Computer Speech and Language (1990) **4**, 193–202

A dictionary for a very large vocabulary word recognition system

Philip F. Seitz*, Vishwa N. Gupta†, Matthew Lennig†, Patrick Kenny, Li Deng‡, Douglas O'Shaughnessy, and Paul Mermelstein†

Institut National de la Recherche Scientifique—Télécommunications
3, place du Commerce, Verdun (Québec) H3E 1H6 Canada

Abstract

It is not too difficult to select a fairly small (on the order of 20 000 words) fixed recognition vocabulary that will cover over 99% of new input words when the task is limited to text in a specific knowledge domain and when one disregards names and acronyms. Achieving such a level of coverage is much more difficult when restrictions on knowledge domain and names are lifted, however. This report describes how we selected a 75 000-word English recognition vocabulary that covers over 98% of words in new newspaper text, including names and acronyms. Observations collected during the vocabulary selection process indicate the limiting factors for coverage of general knowledge domain text such as newspaper stories.

1. Introduction

In principle, there can never be an "unlimited vocabulary" word recognizer for English, any more than one person could have an unlimited vocabulary, since new words enter the vocabularies of all living languages on a continuous basis. However, it has become reasonable to expect an isolated word automatic speech recognizer to have a vocabulary that equals or exceeds that of an average person. This has been our goal during the development of a speaker-dependent, isolated-word recognition system (Gupta, Lennig & Mermelstein, 1988). Rather than having a particular speech-to-text application in mind for our system, such as office correspondence, we intend it to be capable of handling dictation in any knowledge domain that is not highly technical. A kind of text that is generally non-technical and that is wide-ranging in terms of vocabulary, is newspaper stories. In newspaper stories one encounters reports on activities in many different domains, e.g., politics, business, sports, and science and technology, each having its own specialized vocabulary; and of course the set of names and acronyms that

*Present address: Center for Auditory and Speech Sciences, Gallaudet University, Washington, DC 20002–3625 U.S.A.

†Also affiliated with Bell-Northern Research, Montreal.

‡Present address: Department of Electrical Engineering, University of Waterloo, Waterloo, Ontario H2L 3G1 Canada.

0885–2308/90/020193 + 10 $03.00/0

occur in newspaper stories is practically boundless. Therefore, in order to explore the outer limits of what is possible in text coverage by a recognition vocabulary, we set out to maximize text coverage on a new sample of 100 000 words of completely unrestricted newspaper text.

There are many reasons to study relationships among vocabularies' size, linguistic characteristics and text coverage, and there are many applications for very large vocabularies with high general knowledge domain text coverage. Besides speech-to-text, there are applications in text-to-speech, parsing, machine translation, and related computer speech and language areas. The present need for vocabularies with very high general knowledge domain text coverage has been the impetus for re-addressing traditional issues surrounding the estimation of the size, characteristics and text coverage of vocabularies. Present-day computing resources allow us to perform larger experiments more easily than those undertaken by lexicographers, statisticians, psychologists and linguists of 30 years ago.

While the methods and findings of the present report bear upon a number of computer speech and language applications, they have a special relevance to current research in automatic speech recognition because of the particularly severe limitations on vocabulary size, and thus knowledge domain, that are characteristic of contempory recognition systems. Given a very large, transcribed and tagged vocabulary with demonstrably high coverage of unrestricted text, we can begin to address new challenges in automatic speech recognition, such as how to perform an efficient heuristic search through the very large space (Gupta, Lennig & Mermelstein, 1988), how to construct statistical language models from unrestricted text (Gupta, Lennig & Mermelstein, 1989; O'Shaughnessy, Gupta, Lennig, Seitz & Mermelstein, 1990), and how to design a phonological component appropriate for a very large and linguistically diverse vocabulary (Seitz *et al.*, 1990).

The goal of maximizing the text coverage of a recognition vocabulary is equivalent to the goal of minimizing recognition errors caused by input words not being in the vocabulary. For a word recognizer, the input of an out-of-vocabulary word is always an error, no matter how high the probability is that the input word was generated by a string of the system's phoneme models. Out-of-vocabulary errors are particularly troublesome for the statistical language-model component of a recognition system, since there is no way for the language model to correct them and since the wrong word will be used to compute the probabilities of following words. Although very large vocabulary word recognition systems must include real-time and non-real-time interfaces for users to correct recognition errors, to set default spellings for names, and to add words to the dictionary, it is desirable to equip a system with a dictionary that will require as little attention from the user as possible. It is also desirable not to place any *a priori* restrictions on the user with respect to his/her English vocabulary. When these two desiderata are addressed simultaneously, the burden on the dictionary component of the system becomes heavy. However, we will show in this paper that it is possible to achieve a very low incidence of out-of-vocabulary words even on unrestricted newspaper text.

