Trnka [(11)]

[(1)]Czech Technical University in Prague, ČVUT FEL K331, Technická 2, 16627 Praha 6, Czech Republic
pollak@feld.cvut.cz
[(2)]Brno University of Technology, Czech Republic,
[(3)]Lernout & Hauspie, France,
[(4)]ELRA/ELDA, France,
[(5)]SPEX Nijmegen, Netherlands,
[(6)]Technical University of Budapest, Hungary,
[(7)]Philips Speech Processing, Germany,
[(8)]Wroclaw University of Technology, Poland,
[(9)]Siemens AG, Germany,
[(10)]Auditech Ltd., Russia,
[(11)]Slovak Academy of Sciences, Slovakia,

Abstract

This paper describes the creation of five new telephony speech databases for Central and Eastern European languages within the SpeechDat(E) project. The 5 languages concerned are Czech, Polish, Slovak, Hungarian, and Russian. The databases follow SpeechDat-II specifications with some language specific adaptation. The present paper describes the differences between SpeechDat(E) and earlier SpeechDat projects with regard to database items such as generation of phonetically rich sentences, speaker recruitment, etc. The collections of the DBs are in the finishing phase. The DBs will be validated by SPEX and will be distributed by ELRA.

1. Introduction

The great progress in the field of telecommunications over the last decades has brought a need of speech technology in telecommunication services. A typical application is voice input for many different automated services. The reliability of speech recognizers in these applications must be guaranteed by training on realistic data collected directly from the telephone network.

SpeechDat(E) is a project in a series of European projects aiming at the creation of large telephone speech databases (DBs) (van den Heuvel et al., 1998). This project extends successfully finished projects SpeechDat(M) and SpeechDat-II (SpeechDat, http://www.speechdat.org), (Höge et al., 1997), (Draxler et al., 1998). 28 telephone speech DBs were collected under SpeechDat-II; 20 of these DBs were collected from the fixed network (FDB), 5 from mobile network (MDB), and 3 for speaker verification purposes (SDB). These DBs cover all 11 official languages of the European Union, some dialectal variants, and in addition two Slavic languages, namely 1000 speakers Russian FDB and 1000 speakers Slovenian FDB.

The 5 new DBs of five Eastern European languages are currently at the finishing phase within SpeechDat(E). The consortium of 10 partners work on the collection of this data. The project co-ordinator is MATRA NORTEL Communication.

Following DBs are being collected:

- *1000 speakers Czech DB* at the Brno University of Technology and the Czech Technical University in Prague,
- *1000 speakers Slovak DB* at the Institute of Control Theory and Robotics at the Slovak Academy of Sciences in Bratislava,
- *1000 speakers Polish DB* at the Wroclaw University of Technology in co-operation with Siemens,
- *2500 speakers Russian DB* at Auditech Ltd. in St.-Petersburg (this DB is an extension of existing 1000 speaker FDB collected within SpeechDat-II),
- *1000 speakers Hungarian DB* by Philips in co-operation with the Technical University in Budapest.

The most important language specific problems are discussed in the following sections . The details can be found in project deliverables (Galounov and Kochanina, 1999), (Černocký et al., 1999), (Rusko, 1999), (Sadowski and Staroniewicz, 1999), (Vicsi and Virag, 1999).

2. Database Item Design

The corpora were derived from SpeechDat(II) item list [1]. Table 1 illustrates the list of items to be sampled in SpeechDat(E). Figures slightly vary between 1000 and 2500 speaker DBs.

Recommended minimal amount of **application words** is 25 but this amount may be higher because there is not

1000 speakers	*2500 speakers*	*Type*	*Items*
2	**2**	**isolated digits**	- single isolated digit, - sequence of 10 isolated digits,
4	**4**	**digit/number string**	- prompt sheet number, - telephone number, - credit card number, - PIN-code,
1	**1**	**natural number**	- 4 non-zero digit number up to 10 000 000
2	**2**	**money amount**	- local currency, - international currency (US dollar, Euro),
2	**2**	**yes/no question**	- predominantly ’yes’ **(spontaneous)**, - predominantly ’no’ **(spontaneous)**,
3	**3**	**date**	- birth-date **(spontaneous)**, - prompted phrase, - relative and general date expression,
2	**2**	**time**	- time of day **(spontaneous)**, - prompted time phrase,
6	**6**	**application keyword (keyphrase)**	- commands for different teleservices,
1	**1**	**word spotting phrase**	- using embedded application word,
6	**6**	**directory assistance name**	- city of birth/growing up **(spontaneous)**, - city, - company/agency, - surname, - forename and surname, - own forename **(spontaneous)**,
3	**3**	**spelling word**	- artificial sequence, - city name, - own forename **(spontaneous)**,
4	**4**	**phonetically rich word**	
12	**9**	**phonetically rich sentence**	
48	**45**	**Total number of items**	

