The Design for the Wall Street Journal-based CSR Corpus*

Douglas B. Paul and *Janet M. Baker*

Lincoln Laboratory, MIT
Lexington, Ma. 02173

Dragon Systems, Inc.
320 Nevada St.
Newton, Ma. 02160

ABSTRACT

The DARPA Spoken Language System (SLS) community has long taken a leadership position in designing, implementing, and globally distributing significant speech corpora widely used for advancing speech recognition research. The Wall Street Journal (WSJ) CSR Corpus described here is the newest addition to this valuable set of resources. In contrast to previous corpora, the WSJ corpus will provide DARPA its first general-purpose English, large vocabulary, natural language, high perplexity, corpus containing significant quantities of both speech data (400 hrs.) and text data (47M words), thereby providing a means to integrate speech recognition and natural language processing in application domains with high potential practical value. This paper presents the motivating goals, acoustic data design, text processing steps, lexicons, and testing paradigms incorporated into the multi-faceted WSJ CSR Corpus.

INTRODUCTION

As spoken language technology progresses and goals expand, progressively larger, and more challenging corpora need to be created to support advanced research. The SLS DARPA 1994 goals are ambitious, focusing on cooperative speakers, generating goal-directed, spontaneous continuous speech, in speaker-adaptive and speaker-independent modes, for expandable vocabularies (5000 or more words active), moderate perplexity (100-200), with integrated speech and natural language processing, for speakers in a moderate noise environment, using multiple types of microphones, engaged in command/database and dictation applications. In contrast to typical command/database applications, dictation (i.e. interactive speech-driven word processing) tasks focus on cooperative speakers (e.g. speaker dependent/adaptive sustained usage) who generate continuous speech (usually in a somewhat careful fashion to facilitate accurate transcription) verbalizing their words and sentence punctuation. The existing Resource Management[15] and subsequent Air Travel Information System[16] corpora target specific database inquiry tasks, characterized by medium vocabularies (<1500 words) with language model perplexities ranging from 9 to 60. The WSJ corpus described here is designed to advance CSR technology and support the 1994 SLS research goals. A similar read speech corpus in the French language has been successfully completed using text from the newspaper Le Monde[5].

Commencing with serious contractor concerns regarding suitable CSR corpora[12] starting in the mid 1980's, the DARPA SLS Coordinating Committee started considering new corpora requirements in early 1990, with the subsequent formation of the CSR Corpus Committee, culminating in the WSJ Corpus design. The CSR Corpus Committee members include J.M. Baker (Dragon, chair), F. Kubala (BBN), D. Pallett (NIST), D. Paul (LL), M. Phillips (MIT), M. Picheny (IBM), R. Rajasekran (TI), B. Weide (CMU), M. Weintraub (SRI), and J. Wilpon (ATT). A survey taken of the DARPA contractors for CSR research interests disclosed highly diverse, often opposing views of research interest. All contractors, however, cited a common interest in pursuing research on "Domain-independent Acoustic Models", "Domain-independent Language Models", and "Speaker-adaptation".

The outcome of lively meetings and discussions resulted in the definition and preliminary authorization of a major (>400 hrs.) corpus with materials based primarily on WSJ material (backed by WSJ text from 1987-89 provided by the ACL/DCI[9] to enable statistical language modeling) and supplemented by other material (spontaneous dictation, Hansard, etc., shown in Table 1). This corpus will provide a uniquely rich resource, in a carefully crafted structure designed to elicit a highly productive flow of diagnostic research information with an array of comparative test paradigms.

Although this WSJ corpus is large relative to many other available corpora, it should be cautioned that insofar as most research experiments continue to show marked improvement with the increased availability of training data, it is likely that this corpus also will fail to allow us to find or achieve asymptotic performance. Most systems continue to be undertrained or constrained to work in suboptimal lower dimensional spaces, due to their data-starvation. Indeed, this result is not really surprising in light of the much larger amounts of speech data to which young children must be exposed before gaining recognition proficiency of even modest size vocabularies.

The structure, features, and dimensions of this corpus constitute the outcome of a heavily debated consensus process, which satisfies the basic (though certainly not all) different requirements of the different research foci of all parties involved. There are significant portions of this corpus which

*This work was sponsored by the Defense Advanced Research Projects Agency. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government.

will be more heavily used by one or more research groups, and not at all by others. Nonetheless, the common basis and careful structuring of these materials should allow for highly informative intra- and inter-group comparisons. The members of this committee are to be commended and should take pride in their success in jointly exercising a rare "statesmanlike" cooperation to support the legitimate diversity of expert research interests in this field (often overcoming strong pressures of both personal and political convictions to support only their own narrower research interests).