Jelinek (1985) has described the process of selecting and "personalizing" a recognition vocabulary for IBM's Discrete Dictation Recognizer. For the domain of office correspondence, the IBM group was able to obtain 99·5% coverage of new text, *excluding names and acronyms*, from a 20 000 word recognition vocabulary built by processing 1 300 000 words of office correspondence text. That is, there were 20 000 *distinct* words (excluding names and acronyms) in 1 300 000 words of this text, and there is an estimated probability of 0·995 that a word of new text from the office correspondence

domain is in the set of 20 000 words. The vocabulary selection process that we report in this paper is in a similar vein, but our domain is newspaper stories, and we include names and acronyms.

In the next two sections, we describe the procedures we used to increase text coverage, while keeping the size of the vocabulary manageable by pruning useless words. In the following section, we present the results of tests on the text coverage of the vocabulary at each stage of its development. In the final section, we describe some of the properties of the new words that enter the vocabulary on successive stages of augmentation. The properties of out-of-vocabulary words indicate the ultimate limiting factors for a vocabulary's coverage of text in a general knowledge domain such as newspaper stories.

2. Procedures for building the dictionary

2.1. Preliminaries

Our recognizer has no morphological capability, so words that are spelled differently are treated as separate words even when they are differently inflected forms of the same stem, e.g. *Bourassa—Bourassa's*. On the other hand, we have established a correspondence table for words with multiple valid spellings (e.g. *labor—labour*) so the alternate spellings are not counted as distinct words. The alternate spelling correspondence table is especially useful in the case of Canadian English texts, where British and U.S.A. spelling conventions are in competing use (Ireland, 1979). Most hyphens are treated as spaces, that is as word delimiters, so forms such as *pseudo-intellectual* consist of two words; however, a number of hyphenated words, such as *co-operate*, which have alternate unhyphenated spellings (here *cooperate*), are treated as single words and are included in the alternate spelling correspondence table.

2.2. The initial fixed vocabulary

It is reasonable as well as practical to start to build a general knowledge domain vocabulary around an existing dictionary. We used a machine-readable version of *Webster's Seventh New Collegiate Dictionary* (G. & C. Merriam Company, 1967)—henceforth *W7*—which contains 64 920 regular entries (that is, top entries, not counting derived forms listed as second entries under top entries, and not counting words in the dictionary's appendices). As an efficient way to supplement the *W7* vocabulary with frequent inflected forms of words and with names, we included in our initial fixed vocabulary the 12 753 words in Francis & Kucera's *Standard Sample of Present-Day American English* (the "Brown Corpus"; Francis & Kucera, 1979) which do not occur as top entries in *W7*. We also added 376 common given names from an appendix to *W7*, and 123 words that occurred in an approximately 5000 word sample of various newspaper stories and other texts that we were using as scripts for collecting speech data from our subjects. The resulting vocabulary of 78 172 words constituted the initial fixed vocabulary, to be augmented and pruned by methods described below.

2.3. Methods for augmenting and pruning the vocabulary

In order to augment the initial fixed vocabulary with words that improve its coverage, and to discover which words in the initial vocabulary do not contribute to coverage, we

processed 10 300 000 words of text from newspaper stories. We processed these words in four batches. The first, large batch was a 10 000 000 word sample of text from the *Globe & Mail* (Toronto), comprising about 4 months of that daily newspaper's publication. Each of the next three, smaller batches consisted of non-overlapping 100 000-word samples from the *Gazette* (Montreal). For each sample, input words were compared against the initial vocabulary, and if an input word was not in the current vocabulary, it was tagged as a new word. Frequency counts of new words were made for each sample of text, and frequencies of matches with input words were accumulated for all words in the current vocabulary. New words discovered on each run were added to the vocabulary, and thus constituted parts of the vocabularies for subsequent runs.

