Speech, Hearing and Language: work in progress

Volume 14

STUDIES IN THE STATISTICAL MODELLING OF DIALOGUE TURN PAIRS IN THE BRITISH NATIONAL CORPUS

Gordon HUNTER and Mark HUCKVALE

Department of Phonetics and Linguistics
UNIVERSITY COLLEGE LONDON

STUDIES IN THE STATISTICAL MODELLING OF DIALOGUE TURN PAIRS IN THE BRITISH NATIONAL CORPUS

Gordon HUNTER and Mark HUCKVALE

Abstract

This article describes some preliminary investigations into the statistical properties of the transcribed dialogues that were collected for the British National Corpus of English. Our aim has been to look for evidence of linguistic structure which could be used to build better statistical language models for spontaneous human-human dialogues. We have concentrated on pairs of successive, relatively short dialogue turns. We find significant differences in the lexical distributions for dialogues compared to written text, as expected. Further experiments using cache, trigger and cluster-based models applied to pairs of turns found that interpolating such models with a standard trigram model resulted in improvements in perplexity compared with the perplexity scores obtained using a trigram model alone.

1. Introduction

In recent years, practical spoken language interfaces for human-machine interaction, and dialogue systems in particular, have become technologically feasible [12]. The aims of such systems should include allowing users to communicate with the system in as natural a way as possible, minimising effort on the part of, and inconvenience to, the user [8, 13, 14]. However, this should not be at the expense of the system's reliability in terms of recognising the user's responses correctly and taking appropriate action on them.

Thus, the study of spoken dialogue is important in speech technology for a number of reasons: firstly to obtain better word recognition performance in enquiry services, secondly in creating more natural man-machine interactions, and thirdly in understanding the structure and function of dialogue acts. For example, Jurafsky et al. [7] showed that classification of speaker turns as dialogue acts was useful in reducing word errors in those turns. Karis and Dobroth [8] argue that better user interfaces will come from systems which follow established norms of turn taking in human-human dialogues. Cohen [3] demonstrates that further work is needed to determine whether dialogue should be modelled as a grammar, as a plan, or as a co-operative activity.

Spontaneous human-human dialogue is very different in form from written text or even spoken monologue. Although both a text and a dialogue might be assigned to topic areas, the way in which a topic is introduced, presented and argued will be different in dialogue. Dialogue introduces new concepts such as turn-taking and co-operation; and the distribution of phatic utterances and question/answer pairs will be quite different.

Turn-taking in particular is clearly a very important aspect of dialogue, and the above arguments suggest that better modelling of dialogue turns will play a positive role in developing more successful human-machine interfaces. Some studies [15,16] have employed a "Wizard of Oz" approach, where the subject acts as a "user" interacting with the system, presumed to be a computer. The advantages and disadvantages of this approach have been discussed by Jonsson & Dahlback [17]. It assumes that the resulting "dialogues" will be typical of human-machine interaction in their nature.

This would appear to run against the aim of permitting the user to talk to the system in as natural a manner as possible, i.e. of requiring the interface to mimic human-human dialogue as well as possible. The alternative is therefore to use material recorded and/or transcribed from spontaneous human-human dialogue.

Adopting the latter approach, we predict that the statistical modelling of spontaneous dialogue would benefit from being made sensitive to its special character. For example, what lexical or structural relationships occur between adjacent turns; how can the end of a turn or the end of a dialogue be identified; how is back-channel information synchronised to content?

Such questions can now start to be addressed in statistical language modelling because of the recent availability of large corpora of transcribed dialogues. Table 1 gives summary information about some English dialogue corpora.