THE WSJ-CORPUS STRUCTURE AND CAPABILITIES

Specifically, the WSJ corpus is scalable and built to accommodate variable size large vocabularies (5K, 20K, and larger), variable perplexities (80, 120, 160, 240, and larger), speaker dependent (SD) and independent (SI) training with variable amounts of data (ranging from 100 to 9600 sentences/speaker), including equal portions of verbalized and non-verbalized punctuation (to reflect both dictation-mode and non-dictation-mode applications), separate speaker adaptation materials (40 phonetically rich sentences/speaker), simultaneous standard close talking and multiple secondary microphones, variable moderate noise environments, equal numbers of male and female speakers chosen for diversity of voice quality and dialect. In order to collect large quantities of speech data very cost-effectively, it was decided to collect the majority of the recorded speech in a "read" speech mode, whereby speakers are prompted by newspaper text paragraphs. The presentation of coherent paragraph blocks of text provides semantically meaningful material, thereby facilitating the production of realistic speech prosodics. Small amounts of unprompted "spontaneous" speech are provided for comparison (utilizing some naive speakers as well as some who are experienced at dictation for human transcription).

Testing paradigms were carefully constructed to accommodate efficient comparisons of SI and SD performance and variable size vocabulary "open" and "closed" tests to permit evaluation both with and without "out-of-vocabulary" lexical items. The value of variable amounts of training set materials can be directly assessed both within and across speakers. Well-trained speaker-dependent performance provides an upper bound against which the success of different speaker-independent modeling and speaker-adaptive methodologies may be rigorously compared.

Adaptive acoustic and language modeling is easily supported through the following simple though rigorous automatic paradigm: 1) Recognition of a sentence is performed and assessed as usual against existing system acoustic and language models. 2) The system commences to adapt using (supervised) or not using (unsupervised) the correct "clear text" to modify its internal acoustic and language models automatically before proceeding to recognition of the next utterance.

Recognition performance with this kind of automatic adaptation is assessable with standard scoring routines. This mode provides an easy means to maximize performance for speakers by tracking and accommodating to speaker and environmental changes in a dynamic fashion, also simulating (in a reproducible fashion) an interactive system mode where speakers correct system recognition errors, and using systems which can utilize this feedback to improve performance, in a continuous automatic fashion. The results of automatic adaptation can be assessed in an on-going "dynamic" fashion, or stopped after varying amounts of adaptation, for subsequent "static" testing on materials to which the system is not subsequently adapted[1,2,3].

The availability of large amounts of machine-readable text from nearly three years of the Wall Street Journal enables meaningful statistical benchmark language models (including bigrams and trigrams) to be generated, and the results from these to be easily contrasted. By varying the types of language models chosen, the effect on recognition performance of variable perplexities for the same textual materials can be assessed. The availability of this text provides a valuable resource enabling novel language models and language models adapted from other tasks to be developed and evaluated as well.

THE WSJ-PILOT DATABASE

It was judged to be too ambitious to immediately record a 400 hour recognition database. Therefore, a smaller pilot database built around the WSJ task was designed. A joint BBN/Lincoln proposal for the pilot was adopted by the CSR committee. In an attempt to "share the shortage" this proposal provided equal amounts of training data for each of three popular training paradigms. This proposal was also rich enough that it provided for "multi-mode" use of the data to allow many more than just three paradigms to be explored. The original plan was for about a 45 hour database, but the three recording sites, (MIT, SRI, and TI), each recorded about a half share for a total of 80 hours. The resultant database is shown in Table 4 and described below. (About 1.5K additional SI training sentences are not shown in the table.)

THE WSJ TEXT PREPROCESSING

It is important to be able to train a language model that is well matched to the (text) source to be used as a control condition to isolate the performance of the acoustic modeling from the language modeling[12]. (It is always possible to train a mismatched language model, but its effects cannot be adequately assessed without a control matched language model.) Ideally, one would have access to many (tens to hundreds of millions of words) of accurately transcribed spoken speech. Such was not available to us. Therefore, this condition was simulated by preprocessing the WSJ text in a manner that removed the ambiguity in the word sequence that a reader might choose. (This preprocessing is similar to that which might be used in a text-to-speech system[4].) This ensures that the unread (and unchecked) text used to train the language model is representative of the spoken test

material.