3. Incremental increases in vocabulary size, and vocabulary pruning

3.1. Vocabulary size increases

For the first vocabulary augmentation run, in which 10 000 000 words from the *Globe & Mail* were processed, we retained a total of 20 000 new words; 10 000 were words whose first letters were upper-case and which were not the first word in a sentence (suggesting names and acronyms), and 10 000 whose first letter was lower-case. Henceforth we will refer to all of these upper-case words simply as "names." Although we discovered more than 20 000 new words on the run on 10 000 000 words, we retained only the top 10 000 words in frequency for each category in order to keep the task within manageable bounds. Table I shows that there is a skewed distribution of number of occurrences in the sample for these new words; some words that occur very frequently in the text sample were not in the initial vocabulary (e.g. *Mulroney*, the name of the current prime minister of Canada), but most of the new words, both names and lower-case, have probabilities of occurrence that we can estimate from this sample at around once in one million words. This indicates that most of the improvement in text coverage gained by adding these news words will be due to a fairly small set of the most frequently occurring of these words.

In the three vocabulary augmentation runs using 100 000-word samples from *The Gazette*, which followed the first run on 10 million words, there were totals of 2411, 2175,

TABLE I. Central tendencies of number of occurrences of the 10 000 most frequently occurring upper-case and 10 000 most frequently occurring lower-case new words discovered in a vocabulary augmentation run on 10 000 000 words of text from *Globe & Mail* stories

	Names	Lower-case
Cumulative occurrence	351 949	101 944
Maximum No. of occurrences	5049	3151
Minimum No. of occurrences	6	1
Mean No. of occurrences	35·12	10·19
Standard deviation	117	50
Median No. of occurrences	13	4
Modal No. of occurrences	7	2
Inner-quartile range	20	5

and 1957 out-of-vocabulary words discovered. In all three of these runs, names made up about 85% of the out-of-vocabulary words. The distributions of numbers of occurrences of the out-of-vocabulary words were much less skewed in these later runs than in the first run, since all of the very frequently occurring names and lower-case words had already been encountered in the first run on 10 000 000 words.

At the end of the four cycles of vocabulary augmentation described above, the vocabulary size stood at 104 715 words.

3.2. Pruning and implementing the vocabulary

A total of 29 150 words from *W7* were never observed in any of the text samples, and were deleted from the final vocabulary. After inspecting the new words from the text samples, we identified several categories of words that we also deleted from the final vocabulary. There were mis-spellings and words which are not acronyms but which appeared in all upper-case letters (usually because they were the lead word in a story). Words in both these categories were first corrected (or converted to the appropriate case) and checked against the vocabulary. If they were already included in the vocabulary in their corrected forms, their original forms were deleted; otherwise, their corrected forms were included. This accounted for 201 deletions. There was a set of non-pronounceable abbreviations which had to be cross-referenced to full forms (e.g. *Que.* stands for *Quebec*). A speech recognition vocabulary must include only pronounceable words, so, after confirming that the full forms corresponding to these abbreviations were in the dictionary, 113 of them were deleted. Likewise, we retained only one spelling for each of the 25 words with alternate spellings. Of course, all the new words had to be given transcriptions in terms of symbols that can be related unambiguously by rule to the units of recognition. In the course of this transcription task, the linguist performing it encountered 43 names which he could not confidently transcribe, and these were subsequently deleted from the vocabulary as well.

Following the four augmentation cycles and the pruning procedures described above, the final recognition vocabulary consists of 75 183 words.

4. Text coverage by the recognition vocabulary

4.1. Text coverage estimation procedure

To estimate the text coverage of the vocabulary at each stage of its augmentation, we used a fourth 100 000-word sample of text from *The Gazette*.[1] We compared each word from this sample first against the initial fixed vocabulary, and then against each of the augmented vocabularies *minus* all the words pruned from the final vocabulary.

4.2. Incremental text coverage

Table II traces the text coverage of the recognition vocabulary through the stages of its development.

Figure 1 presents the data from Table II in graphic form.

[1]Ideally, we would have liked to test the text coverage of our vocabulary on a sample of text from a newspaper other than one of those used for vocabulary augmentation. Unfortunately, we were not successful in our attempts to obtain computer-readable text from other newspapers.