Table 1: Structure of items for SpeechDat(E) databases.

always a simple translation of one English application word for some languages, e.g. “re-dial” - “opakovat volbu” for Czech, “wybierz ponownie” for Polish. The larger set of application words (33 words) was designed for Russian DB.

There are also more possible ways of the translation for some application words, often depending on the context. We can see one example from Slovak DB. The word “operator” can be translated as “operátor” but the older expression “spojovateľka” is used more often, so it was included in the Slovak DB too. The similar case is for the word “cancel”, two possible translations “zrušiť” and “storno” were used in the DB items.

Generally, some difficulties in item design appeared with respect to the **high inflective nature** of all collected languages. The situation for Russian, Czech, Slovak, and Polish were very similar due to the common Slavic origin. Compared to languages of Romance or Germanic origin, the differences are not very significant for verb conjugation. However, nouns, adjectives, and numerals have got 6 or 7 singular and 6 or 7 plural cases.

It consequently caused some problems to cover some words in single items like natural numbers, times, money amounts, etc. One example from Czech corpus: the numeral “hundred - sto” can have 4 different forms “jedno sto (100), dvě stě (200), tři sta (300), pět set (500)”. All these forms should be well covered in the DB but it starts being difficult for 1000 speakers DB. A similar case can be found for “thousand - tisíc, hundredth part - setina, etc.”, and again a similar situation exists for Polish, where the form of names of currencies (in money amounts items) change depending on number, e.g. “jeden cent - one cent”, “trzy centy - three cents” and “pięċ centów - five cents”.

There are also two **different forms for surname** with respect to gender in Slavic languages. Generally, they may sound quite differently so they must be both included in the DBs.

For example in Polish some surnames have different forms depending on gender (e.g. with endings ”-ski”, ”-cki” for male and ”-ska”, ”-cka” for female). Different extensions of surname can be found also in Russian, e.g. “Sokolov, Kuzmin” for male surnames, “Sokolova, Kuzmina” for female ones. Moreover, in this case the stress may fall on the same syllable or it may be shifted, e.g. “Sokolov, Sokolova”, the syllable “lov” is stressed, while in “Kuzmin, Kuzmina” the last syllable is stressed in both cases.

In Hungarian, the surname (family name) stands at the first place and the forename (Christian name) follows it. Instead of using “Mrs.” several variations are accepted for female names. Mostly for married women, husband’s family name stands at first place and the “né” suffix is added to the husband Christian name at the second place, e.g. “Kovács Sándorné” which means in English “Kovacs Sandor’s”.

An important task of DB design was the creation of corpora of **phonetically rich material**. The corpora of phonetically rich material were obtained in different way for each language:

- *Czech* - by processing of newspaper texts downloaded from different Internet WEB-sites.

  The first original corpus of the texts contained more than two million sentences. Firstly, several filters were applied to exclude unusable sentences, i.e. too long or too short, containing digits, abbreviations, parenthesis, etc. Then the sentences were transcribed into sequences of phones. This transcription reflected the most probable pronunciation. Secondly, all sentences were scored by number expressing its contents of rare units and then they were sorted by descending score. A sub-corpus of sentences with the highest score was selected. Word frequencies were taken into account in this selection to suppress too many repetitions of words. Finally, resulting 8.000 sentences were hand-checked for hard-to-pronounce words, offensive contents, grammatical errors etc. All the previous automatic processing was repeated on this small clean corpus and the final set of 5.300 sentences was obtained.

- *Slovak* - For the Slovak database 2949 sentences were designed. They were taken from texts of different styles, such as books, encyklopedias and newspapers. A part of the sentence corpus is taken from the train information system and a part also from legal literature. Sentences from fiction were introduced by including hundreds of pages of text of the ”Literárny týždenník” (Literary weekly magazine). The sentences are up to 10 words long.