Corpus	English	Signal	Speakers	Style
SWITCHBOARD	American	Telephone	2400	Given topic
CALLHOME-English	Northern American	Telephone	120	Conversation
CALLFRIEND-English	American	Telephone	60	Conversation
BRAMSHILL	British	Microphone	~200	Given topic
HCRC Maptask	British	Microphone	64	Given topic
BNC	British	N/A	> 124	Conversation

Table 1: *Some corpora of transcribed dialogues in English*

The SWITCHBOARD corpus [5] has been the target of a lot of work in language modelling, for example work in the recognition of dialogue acts [7] or in the use of grammatical constraints for statistical language models [2]. However this corpus is quite small in text processing terms. It contains only about 3 million words, and was collected by prompting the speakers with one of about 70 possible topics of conversation. On the other hand, the spoken dialogue portion of the British National Corpus (BNC) [1] contains about 7 million words and was collected from everyday conversation. This corpus appears to be an interesting source of information about the structure and form of everyday dialogue, although the audio recordings themselves were not collected under controlled conditions and may not be useful for recognition applications.

This paper describes some initial studies to characterise the statistical properties of the spoken dialogue portion of the BNC, with the aim of investigating the extent to which surface lexical information in one turn can be used to predict the lexical content of the following turn in the same dialogue. In the context of human-computer dialogue systems, the first turn can be interpreted as the computer's (known) prompt and the following turn as the human user's response, which the computer must interpret. A subsidiary aim is to learn more about the structure of human-human dialogue through statistical models and, in the longer term, correlate our findings with the outcomes of the linguistic theory of conversational analysis. For example, does the presence of a

wh- question in one turn correlate well with words characteristic of an answer in the following turn? In this present study, we have looked at the lexical distribution of dialogue turns, the text perplexity using trigram language models and the utility of cache, trigger pair and cluster-based language models.

2. BNC Dialogue Data

The British National Corpus [1] is a large database of modern British English, compiled during the early 1990's. It is composed of both written material (totalling around 90 million words), transcriptions of spoken monologue (around 1.9 million words) and transcriptions of spoken dialogue (around 7.7 million words). The dialogue material was recorded in 672 different situations, from both "context governed" sources (business meetings, medical consultations, etc.) involving an unspecified total number of speakers and "demographic sampled" material from spontaneous conversations involving 124 speakers designed to be representative of the UK's speakers of British English in terms of age, gender, social group and region. Thus, the dialogue material is comprised of 672 distinct transcriptions, containing a total of over 880,000 sentences and over 7.7 million words. The audio recordings have been deposited at the National Sound Archives of the British Library.

The BNC spoken material was orthographically transcribed, word-tagged using an automatic system and "marked-up" using Standard Generalised Mark-up Language (SGML). The SGML not only indicates which speakers are involved in a conversation, but also their social relationship. The mark-up also indicates dysfluencies and occasions where the speech of the two participants overlapped. In the studies reported here we have pre-processed the transcriptions, looking only at sections of transcription involving a single pair of speakers, and starting a new logical “dialogue” whenever a new speaker joined the conversation. We have made all overlapping speech sequential, removed dysfluencies and then deleted turns which became empty as a consequence of these changes. We hope to return to these issues in later studies. We have not used the word-tag information.

	Minimum	Mode	Median	Mean	Maximum
Dialogues in a file	1	1	20.5	136.4	3115
Turns in a dialogue	1*	2	2	6.19	2326
Words in a dialogue	1*	9	19	79.36	21123
Words in a turn	1	1	5.5	13.0	18575
Proportion of words in a dialogue by first speaker	0.00*	0.500	0.500	0.499	1.00*

Table 2: *Some summary descriptive statistics for the BNC dialogue material.*

* These "dialogues" are clearly not true dialogues. Such "pseudo-dialogues" are probably due to an error in the transcription or mark-up within the BNC material.