TABLE II. Text coverage by the recognition vocabulary after each stage of its augmentation

Augmentation	Initial	*Globe & Mail*	*Gazette*1	*Gazette*2	*Gazette*3
Per cent text coverage	93·52	97·59	97·82	98·04	98·12
Lower-case only	98·68	99·67	99·68	99·68	99·69
Vocabulary size ('000s)	49	69	71·5	73·5	75

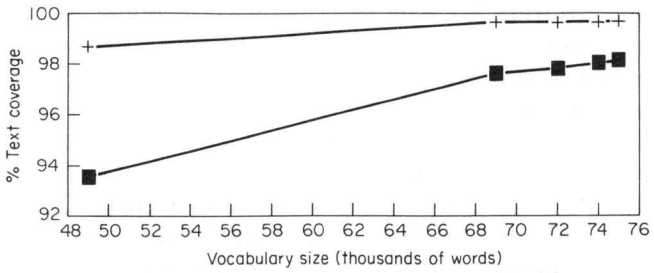

Figure 1. Text coverage by the recognition vocabulary after each stage of its augmentation (cf. Table II). —■— Lower-case and names, —+—Lower-case only.

The initial fixed vocabulary, whose size is around 49 000 after pruning, is clearly lacking in coverage of names, as indicated by the large difference between overall text coverage and coverage of lower-case words only. Despite the large size of the initial fixed vocabulary, and despite the fact that approximately one-fourth of it consists of words from the Brown Corpus, its lower-case-only text coverage is worse than the 99·5% coverage that Jelinek (1985) reported with a 20 000-word vocabulary built by processing 1 300 000 words of text. This difference is probably due to the "personalization" factor: Whereas Jelinek's group built their vocabulary using text from the domain of their system's application (viz. office correspondence), our initial vocabulary had a much smaller amount of "training" on text from the domain that its coverage was tested on (viz. newspaper stories). There are some newspaper stories represented in the Brown Corpus, but they are American (rather than Canadian) and are over 25 years old.

We should expect that our fixed vocabulary will have poorer coverage of a 100 000-word sample of text from the 1995 publication year of the *Wall Street Journal* than it does on our 100 000-word sample of text from *The Gazette* which is from the same publication year as the 300 000 words from *The Gazette* that were used to augment the vocabulary. Thus the "personalization" issue remains even for text in a very general knowledge domain.

There is a large improvement in text coverage after the first, large augmentation of the vocabulary, especially with respect to names. We expected to see another leap in coverage of names after the first *Gazette* augmentation, since Toronto and Montreal newspapers might be expected to have somewhat different frequently-occurring sets of proper names. This was not the case, however. Local names apparently occur much less

frequently than "global" names in newspaper text, so local names are relatively unimportant for newspaper text coverage.

By the last augmentation, the text coverage increase curve flattens out. By this point we can expect less than 2% recognition errors due to out-of-vocabulary words. The probability that an out-of-vocabulary word is a name can be estimated as 0·83 (see Table II).

The passage of newspaper text shown in Table III illustrates the kinds of words that entered the vocabulary from various sources.

TABLE III. Example text from *Globe & Mail* (9 March 1989), showing words from *Webster's* in Roman type, words from the *Brown Corpus* in italics, and words from vocabulary augmentation on newspaper text in bold face. The word in all capitals is not in the vocabulary.

Regina: Saskatchewan Premier Grant **Devine**, *declaring* that socialism is about to meet its Waterloo, *unveiled plans* yesterday to offer *shares* in three provincial Crown *corporations* with *assets* worth about two billion *dollars*. In a Throne Speech to open a new legislative session, his Government *announced* that *investors* will be *offered shares* in the Potash Corporation of **Saskatchewan**, **Saskatchewan** Government Insurance, and the provincial natural gas monopoly known as SASKENERGY. Mr **Devine** *predicted* a tremendous battle with the opposition New *Democrats* over his **privatization** *plans*.

5. Limiting factors for general knowledge domain text coverage

Given that it is in principle impossible to achieve 100% general knowledge domain text coverage by a recognition vocabulary—except, of course, due to sampling error—we would like to know what kinds of out-of-vocabulary words are encountered even after extensive augmentation of a large vocabulary. Knowing the characteristics of the ultimately irreducible residue of out-of-vocabulary words will help us to conceive design improvements for the vocabulary as well as heuristics for recovering from out-of-vocabulary recognition errors.

5.1. *Methods for analyzing factors involved in text coverage*

For each vocabulary augmentation run, we classified new words according to sets of morpho-syntactic/semantic/lexical categories appropriate to lower-case words and names. These two sets of categories, which we established through investigation of new words discovered on vocabulary augmentation runs, are given in Table IV.