- *Hungarian* - The basic material for the creation of Hungarian phonetically rich sentences was newspaper text. Its size was about 1.6 MB and it contains about 14000 sentences. Firstly, the text was cleaned from extra characters, meaningless words, page numbers, etc. Then it was converted into a string of phonemes with a special algorithm which was developed at the Technical University in Budapest. The statistical analysis followed and 2400 sentences were then chosen. Each sentence is repeated 5 times in the DB now.

- *Polish* - The Polish corpus of 1536 sentences was collected from various sources with a special care for good coverage of rare phonemes. The corpus was divided into twelve sets each with special respect for containing particular group of the most rare phonemes. Finally the computer program organized the phonetically rich sentences in sets of 12 (each set for a separate answer sheet), where each set contains at least 2 examples of each Polish phoneme. As the result 1280 sets of sentences were obtained. A similar procedure was applied for the phonetically rich words.

- *Russian* - For Russian DBs, the phonetically rich sentences were selected from various novels by Russian authors and a number of newspaper and magazine articles on various topics. The selection was carried out by experts; the selected material was further phonetically balanced with the help of specific software and the necessary corrections were made.

3. Speaker Coverage

Speaker coverage is balanced with respect to gender, age, and dialect. For gender it is on 50% – 50% basis with allowed tolerance 5%. The requirements for the age coverage are summarized in table 2.

Age group	*Speakers in DB*
0 – 15	min. 1 %
16 – 30	min. 20 %
31 - 45	min. 20 %
46 - 60	min. 15 %
> 60	optional

Table 2: Age groups coverage.

3.1. Dialect balance

While the coverage with respect to sex and age was quite clear, the dialect balance of DBs causes some problems. It should be proportional to population in defined dialectal regions. But the definition of these regions was difficult for some languages. These problems were caused generally by great movement of people during last 50 years.

- *Czech* : The definition of Czech dialectical regions was relatively easy. Together with Assoc. Prof. Zdena Hladká from Masaryk University Brno, 5 regions were defined. The population in defined regions seems to be quite stable, except the movement of younger people from the country to cities. Nevertheless, since the distances in Czech Republic are not very long, people usually do not lose contacts with their place of origin. From this point of view, we have chosen the place of basic school finishing as the criterion of dialect coverage.

- *Slovak* : Slovakia still has many (about nine) dialect regions because the Slovak literary language is very young. All of them are covered in the DB, but for the practical purposes it is better to have a smaller number of integrated dialect regions. After a discussion with phoneticians we decided to geographically divide

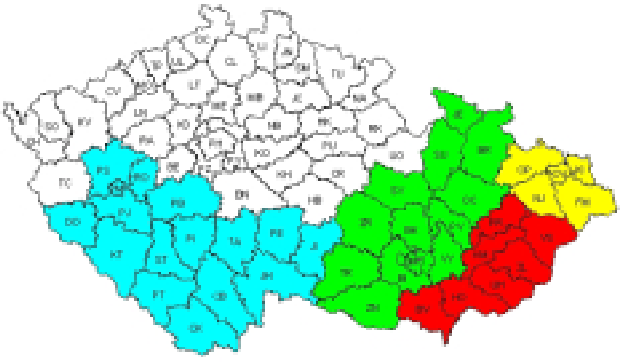

Figure 1: Dialect regions for the recording of Czech DB.

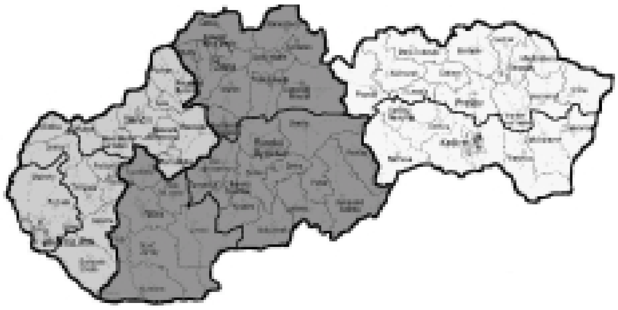

Figure 2: Dialect regions for the recording of Slovak DB.

Slovakia into three dialect regions. Their borders follow the borders of the recent administrative regions – counties.