Some descriptive statistics relating to the content of the BNC dialogue material in terms of dialogues, turns and words are shown in Table 2. Many of the files in the BNC contain contributions from more than 2 speakers and thus a single file may contain several logical "dialogues" as defined above. In fact, the length of the files

varied extensively - ranging from very short 2 speaker interchanges to long debates involving many speakers – in a highly skewed manner. Hence, despite the modal number of dialogues per file being 1, the median number was 20.5, and 70th and 80th percentiles were 66.1 and 154.4 respectively. One file contained over 3100 distinct dialogues, according to the above definition. These unusually high values are in part due to this way in which a "dialogue" was defined, which was largely for simplicity and computational convenience. However, a consequence of this is that a conversation involving (say) just 3 speakers could consist of a large number of dialogues according to our definition. Each time one speaker of the three is replaced by another results in the start of a new dialogue. In all, the 672 BNC files contained a total of 91,650 such dialogues. In a similar manner, although less directly affected by the way we defined a dialogue, the number of turns within a single dialogue showed wide, highly non-normal, variation. The modal number of turns in a dialogue was 2, as was the median. The mean number was 6.2, and 90th percentile 8. However, one dialogue contained as many as 2,326 turns. The number of words in a dialogue also showed extremes. The modal number was just 9, and median 19, but the mean was 79.4 (larger than the 80th percentile which was 55), and one dialogue contained 20,023 words. The dialogues varied considerably in their "balance" - the proportion of word spoken by each speaker. However, there was no evidence that, on the whole, either the first or second speaker to enter the conversation tended to dominate. The mean, mode and median proportion of words spoken by the first speaker were all 0.50, with 80% of the dialogues having the first speaker contributing between 11.7% and 88.9% of the dialogue's total words. These cases were deemed to be reasonably well-balanced.

3. Lexical Distribution

The BNC dialogue material contains a total of 49,989 distinct word types. An equivalently-sized sample of the BNC text corpus contained 104,827 types. An 80Mword sample of the BNC text corpus contained 352,860 types. There is also a marked difference between dialogue and text in the growth in coverage with size of lexicon, see Figure 1.

The smaller dialogue lexicon provides two advantages for dialogue language modelling. Firstly, since it includes all the distinct words which might be encountered within both the training and the test data, there is no issue about how to deal with out-of-vocabulary words – a feature which often causes problems in statistical language modelling in situations where the vocabulary required, due to size constraints, may not be closed. Secondly, the relatively small vocabulary means there will be fewer problems in smoothing n-gram language models.

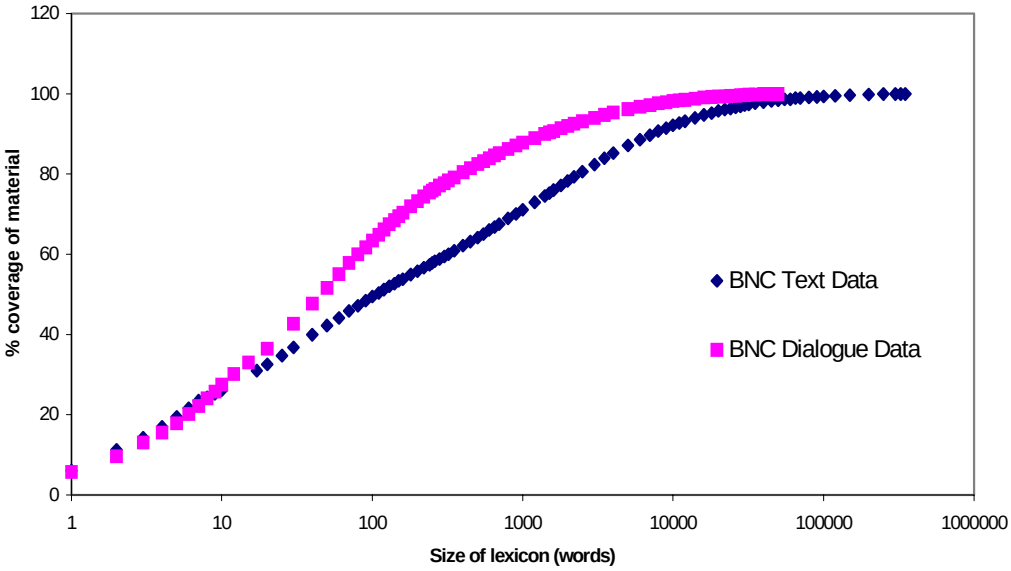

Figure 1: *The proportions of BNC text and dialogue material covered by lexica of various sizes.*