The original WSJ text data were supplied by Dow Jones, Inc. to the ACL/DCI[9] which organized the data and distributed it to the research community in CD-ROM format. The WSJ text data were supplied as 313 1MB files from the years 1987, 1988 and 1989. The data consisted of articles that were paragraph and sentence marked by the ACL/DCI. (Since automatic marking methods were used, some of the paragraphs and sentence marks are erroneous.) The article headers contained a WSJ-supplied document-control number.

The preprocessing began with integrity checks: one file from 1987 and 38 from 1988 were discarded due to duplication of articles in the same file (1987) or duplication of data found in other files (1988). 274 files were retained, which yielded 47M with-verbalized-punctuation words from 1.8M sentences. (The yield is on the order of 10% fewer words in the non-verbalized-punctuation version.) Each file contain a scatter of dates, usually within a few days, but sometimes up to six months apart. Each file was characterized by its most frequent date (used later to temporally order the files).

Since the CSR Committee had decided to support both with and without verbalized punctuation modes, it was necessary to produce four versions of each text: with/without verbalized punctuation × prompt/truth texts. (A prompt text is the version read by the speaker and the truth text is the version used by the training, recognition, and scoring algorithms.) The preprocessing consisted of a general preprocessor (GP) followed by four customizing preprocessors to convert the GP output in the four specific outputs. The traditional computer definition of a word is used—any white-space separated object is a word. Thus, a word followed by a comma becomes a word unless that comma is separated from the word. (Resolution of the role of a period or an apostrophe/single quote can be a very difficult problem requiring full understanding of the text.)

The general preprocessor started by labeling all paragraphs and sentences using an SGML-like scheme based upon the file name, document-control number, paragraph number within the article, and sentence number within the paragraph. This marking scheme, which was carried transparently though all of the processing, made it very easy to locate any of the text at any stage of the processing. A few bug fixes were applied for such things as common typos or misspellings. Next the numbers are converted into orthographics. "Magic numbers" (numbers such as 386 and 747 which are not pronounced normally because they have a special meaning) are pronounced from an exceptions table. The remaining numbers are pronounced by rule—the algorithms cover money, time, dates, "serial numbers" (mixed digits and letters), fractions, feet-inches, real numbers, and integers. Next sequences of letters are separated: U.S.→U. S., Roman numerals are written out as cardinals or ordinals depending on the left context, acronyms are spelled out or left as words according to the common pronunciation, and abbreviations (except for Mr., Mrs., Ms., and Messrs.) are expanded to the full word. Finally, single letters are followed by a "." to distinguish them from the words "a" and "I". This output is the input to the four specific preprocessors.

The punctuation processor is used in several modes. In its normal mode, it is used to produce the with-verbalized-punctuation texts. It resolves apostrophes from single quotes (an apostrophe is part of the word, a single quote is not), resolves whether a period indicates an abbreviation or is a punctuation mark, and separates punctuation into individual marks separate from the words. This punctuation is written out in a word-like form (eg. ,COMMA) to ensure that the speaker will pronounce it. This output is the with-punctuation prompting text. Until this point, the text retains the original case as suppled on the CD-ROM. If one wishes to perform case-sensitive recognition (ie. the language model predicts the case of the word), this same text can be used as the with-punctuation truth text or if one wishes to perform case-insensitive recognition, the text may be mapped to upper-case. (A post-processor is supplied with the database to perform the case mapping without altering the sentence markings.) Initial use of the database will center on case-insensitive recognition.

The without-punctuation prompting text is very similar to the GP output. Only a few things, such as mapping "%" to "percent", need to be performed. This text contains the mixed case and normal punctuation to help the subject speak the sentence. (The subject is instructed not to pronounce any of the punctuation in this mode.) The punctuation processor is used in a special mode to produce the without-punctuation truth-text. It performs all of the same processing as described above to locate the punctuation marks, but now, rather than spelling them out, eliminates them from the output. (Since the punctuation marks do not appear explicitly in the acoustics, they must be eliminated from the truth texts. Predicting punctuation from the acoustics has been shown to be impractical—human transcribers don't punctuate consistently, and, in an attempt to perform punctuation prediction by the language model in a CSR, IBM found a high percentage of their errors to be due to incorrectly predicted punctuation[14]. People dictating to a human transcriber verbalize the punctuation if they feel that correct punctuation is important: e.g. lawyers. They also verbally spell uncommon words and issue formatting commands where appropriate.) This without-punctuation truth text is again mixed case and can be mapped to upper case if the user desires.