- *Hungarian* : During the last decades Hungarian has become quite uniform in Hungary, although there are some slight but characteristic accents within different parts of the country. They are merged into four regions: NORTHERN involving "Palóc, Jász, and Északkeleti", WESTERN involving "Nyugati and Dunántúli", TISZA involving "Tiszai", and SOUTHERN involving "Déli". The names of regions are the English translation of the dominant dialect name.

- *Polish* : Since the population of Poland is mixed up it was impossible to divide Poland into definite number of dialectal regions. Thus it was decided to divide the country into eight geographical regions that only roughly correspond to dialectal regions. The dialectal coverage was based on the place of speakers primary school.

- *Russian* : For the purposes of the DB, 4 dialectical regions have been defined. The dialectal variation in Russia is not very great, considering the size of the country. The dialectal features are less evident in the speech of citizens of big cities and among younger people. The accent is determined according to three parameters: 1) where the speaker spent his childhood; 2) how long the speaker has been living in the dialectal region; 3) the presence of certain dialectal features as defined by an expert.

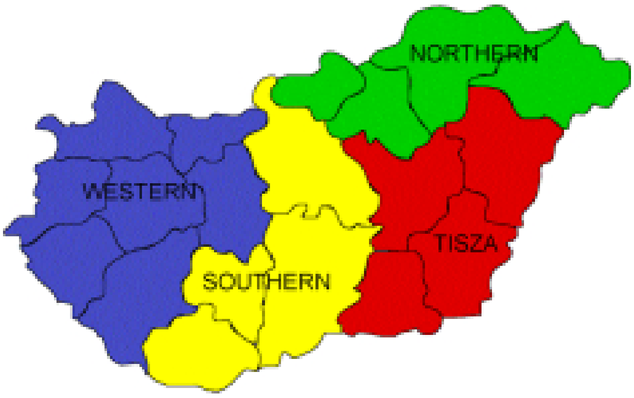

Figure 3: Dialect regions for the recording of Hungarian DB.

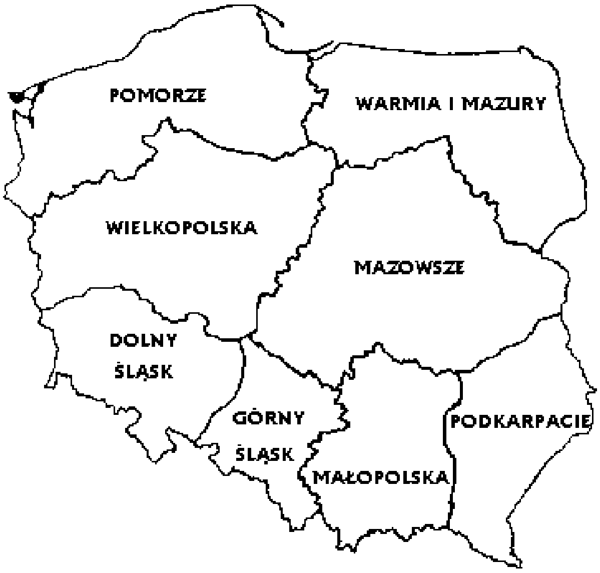

Figure 4: Dialect regions for the recording of Polish DB.

3.2. Recruitment strategy

- *Czech* : Due to bad experiences of Czech people with unserious publicity campaigns, and to lack of special agencies (and of funding to pay them), the snowball recruitment strategy has been adopted. Students of Brno and Prague universities (from different places of Czech Rebuplic), relatives and friends were asked to recruit small number of speakers (usually 20), which might themselves become also recruiters. The small present (a camera film) was offered for a completed call and also for a certain number of recruited speakers. The calls from 1000 speakers were collected within 4-5 months.

- *Slovak* : The speakers in the DB were recruited from the employees of the Slovak Academy of Sciences, from the teachers and the students of several Universities in Slovakia, as well as their relatives and friends. Members of some organizations, such as The Slovak Acoustic Society, The Slovak Cybernetic Society and some cultural institutions were asked for a help too. Every speaker, whose call has been successfully recorded was given a gift - a camera-film. As there are too many non-serious lotteries and telephone games in Slovakia, the most effective way of recruiting speakers was their personal contact with informed persons - recruiters - who knew, how to explain them the need of recording the database. These recruiters were paid according to the number of persons they have recruited to call. In this case the speakers were not given any gift. The recorded data were regularly checked to follow the desired dialect, age and gender distributions.