4. Data Preparation and Baseline Model

Our preliminary investigations indicated that, although identified as "dialogue material", a considerable portion of the spoken component of the BNC consisted of very long speaker turns - in effect, to a close approximation, monologue. Initial studies indicated that little benefit could be obtained from using statistical language models based on trigger pairs or a cache between turns of such material. Instead, we decided to focus on pairs of successive, relatively short turns from the same dialogue.

The set of BNC dialogue material files, after the SGML mark-up had been removed, were divided into pairs of successive speaker turns, and only pairs totalling less than 200 words were retained for use in the experiments described here. This yielded at total of approximately 470,000 pairs of turns.

To obtain baseline perplexity figures for the dialogue data, trigram language models were constructed using the CMU-Cambridge Language Modelling Toolkit [4]. The dialogue data was divided into ten subsets of each of Training (423,000 turn pairs), Development (23,500 turn pairs) and Evaluation (23,500 turn pairs) sets using a cross-validation procedure. The dialogue lexicon was used throughout. Good-Turing smoothing [23] was applied to correct for biasing the models too much towards commonly observed words, with singleton cut-off used for computational convenience.

Separate trigram models were constructed for the first turns of pairs and for second turns of pairs. Averaged across the 10 rotations of the validation process, the perplexity scores obtained for these models applied to the test data were almost identical: 187.61 for first turns and 187.69 for second turns. (This contrasts with the much higher figure of 289.61 obtained for a model trained on both turns of the pairs, which is probably due to replication of turns between successive pairs). The trigram model for the second turns was used both as a baseline and for interpolation with other models in the later experiments.

5. Cache-Based Models

Previous modelling results with text data have shown that cache models are a simple but effective means of tracking how the lexical likelihoods vary with the topic of the current document, e.g. [6,9]. A typical cache model maintains a history of recent words used in the current document, and estimates a dynamic unigram language model using solely those words. This is then interpolated with a static trigram model built from a large quantity of topic-independent text. Experimenting with various cache sizes, Clarkson & Robinson [18] found that a cache of 500 words performed better than any other.

It might be expected that our dialogues would also benefit from a similar approach: that the likelihood of words used in the dialogue would be affected by which words had been used earlier. In particular, it would seem likely that the content of the second of a pair of consecutive dialogue turns would be closely related to that of the first turn of the same pair. To evaluate this, we constructed a cache model applied to pairs of dialogue turns, with a sliding window of at most 500 words used to construct two caches: one for the first turn of the pair, the other for that portion of the current (second) turn which had already been encountered. The caches were reset at the start of each new pair of turns. The resulting cache models were interpolated with a trigram model trained on the content of second turns of pairs. Interpolation parameters were learned using the *interp* program in the CMU toolkit, which performs Expectation-Maximisation (EM) training [19,20] on matched probability streams. The interpolation parameters were trained on the Development test sets of the dialogue data and applied to the Evaluation sets using a 10-fold cross-validation procedure.

The average perplexity, weighted according to the sizes of the datasets, across the ten rotations, was found to be 166.64 for the interpolated cache - trigram model, in contrast to the corresponding average perplexity of 187.69 for the trigram model when applied to the same data alone. This represents an improvement of 11.2% over the baseline (trigram only) figure.

6. Models Based on Word Trigger Pairs

Since we expect dialogue turn pairs to have particular lexical structural patterns as well as well-defined topics, we have investigated the occurrence and application of word-trigger pairs [10] across turns in a dialogue.