WSJ TEXT SELECTION INTO DATABASE PARTS

Next it was necessary to divide the text into sections for the various parts of the database. Since the plan called for the pilot to become a portion of the full database, all text processing and selection were performed according to criteria that were consistent with the full database.

Ninety percent of the text, including all of the Penn Treebank[17] (about 2M words) were reserved for training, 5% for development testing, and the remaining 5% for evaluation testing. The non-treebank text files were temporally

ordered (see above) and 28 were selected for testing—the odd ordinal files for development testing and the even ordinal files for evaluation testing. (The Treebank included the 21 most recent files so it was not possible to simulate the real case—train on the past and test on the "present").

All of the non-test data, with the exception of the sentences recorded for acoustic training, is available for training language models. The acoustic training data is eliminated to allow a standard sanity check: CSR testing on the acoustic training data without also performing a closed test on the language model.

WSJ TEXT SELECTION FOR RECORDING

Next the recording sentences were selected. Separate sentence "pools" were selected from the appropriate text sections for SI train (10K sentences), SD train (20K sentences), 20K-word vocabulary test (4K development test and 4K evaluation test sentences), and 5K-word vocabulary test (2K development test and 2K evaluation test). It was originally hoped that the 5K vocabulary test set could be formed as a subset of the 20K test set, but this was not possible—thus the 4 test sets are completely independent.

The recording texts were filtered for readability. (The WSJ uses a lot of uncommon words and names and uses complex sentence structures that were never intended to be read aloud.) The first step was to form a word-frequency list (WFL) (ie. a frequency-ordered unigram list) from all of the upper-case with-punctuation truth texts. This yielded a list of 173K words. (For comparison, mixed case yields 210K words). Next, a process of automated "quality filtering" was devised to filter out the majority of the erroneous and unreadable paragraphs. This filtering is applied *only* to the recorded texts, not to the general language model training texts. Since many typos, misspellings and processing (both ACL-DCI and preprocessing) errors map into low frequency words, any paragraph which contained an out-of-top-64K-WFL word or was shorter than 3 words was rejected. (The top 64K WFL words cover 99.6% of the frequency-weighted words in the database.) Any paragraph containing less than three sentences or more than eight sentences was rejected to maintain reasonable selection unit sizes. Any paragraph containing a sentence longer than 30 words was rejected as too difficult to read[1]. Because the WSJ contains many instances of certain "boiler-plate" figure captions which would be pathologically over represented in the test data, duplicate sentences were removed from the test sentence pools. Finally human checks verified the high overall quality of the chosen sentences. Note that this does not mean perfect—there were errors in both the source material and the preprocessing.

[1] One of the authors (dbp) has recorded about 2500 WSJ sentences. The most difficult sentences to record were the longest ones. After a little practice, verbalized punctuation sentences were only slightly harder to read than the non-verbalized punctuation ones. This slight additional difficulty can be accounted for by the fact that the verbalized punctuation sentences average about 10% longer than the non-verbalized punctuation ones.

The 20K test pools were produced by randomly selected quality-filtered paragraphs until 8K (4K dev. test and 4K eval. test) sentences were selected. This produced a realized vocabulary of 13K words. Since this data set was produced in a vocabulary insensitive manner, it can be used without bias for open and closed recognition vocabulary testing at any vocabulary size up to 64K words. (However, using it for open vocabulary testing at any vocabulary size less than 20K will yield a large number of out-of-vocabulary errors—the top-20K of the WFL (the 20K open vocabulary) has a frequency weighted coverage of 97.8% of the data.)

Attempts to produce the 5K vocabulary test pools by the same method produced too few sentences to be useful (~1200). Thus it was necessary to use a vocabulary sensitive procedure—paragraphs were allowed to have up to 1 out-of-top-5.6K-WFL words. This produced the highest yield (~4K sentences with a realized vocabulary of 5K words) and reduces, but does not completely eliminate the tail of the word frequency distribution. This test set allows open and closed vocabulary testing at a 5K-word vocabulary, but would be expected to yield somewhat biased test results if used at larger test vocabularies[10,14]. The top-5K of the WFL (the 5K open vocabulary) has a frequency weighted coverage of 91.7% of the data.