- *Hungarian* : A subcontracting of two big companies which recruited among their employees were used. The Hungarian Railway Company (MÁV Rt.) organized all speakers from the regions Tiszai, Western, Northern, and some people from the Southern. The MATÁV Rt. (telecommunications company) organized the speakers from the capital. Only one responsible contact person was in each company. Age, sex, and environment were defined on each prompt-sheet-cover and the organizer had to find an employee according to these parameters. Unfortunately, there were some missing, unusable or incorrect calls. To fill these gaps, the missing prompt sheets were re-printed and they were delivered to secondary schools via Internet or delivered by students of the University.
- *Polish* : The recruitment was done in all eight regions and in each region a group of organizers was selected. Each organizer was responsible for providing 20 speakers from his/her region (group should be sex balanced and should be collected among people of given age ranges). The speakers were given the instruction sheet and also briefed in person by the organizer. Additionally the organizers received information how to contact responsible person and instructions how to train the speakers (with tape which includes the sample material). The relevant questions were a part of the recording session. The answers to the questions in the prompt sheet included the information about the speaker's primary school place, city of call, and also the information about the recording environment and type of phone set. The information about the gender of the speaker is determined at the transcription stage. The age of the speaker is to be known from the spontaneous date item, where the speakers are requested to say their date of birth.
- *Russian* : The snowball method was mostly used for the collection of Russian DB. Personal contacts in different cities were asked to recruit more people. Professors from universities in different parts of the country were asked to recruit their students and their relatives. A number of speakers were found with the help of a social research agency. Recruiters received certain payment; some of the individual callers also received small payment.

4. Annotations

The recorded data are annotated with orthographic transcription rules based on those used in SpeechDat DBs. The label file is in SAM format. Firstly, a suitable annotation tools were chosen. The possibilities of adaptation of existing annotation tools developed within SpeechDat(II) were checked. All partners wrote finally their own tools. It seemed to be more efficient than too complicated language oriented adaptation of an existing tool.

Secondly, the annotations of some untrivial phenomena had to be solved. An important deviation which should be mentioned as an example appeared for Czech and Slovak languages, i.e.:

1. "Ch" is spelled as a unique letter. It was necessary to take this into account analyzing the number of occurrences of different letters in spelled items.
2. There is not unambiguous way to spell letters. Two ways of spelling coexist: the official one, spelling for example "B" as "bé", and the unofficial (but widely used also by educated people), where simply the phonetic form is read ("b", followed by a brief schwa). Moreover, for some letters up to 4 different forms of spelling the same letter were used by speakers. All these spelling forms had to be taken into account in the annotation of calls.

5. Validation

The SpeechDat(E) project is featured by a thorough validation protocol. The specifications which the databases should meet are evaluated by an independent validation center, SPEX. Validation proceeds in three steps:

1. Prevalidation of a small database of 10 speakers. The objective of this stage is to detect serious errors before the actual recordings start.
2. Validation of complete databases. The database is checked against the SpeechDat(E) specifications and a validation report is generated.
3. Revalidation of complete databases. In case the validation report shows that corrections of a database are necessary or desirable, then (part of) the database can be offered for a second validation, and a new report is written.

The final validation report is put onto the final CDs as part of the database.

As in the predecessor project SpeechDat(II), validation comprises all relevant aspects of a database: quality of the transcriptions and of the documentation; signal quality; completeness of the lexicon; speaker and recording environment distributions; correctness of the directory structure, file names and of the format of the label files.

The validation criteria were adopted from SpeechDat(II) (van den Heuvel, 1996). A number of modifications were applied, which are described in (van den Heuvel, 1999). The most important deviations are the following:

– An additional maximum of 5% of the phonetically rich sentences may contain corrupted speech only;
– Natural numbers may exceed 1 Million, provided they do not contain more than 4 significant digits;
– Phonetically rich sentences: each unique sentence should not appear more than 10 times;
– Phonetically rich words: each unique word should not appear more than 5 times;

- Line lengths in label files may exceed 80 characters (extension of standard SAM format);
- The distribution of the dialect regions among the calls should be proportional to that of the population with a deviation of 5% at the maximum, and a minimum representation of 5% of the calls for each dialect region.