A sliding window containing a maximum of 500 words previous to the target word, either in the current turn or the previous turn, was chosen as the word-trigger history, with the window being reset after every pair of dialogue turns. We did not include in the history the two words immediately previous to the target (as these would form part of the trigram model with which the trigger model would later be interpolated). We also only looked at intermediate frequency words for the selection of triggers: our arbitrary criterion was to use a lexicon of only 10,000 words which consisted of the most commonly-found words, excluding the 50 most common. From the training set of 7Mword of dialogue transcription we found 4,957 pairs with an average mutual information [10] greater than 10^{-5} bits. From an equivalently sized quantity of ordinary text material from the BNC we found 3,209 pairs. Only 900 pairs were common to the two styles.

From the list of potential trigger pairs, a criterion of a maximum of 10 triggering words per target word was imposed and the 2,800 triggers both satisfying this

constraint and showing the highest mutual information were used to construct an exponential probability model. The 2,800 parameters of the model were evaluated from training data using the maximum entropy framework [10] and the Generalised Iterative Scaling (GIS) algorithm [22]. The resulting exponential model was interpolated with the baseline trigram model for the content of second turns of pairs using the EM algorithm [19,20] and a reserved set of data.

Across 10 cross-validation rotations, the interpolated trigram-trigger model applied to test data gave an average perplexity of 182.81, approximately 2.6% lower than the corresponding average baseline perplexity figure, 187.69, for the trigram model alone applied to the same test data. This result is rather disappointing compared with the improvement obtained for the interpolated trigram-cache model. A trigger-based model should be able to exploit more general correlations between words than the simple repetitions of words used by cache models, so it would have been expected that the trigger-based model would have given a better result.

7. Cluster-Based Models

Clarkson & Robinson [18] applied a "mixture" model with adaptive weights, based on clusters formed from (mostly text) material from the BNC. They found that, particularly if the resulting cluster-based models were combined with the "full" trigram model trained on the complete set of training data, such models could give a reduction in perplexity of up to 24% over the baseline performance of the "full" trigram model alone. The approach followed in that work assumed that the current topic (used as the basis for clustering in that study) changes slowly as we pass through any document, so that it is possible to adapt the language model as we progress - say refining the model after each 10% of the text. Of course, only that part of the text which has already been seen can be used in the adaptation process.

Typical cluster-based adaptation processes involve initially clustering documents (or parts of documents) – in our case pairs of consecutive turns - reserved for training according to some criterion. A distinct language model is then built for each cluster. Interpolation weights for selecting an appropriate combination of cluster language models can then be computed using the Expectation-Maximisation (EM) algorithm [19,20] applied to previously-seen text.

In this study, we built 10 clusters of pairs of dialogue turns using a k-means algorithm with a metric based on a combination of frequencies of individual words within each pair of turns, the length of each pair and the number of distinct turn pairs in which any given word occurs [21]. We have carried out three experiments, using different strategies for determining the clusters, where several trigram language models are constructed for second turns of pairs, the model being selected according to which cluster the corresponding first turn of the pair is allocated. Then, in the "recognition" phase, one of these language models is chosen for the second turn of each pair according to the cluster most appropriate to the first turn of that pair, which is already known. Thus, we are trying to adapt the language model to be applied to the second turn of a pair according to knowledge about the corresponding first turn. Since the amount of data used to train each cluster model is relatively small, the chosen specific cluster language model was interpolated with the baseline "full" language model using the EM algorithm [19,20] applied to reserved data to obtain the weights. In addition, an "oracle"-type experiment, where the language model to be applied to the second turn of each pair was selected according to which cluster the *second* turn itself would

be allocated was carried out. This is cheating, since the choice of language model used to predict the content of the second turn then relies on information held in the same turn. In a real automatic speech recognition system, such as a dialogue system, this would not be possible – the system does not know the content of the user's current utterance in advance. However, an experiment of this type can provide an upper bound to the perplexity reduction which might be possible from the technique described above if we were able to predict the correct cluster for the second turn of a pair perfectly from the content of the corresponding first turn alone.