Finally, the evaluation test paragraphs were broken into four separate groups. This was done to provide four independent evaluation test sets.

The recording sites selected a randomly chosen subset of the paragraphs from the pool corresponding to the database section being recorded (with replacement between subjects) for each subjects to read. The sentences were recorded one per audio file. All subjects recorded one set of the 40 adaptation sentences.

OTHER WSJ DATABASE COMPONENTS

The above describes the selection and recording of the acoustic portion of the WSJ-pilot database. Additional components—such as a dictionary and language models—are required to perform recognition experiments. Dragon Systems Inc., under a joint license agreement with Random House, has provided a set of pronouncing dictionaries—totaling 33K words—to cover the training and 5K and 20K-word open and closed test conditions. This dictionary also includes the 1K-word Resource Management[15] vocabulary to allow cross-task tests with an existing database. MIT Lincoln Laboratory, as part of its text selection and preprocessing effort, has provided baseline open and closed test vocabularies based upon the test-set realized-vocabularies and the WFL for the 5K and 20K test sets. Lincoln has also provided 8 baseline bigram back-off[8,11] language models (5K/20K words × open/closed vocab. × verbalized/non-verbalized punct.) for research and cross-site comparative evaluation testing. Finally language model training data and utilities for manipulating the processed texts have been made available to the recording and CSR research sites.

NIST compiled the data from the three recording sites (MIT, SRI, and TI), formatted it, and shipped it to MIT where WORM CD-ROMS were produced for rapid distribution to the CSR development sites.

CONCLUSION

The WSJ Corpus and its supporting components have been very carefully and efficiently designed by the joint efforts of the DARPA SLS CSR Committee to support advanced strategic CSR research of many different types. It is hoped that eventually, these materials will be instrumental in facilitating the speech recognition research community to create spoken language technology capabilities suited to broad practical application.

REFERENCES

1. J. M. Baker, "DragonDictate-30K: Natural Language Speech Recognition with 30,000 Words," EUROSPEECH 89, Paris, September 1989.
2. J. M. Baker, Presentation at ESCA Workshop on Performance Evaluation and Databases Noordwijkerhout, Netherlands, September. 1989.
3. J. M. Baker, Presentation at the Kobe Workshop on Performance Evaluation and Databases,Kobe Japan November 1990
4. J. Allen., M. S. Hunnicutt, and D. Klatt, "From Text to Speech: The MITalk System, Cambridge University Press, New York, 1987.
5. J. L. Gauvain, L. F. Lamel, and M. Eskénazi, "Design Considerations and Text Selection for BREF, a large French Read-Speech Corpus," ICSLP 90, Kobe, Japan, November 1990.
6. H. W. Hon and K. F. Lee, "On Vocabulary-Independent Speech Modeling," Proc. ICASSP90, Albuquerque, New Mexico, April 1990.
7. F. Jelinek and R. Mercer, personal communication.
8. S. M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," ASSP-35, pp 400-401, March 1987.
9. M. Liberman, "Text on Tap: the ACL/DCI," Proceedings October, 1989 DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, October, 1989.
10. R. Mercer, personal communication.
11. D. B. Paul, "Experience with a Stack Decoder-Based HMM CSR and Back-Off N-Gram Language Models," Proc. DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, Feb. 1991.
12. D. B. Paul, J. K. Baker, and J. M. Baker, "On the Interaction Between True Source, Training, and Testing Language Models," Proceedings June 1990 Speech and Natural Language Workshop, Morgan Kaufmann Publishers, June, 1990.
13. D. B. Paul, CSR results presented at the October 91 SLS Mid-Term Workshop, CMU, October 1991.
14. M. Picheny, personal communication.
15. P. Price, W. Fisher, J. Bernstein, and D. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," ICASSP 88, New York, April 1988.
16. P. Price, The ATIS Common Task: Selection and Overview," Proceedings June 1990 Speech and Natural Language Workshop, Morgan Kaufmann Publishers, June, 1990.
17. B. Santorini, "Annotation Manual for the Penn Treebank Project," Technical Report, CIS Department, University of Pennsylvania, May 1990.

WSJ-spon Spontaneously spoken data. The subjects simulate dictating a short WSJ-like article. (Included in the WSJ-pilot database.)

Hansard-read Read data from the Hansard database.

Radiology-read Read radiology (medical) reports.