In order to keep the analysis manageable, we only classified the 100 top-frequency new words from each of the lower-case and name sets for each augmentation run. On the three augmentation runs involving 100 000 words samples, the words classified were the N new words with frequencies over 1 plus a random sample of the (100-N) new words with frequencies of 1. This procedure was followed for names and lower-case words separately.

TABLE IV. Categories of out-of-vocabulary words

Categories of lower-case out-of-vocabulary words
1. Abbreviations cross-referenced to words in vocabulary (e.g. *pfd* = "preferred")
2. Spelling variations of words in vocabulary (e.g. *labour* = "labor")
3. Mis-spellings
4. Word fragments from hyphenated words (e.g. *co-, -tech*)
5. Fragments due mainly to French (e.g. *de, le*)
6. Plurals of common nouns whose stems are already in vocabulary
7. Stems of nouns/verbs/adjectives/adverbs (e.g. *goalminder*)
8. Inflected forms of verbs whose stems are in vocabulary
9. Possessive forms of nouns whose stems are in the dictionary
10. Clearly French words (e.g. *meubles, renseignements*)
Categories of out-of-vocabulary names
1. Abbreviations cross-referenced to words in vocabulary (e.g. *Rd.*)
2. Non-inflected place names (e.g. *Oshawa*)
3. Non-inflected family names (e.g. *Tremblay*)
4. Non-inflected acronyms (e.g. *TSE*)
5. Non-inflected names of organizations (e.g. *Unisys*)
6. Inflected forms of all names (e.g. *Bears', Chicago's*)
7. Name fragments (e.g. *Buenos* from "Buenos Aires")
8. Capitalized adjectives, all forms (e.g. *Manitoban*)

5.2. *Characteristics of out-of-vocabulary words in successive vocabularies*

Figs. 2 and 3 show the relative proportions of lower-case and name out-of-vocabulary words discovered on successive augmentation runs. Recall that all of the out-of-vocabulary words discovered on a given run are added to the vocabulary, and thus constitute part of the vocabulary for subsequent runs.

The trend in incremental coverage of the major categories of lower-case words, shown in Fig. 2, is what we might expect given the vastness and constantly changing nature of the English lexicon. Plural forms of nouns whose stems are already in the vocabulary—the most frequent category of new words involving inflectional morphology—make up a steadily declining percentage of out-of-vocabulary words over the series of augmentations. At the same time, out-of-vocabulary stems (singular nouns, least-marked forms of verbs and adjectives) steadily increase as a proportion of all new words. As text coverage increases, the likelihood decreases that out-of-vocabulary words are inflected forms of stems already in the vocabulary. This situation has implications for the methods we choose to improve a general knowledge domain recognition vocabulary's coverage while not letting its size balloon. Specifically, it informs us that text coverage will not be much improved if we automatically generate and include all of a stem's inflected forms. For example, among the 75 183 words in the vocabulary, there are 28 155 singular common nouns and only 5764 plural common nouns; by providing plural forms to the 22 391 singular nouns lacking corresponding plurals, we would increase the vocabulary's size by around 30%, a size increase not justified by the fact that there is a probability of only about 0·0191 that an out-of-vocabulary word is a plural of a stem already in the vocabulary. Once acceptable coverage of lower-case words has been achieved through incremental augmentation, it is not profitable to undertake determinis-

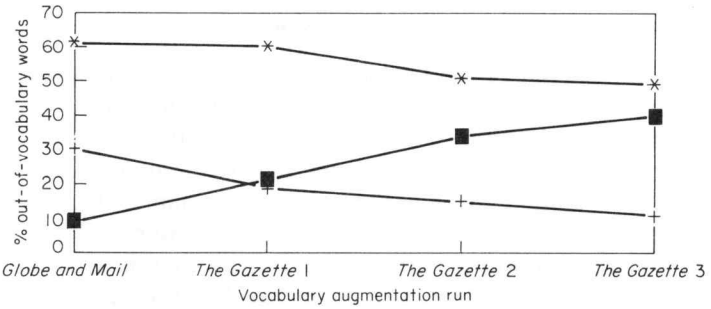

Figure 2. Relative proportions of major categories of lower-case out-of-vocabulary words on successive vocabulary augmentation runs. —■— New Stems, —+— Plural nouns, —*— All others.