At present, the prevalidation of the five corpora has been carried out. The prevalidations took place in the period July - September 1999. There were a number of observations which applied to several databases. The most relevant of these are listed below.

- Word-level punctuation (e.g. hyphens and apostrophes) should be used in the orthographic transcriptions;
- Sentence-level punctuation (e.g. full stop, colon, comma) should be omitted;
- The symbols for unintelligible parts and recording truncations should not be omitted in the orthographic transcriptions;
- All symbols for non-speech acoustic events should be used;
- Transcriptions should be in an ISO 8859-n character set. In case another coding page is used, the software to convert the transcriptions (and the lexicon) to the corresponding ISO 8859 character set should be provided;
- The word spotting phrase should contain at least three words;
- A frequency count of the phones in the full database should be part of the documentation. These counts should pertain to all read items;
- The definition of a list of test sessions should be included in the documentation.

6. Distribution

Most of the SpeechDat(E) databases will be distributed through the European Language Resources Association (ELRA). ELRA was established as a non-profit association in Luxembourg in February 1995, to provide a European-wide, open platform for the selection and distribution of speech, text and terminology resources to be embedded in language enabled systems, and to promote the use of Language Resources within the Human Language Technologies sector (HLT). ELRA has been granted the rights to distribute most of the speech data bases collected within the European funded projects, in particular SpeechDat(M) and SpeechDat-II. Linguistic Resources are universally acknowledged to be critical for the development of robust, broad-coverage, and cost-effective applications for all sectors of HLT, in particular those addressing multi-lingual issues. ELRA has already finalized a distribution agreement with the owners/producers of 4 Eastern European databases. Interested parties will have to enter into only one agreement with ELRA.

7. Conclusions

The 5 new very large telephone DB for Czech, Polish, Slovak, Hungarian, and Russian are being collected and prepared for the validation procedure. These DBs should promote the creation of user friendly voice-driven services of telecommunications operators. It starts to be very important especially from the point of view of the liberalization of telecommunications markets in some countries (e.g. Czech Republic 2001, Slovakia 2002). The databases will be mostly distributed by the European Language Resources Association (ELRA/ELDA).

Acknowledgements

SpeechDat(E) project is supported by INCO- Copernicus Project No. 977017. Most of the resulting DBs will be distributed by the ELRA which support the validation of DBs.

8. References

Draxler, C., H. van den Heuvel, and H. Tropf, 1998. SpeechDat expieriences in creating large multilingual speech databases for teleservices. In *First International Conference on Language Resources and Evaluation, LREC'98*. Granada (Spain).

Galounov, V. and J. Kochanina, 1999. Definition of corpus, scripts, standards, and specification of environment/speaker coverage for Russian. Technical report, SpeechDat(E). Deliverable ED1.12.1, workpackage WP1.

Höge, H., H. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach, and K. Choukri, 1997. European speech databases for telephone applications. In *Proc. of ICASSP*.

Rusko, M., 1999. Definition of corpus, scripts, and standards for fixed networks. Technical report, SpeechDat(E). Deliverable ED1.12.3, workpackage WP1.

Sadowski, J. and P. Staroniewicz, 1999. Definition of corpus, scripts, standards, and environmental and speaker specific coverage applied for speech databases for Polish. Technical report, SpeechDat(E). Deliverable ED1.12.4, workpackage WP1.

SpeechDat, http://www.speechdat.org. WEB-page of all SpeechDat projects.

van den Heuvel, H., 1996. Validation criteria. SpeechDat Technical Report SD1.3.3, SPEX.

van den Heuvel, H., 1999. Validation criteria. SpeechDat(E) deliverable ED1.4.2, SPEX.

van den Heuvel, H., V. Galounov, and H. Tropf, 1998. The SpeechDat(E) project: Creating speech databases for Eastern European languages. In *Proc. of Workshop on Speech Database Developement or Central and Eastern European Languges*. Granada (Spain).

Černocký, J., P. Pollák, and V. Hanžl, 1999. Definition of corpus, scripts, standards, and environmental and speaker specific coverage applied for speech databases. Technical report, SpeechDat(E). Deliverable ED1.12.2, workpackage WP1.

Vicsi, K. and A. Virag, 1999. Definition of corpus, scripts, standards, and environmental and speaker specific coverage applied for speech databases for Hungarian. Technical report, SpeechDat(E). Deliverable ED1.12.5, workpackage WP1.