The various experiments, using different methods for developing the clusters, were as follows:

Experiment "F" - the clusters were built according to the content of both first and second turns of each turn pair in the training set. In testing, a model was chosen for the second turn of the pair according to which cluster the corresponding first turn would be assigned. This model was interpolated with the "full" language model for second turns. Once results had been compiled for each of the 10 clusters, an average perplexity, weighted according to the sizes of the clusters, was computed.

Experiment "T" - similar to experiment "F" above, but the clusters were built according to the content of the first turns of the pairs only.

Experiment "R" - similar to experiment "F" above, but only the content of second turns of the pairs was used in constructing the clusters. However, during testing, a language model was chosen for each second turn according to which cluster the corresponding first turn would be allocated, just as for experiments "F" and "T".

Experiment "O" - the "oracle" experiment, where only the content of second turns of pairs is used to build the clusters and, in testing, the language model for the second turn is chosen according to which cluster that turn would be assigned.

In each experiment, and for each cluster, interpolation of the appropriate cluster-based model with the "full" trigram model gave a modest reduction in perplexity over the baseline of the "full" trigram model alone. As expected, the "oracle" model, which relies on information within the second turn of the pair and is therefore invalid for predicting its content, gave the largest improvement over the baseline. This figure represents the greatest reduction which could be obtained using this type of approach. A summary of the weighted average perplexities calculated from each experiment are given in Table 3 below.

Experiment	Perplexity		Percentage Reduction in Perplexity over Baseline
	Baseline LM	Interpolated Baseline-Cluster LM	
"F"	218.54	212.99	2.54 %
"T"	218.54	212.92	2.57 %
"R"	218.70	213.22	2.51 %
"O" (Oracle)	214.62	198.88	7.33 %

Table 3: *Perplexities of baseline and interpolated baseline-cluster language models. See text for description of experiments. These results are of a single "rotation" alone*

The three ordinary approaches to clustering make rather different assumptions regarding the nature of the turns within each pair. The "F"-type approach assumes that the two turns of a pair are alike. Thus, in this case, it should not make much difference, with regard to assignment of pairs to clusters or choosing clusters based on information in the turns, whether clusters are selected on the basis of the first turn of a pair, the second turn of a pair, or both turns together. The "T"-type method assumes that the second turns of a pair are strongly dependent on the first. Therefore, selecting a second turn language model on the basis of which cluster the corresponding first turn belongs to is a logical step. By contrast, the "R"-type approach assumes that both turns of the pair are strongly dependent on each other, so that even though pairs have initially been assigned to clusters on the basis of their second turns, the content of the first turn of a pair should contain sufficient information to decide which cluster is appropriate for modelling the second turn of the pair. The results of these experiments suggest that there is little reason to prefer one method over the others.

8. Conclusions

Our investigations into the statistical properties of successive dialogue turns within the BNC have shown that improvements in language model perplexity can be obtained through the use of clustering, word trigger pairs or a cache. However, these studies are still at a relatively early stage and many possible parameters, such as the size of the cache or the number of triggers or clusters used, can be adjusted. Furthermore, combinations of these types of models may lead to additional, if modest, improvements. Triggers may also be useful in predicting the cluster most appropriate for the second turn of a pair. It is also intended that the linguistic significance of the trigger pairs used and of the turn clusters found be investigated – for example, whether there are obvious semantic or lexical links between words of a trigger pair or words in the same cluster.

Acknowledgements
Gordon Hunter has been supported by a research studentship from the UK Engineering and Physical Sciences Research Council.

References

Burnard, L., Users' Reference Guide for the British National Corpus, Oxford University Computing Services, Oxford, U.K., 1995.

Chelba, C., Jelinek, F., "Recognition Performance of a Structured Language Model", *Proceedings EuroSpeech-99*, 1999.

Cohen, P., "Dialogue Modeling", in R. Cole et al. (eds) *Survey of State of the Art in Human Language Technology*, Cambridge University Press, 1998.