CALS-read (Computer-aided Acquisition & Logistic Support) Read repair manuals.

DART-read (Database Query for Material Routing) Read database queries.

USENET-read Read computer bulletin board messages.

NPR-read (National Public Radio) Read transcriptions of radio programs.

Blind-read Taped recordings for the blind.

Table 1. Proposed complementary datasets for the Large Vocabulary CSR database.

Vocab	Word Coverage
5K	91.7%
20K	97.7%
64K	99.6%
173K	100.0%

Table 2. Frequency-weighted upper-case word coverage (from the word-frequency list).

TRAINING:	SI-160	SI-16/SD-2400	LSD-9600
Train	160a ⋆ 260 = 41600	16b ⋆ 2400 = 38400	4b′ ⋆ 7200 = 28800
Adaptation	160a ⋆ 40 = 6400	16b ⋆ 40 = 640	

Est. total training data: SI-160: 86 hrs, SI-16: 79 hrs, SD-2400: 5 hrs/spkr, LSD-9600: 20 hrs/spkr

DEVELOPMENT TEST:	SI	SD
Read text, 5K	32c ⋆ 100 = 3200	16b ⋆ 100 = 1600
Read text, 20K	32c ⋆ 100 = 3200	16b ⋆ 100 = 1600
Spontaneous	32c ⋆ 100 = 3200	16b ⋆ 100 = 1600
Read spontaneous	32c ⋆ 100 = 3200	16b ⋆ 100 = 1600
Adaptation	32c ⋆ 40 = 1280	

EVALUATION TEST:	SI	SD
Read text, 5K	32d ⋆ 100 = 3200	16b ⋆ 100 = 1600
Read text, 20K	32d ⋆ 100 = 3200	16b ⋆ 100 = 1600
Spontaneous	32d ⋆ 100 = 3200	16b ⋆ 100 = 1600
Read spontaneous	32d ⋆ 100 = 3200	16b ⋆ 100 = 1600
Adaptation	32d ⋆ 40 = 1280	

Table 3. The plan for the WSJ portion of the full database. Format: no. spkr ⋆ no. sent = total no. sent. The letters following the number of speakers indicate the speaker sets (b′ is a subset of b). The data in all sections, except for adaptation, is half verbalized punctuation and half non-verbalized punctuation. Training times do not include the adaptation data. Times based on 7.4 sec/sentence. Total database size: 157K sentences=323 hrs=37 GB.

TRAINING:	SI-84	SI-12/SD-600	LSD-2400
Train	84a ⋆ 100† = 7240	12b ⋆ 600 = 7200	3b′ ⋆ 1800 = 5400
Adaptation	84a ⋆ 40 = 3660	8b ⋆ 40 = 320	

Total training data: SI-84: 15.3 hrs, SI-12: 14.3 hrs, SD-600 ~1.2 hrs/spkr, SD-2400: ~4.8 hrs/spkr
†Some speakers recorded 50 sentences.

DEVELOPMENT TEST:	SI	SD
Read text, 5K	10c ⋆ 80 = 800	12b ⋆ 80 = 960
Read text, 20K	10c ⋆ 80 = 800	12b ⋆ 80 = 960
Spontaneous	10c ⋆ 80 = 800	12b ⋆ 80 = 960
Read spontaneous	10c ⋆ 80 = 800	12b ⋆ 80 = 960
Adaptation	10c ⋆ 40 = 400	

EVALUATION TEST:	SI	SD
Read text, 5K	10d ⋆ 80 = 800	12b ⋆ 80 = 960
Read text, 20K	10d ⋆ 80 = 800	12b ⋆ 80 = 960
Spontaneous	10d ⋆ 80 = 800	12b ⋆ 80 = 960
Read spontaneous	10d ⋆ 80 = 800	12b ⋆ 80 = 960
Adaptation	10d ⋆ 40 = 400	

Table 4. The WSJ-Pilot database. Format: no. spkr ⋆ no. sent = total no. sent. The letters following the number of speakers indicate the speaker sets (b′ is a subset of b). The average sentence length is ~7.4 sec. (Verbalized punctuation sentences tend to be somewhat longer than average and non-verbalized punctuation sentences somewhat shorter.) The data in all sections, except for adaptation, is half verbalized punctuation and half non-verbalized punctuation. Training times do not include adaptation data. Total database size: 39K sent=80 hrs=9.2 GB.