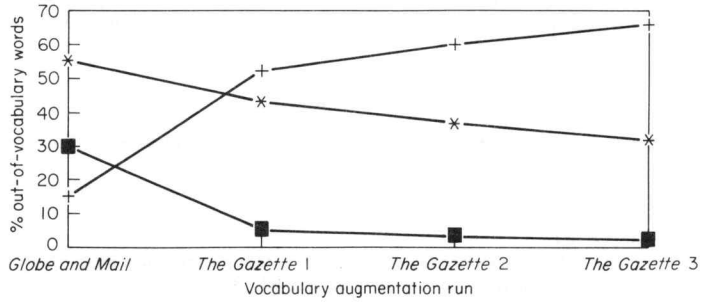

Figure 3. Relative proportions of major categories of out-of-vocabulary names on successive vocabulary augmentation runs. —■— Place names, —+— Family names, —*— All others.

tic augmentation of the vocabulary with respect to inflected forms, since the major source of out-of-vocabulary recognition errors involving lower-case words will be due to new stems.

The trend in characteristics of out-of-vocabulary names, shown in Fig. 3, is even clearer than the trend for lower-case words. Over the series of vocabulary augmentations, the proportion of out-of-vocabulary names that are family names rises from 15% to 66%, while place names, which at the first augmentation were the most frequently occurring category of out-of-vocabulary word, fall from 30% to 2% of all out-of-vocabulary names. For the recognition vocabulary as it stands, we can estimate that there is a probability of 0·54 that an out-of-vocabulary word is a family name. This prominent characteristic of family names suggests that they be dealt with on a special basis. For example, family names might be the one area of the recognition vocabulary that a user or user group needs to "personalize." Alternatively, users might need to be informed of a single restriction on their input to the recognizer: "unusual" family names should be spelled in order to obtain the best performance. Fortunately, these restrictions place no more burden on the user than he/she could expect to encounter with a human stenographer.

6. Conclusion

In his influential discussion of the problem of selection of a recognition vocabulary, Jelinek (1985) wrote that a large, fixed recognition vocabulary can cover only part of the words that users will need to input; the non-covered vocabulary, according to Jelinek, "...must be obtained through active participation by the user." Our objective in the development of a recognition vocabulary has been to minimize the active participation in vocabulary selection required of the user. To this end, we have lifted the restrictions on knowledge domain and proper names and acronyms that have been built into other large recognition vocabularies. We have demonstrated that it is possible to achieve over 98% text coverage—which we consider adequate for most speech-to-text applications—in a general knowledge domain, including names and acronyms, with a fixed vocabulary of 75 000 words.

This research was supported by the Natural Sciences and Engineering Research Council of Canada. The authors express their thanks to the C. & G. Merriam Company, The Toronto *Globe & Mail*, and the Montreal *Gazette* for providing the computer-readable texts used in this study.

References

C. & G. Merriam Company. (1967). *Webster's Seventh New Collegiate Dictionary*. Springfield, Massachusetts, U.S.A.

Francis, W. N. & Kucera, H. (1979). *Manual of Information to Accompany a Standard Sample of Present-Day Edited American English, for Use with Digital Computers*. Brown University Department of Linguistics, Providence. RI.

Gupta, V. N., Lennig, M. & Mermelstein, P. (1988). Fast search strategy in a large vocabulary word recognizer. *Journal of the Acoustical Society of America*, **84**, 2007–2017.

Gupta, V. N., Lennig, M. & Mermelstein, P. (1989). A language model for very large vocabulary speech recognition. Unpublished manuscript.

Ireland, R. J. (1979). Canadian Spelling: An Empirical and Historical Survey of Selected Words. Ph.D. Thesis, York University, York, Ont.

Jelinek, F. (1985). The development of an experimental discrete dictation recognizer. *Proceedings of the IEEE*, **73**, 1616–1624.

The Montreal *Gazette* (1985–1987). Infomart, Montreal, Province of Quebec, Canada.

O'Shaughnessy, D., Gupta, V., Lennig, M., Seitz, F. & Mermelstein, P. (1989). Language modeling for very-large-vocabulary speech recognition. *Journal of the Acoustical Society of America*, **86**, 575.

Seitz, P. F., Gupta, V. N., Lennig, M., Deng, L., Kenny, P., O'Shaughnessy, D., and Mermelstein, P. (1990). Phonological rule set complexity as a factor in the performance of a very large vocabulary automatic word recognition system. Submitted to *Journal of the Acoustical Society of America*.

The Toronto *Globe & Mail* (1985). Infoglobe, Toronto, Ontario, Canada.