Clarkson, P. & Rosenfeld, R., "Statistical Language Modeling Using the CMU-Cambridge Toolkit", *Proceedings of Eurospeech'97*, Vol. 5, pp2707-2710, 1997

Godfrey, J.J., Holliman, E.C. & McDaniel, J., "Switchboard: Telephone Speech Corpus for Research & Development", *Proceedings of ICASSP-92*, Vol.1, pp 517-520, 1992.

Iyer, R.M. & Ostendorf, M., "Modeling Long Distance Dependence in Language: Topic Mixtures Versus Dynamic Cache Models", *IEEE Transactions on Speech & Audio Processing*, Vol.7, pp 30-39, 1999.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., and van Ess-Dykema, C., "Switchboard Discourse Language Modeling Project – Final Report", John Hopkins University, 1998.

Karis, D., and Dobroth, K.M., "Automatic Services with Speech Recognition over the Public Switched Telephone Network: Human Factors Considerations", *IEEE Journal on Selected Areas in Communications*, Vol. 9, pp574-585, 1991.

Kuhn, R. & De Mori, R., "A Cache-Based Natural Language Model for Speech Recognition", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, Vol.12, pp 570-583 (Corrections in Vol. 14, pp 691-692), 1990.

Rosenfeld, R., "A Maximum Entropy Approach to Adaptive Statistical Language Modeling", *Computer Speech & Language*, Vol. 10 , pp 187-228, 1996.

Rosenfeld, R., "Two Decades of Statistical Language Modelling: Where do we go from here?" *Proceedings of the IEEE*, Vol. 88, No.8, pp1270-1278, 2000.

Gorin, A.L., Riccardi, G. & Wright, J.H. "How May I Help You ?", *Speech Communication*, Vol. 23, pp 113-127, 1997.

Allen, J.F., Byron, D.K., Dzikovska, M., Ferguson, G., Galescu, L. & Stent, A. "Towards Conversational Human-Computer Interaction", *AI Magazine*, Vol.22, No. 4, pp 27-37, 2001.

Gauvain, J-L & Lamel, L. "Large-Vocabulary Continuous Speech Recognition: Advances and Applications", Proceedings *of the IEEE*, Vol. 88, No. 8, pp 1181-1200, 2000.

Dahlback, N., Jonsson, A, & Ahrenberg, L. "Wizard of Oz Studies - Why and How", *Knowledge-Based Systems*, Vol.6, No. 4, pp 258-266, 1993

Pirker, H., Loderer, G., Trost, H. "Thus Spoke the User to the Wizard", *Proceedings of Eurospeech '99* (Budapest), Vol. 3, pp 1171-1174, 1999.

Jonsson, A. & Dahlback, N. "Distilling Dialogues - A Method Using Natural Dialogue Corpora for Dialogue Systems Development", *Proceedings of the 6th Applied Natural Language Processing Conference* (Seattle), pp 44-51, 2000.

Clarkson, P.R. & Robinson, A.J. "Language Model Adaptation Using Mixtures and an Exponentially-Decaying Cache", *Proceedings of ICASSP'97*, Vol.2, pp 799-802, 1997.

Dempster, A.P., Laird, N.M & Rubin, D.B. "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, Vol. 39, No. 1, pp 1-38, 1977.

Jelinek, F. "Self-Organised Language Modeling for Speech Recognition" in *Readings in Speech Recognition* (Ed. A. Waibel & Kai-Fu Lee), pp 450-506, Morgan Kaufman, 1990

Robertson, S.E. & Sparck-Jones, K. "Simple, Proven Approaches to Text Retrieval", Technical Report TR356, University of Cambridge Computer Laboratory (http://www.cl.cam.ac.uk/TechReports/TRIndex.html), 1997.

Darroch, J.N. & Ratcliff, D. "Generalised Iterative Scaling for Log-Linear Models", *Annals of Mathematical Statistics*, Vol. 43, pp 1470-1480, 1972

Katz, S.M. "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", *IEEE Transactions on Acoustics, Speech & Signal Processing*, Vol. 35, No. 3, pp 400-401, 1987.