ETH TREC-6:
Routing, Chinese, Cross-Language and Spoken Document Retrieval

Bojidar Mateev, Eugen Munteanu,
Páraic Sheridan, Martin Wechsler, Peter Schäuble

Swiss Federal Institute of Technology (ETH)
CH-8092 Zürich, Switzerland

Abstract

ETH Zurich's participation in TREC-6 consists of experiments in the main routing task, both manual and automatic runs in the Chinese retrieval track, cross-language retrieval in each of German, French and English as part of the new cross-language retrieval track, and experiments in speech recognition and retrieval under the new spoken document retrieval track. This year our routing experiments focused on the improvement of the feature selection strategy, on query expansion using similarity thesauri, on the grouping of features and on the combination of different retrieval methods. For Chinese retrieval we continued to rely on character bi-grams for indexing instead of attempting to segment and identify individual words, and we introduced a new manually-constructed stopword list consisting of almost 1,000 Chinese words. Experiments in cross-language retrieval focused heavily on our approach using multilingual similarity thesauri but also included several runs using machine translation technology. Finally, for the spoken document retrieval track our work included the development of a simple speaker-independent phoneme recogniser and some innovations in our probabilistic retrieval functions to compensate for speech recognition errors.

1 Introduction

The introduction of two new tracks to the TREC-6 evaluation have helped ETH Zurich greatly in evaluating the research in spoken document retrieval (e.g. [Wechsler and Schäuble, 1995]) and cross-language information retrieval (e.g. [Sheridan and Ballerini, 1996]) that we have been conducting for the past number of years. We have therefore participated fully in both of these tracks and, even based on the preliminary results released before the conference, have learned a great deal from the experience. Additionally to these two new tracks, TREC-6 sees our first official submissions to the Chinese retrieval track. These include both automatic and manual runs and build upon work we completed just after the deadline for submissions in last year's evaluation. We also continue to refine our approach to the routing task in the search for performance improvements. Our routing submissions therefore constitute the investigation of several sources of possible improvement over our TREC-5 submissions.

One aspect of our approach to the routing task that we specifically examined for improvements was the use of the U-measure for feature selection [Ballerini et al., 1996]. Further endeavours for the routing task centred on examining the grouping of semantically related features, the use of similarity thesauri and the combination of different retrieval methods. More details of our routing work and experiments are presented in section 2.

For our work on Chinese retrieval we already had in place an indexing module which indexed Chinese texts using character bi-grams. For our initial experiments last year we simply indexed everything using bi-grams and used our standard SPIDER retrieval functions, without any consideration of Chinese language characteristics. We have now included a manually constructed stopword list for Chinese, consisting of almost 1,000 words. Not only did we remove the stopwords from our index, we also used stopwords as obvious word boundaries, indexing only character strings between stopwords. Our Chinese indexing module additionally recognises English words in Chinese text and maintains these as complete units instead of breaking them into bi-grams. Details of our automatic and manual Chinese experiments are presented in section 3.

The cross-language retrieval track provided document collections in each of English, French and German and topic descriptions also in each language. Our work for this track covered all of these languages for the monolingual runs and also English-to-German, German-to-English, German-to-French, and French-to-German cross-language runs. We did not work with French-English cross-language combinations. Most of our cross-language work was based on further examination of our approach using similarity thesauri, especially with respect to the construction of similarity thesauri over bi-lingual corpora of varying quality. We did however perform additional runs using machine translation technology to automatically translation queries between English and German. More details are given in section 4.

We participated in the full Spoken Document Retrieval (SDR) track with our probabilistic weighting approach based on matching documents and queries at the phoneme level. We also built a simple speaker-independent phoneme recogniser using the HTK Toolkit [Young et al., 1993]. Much of our work for this track was centred on refining the probability estimation module used in our approach to retrieval from errorful documents (in this case, the output from speech recognition) and the weighting function used in ranking documents. This is described more fully in section 5.

2 Routing Experiments

In our TREC-5 paper we introduced a method for feature selection based on the U-measure [Ballerini et al., 1996]. The features with the highest value of μ_U were selected and included in the query profile for each topic. This year, for each query the 300 one-word features and the 300 phrases with the highest value of μ_U were selected. Additionally, if a query had a title part then the one-word features from the title are automatically selected and included at the top of the features selected by μ_U value. The selection of features with this method was then used as a source of information for grouping features.

Our aim in grouping features was to gather together *related* one-word features from a pre-selected set. Features can be considered to be related in many different ways for this task. Our approach is inspired by the idea of a *lexical space* as described in [Zavrel and Veenstra, 1995], though we aim to group together semantically related features instead of the syntactic classes described therein. The method depends on two sets of one-word features: a set of *focus* features and a set of *context* features. For our routing experiments we use query-dependent selections for the two features sets, making use of the feature selection method described above. Table 1 shows that using query-dependent feature sets in the creation of lexical spaces (feature groupings) indeed has the desired result of generating query-specific feature groupings. This example shows the feature groups generated around the feature *"dump"* which occurs in three queries.

Query Nr.	5	12	10001
Query	Dumping charges against Japan	Water Pollution	Soil Pollution
Dump...	antidump	discharg	dispos
	penalti	water	discharg
	punit	cleanup	mar
	impos	untreat	widespread
	price	pollut	pollut

Table 1: Features grouped with the feature "dump" for different queries.

Our official submissions for the TREC-6 evaluation used a combination of three different methods for routing retrieval, with the particular parameters of each method adjusted over the training documents from the AP, FBIS, WSJ and ZF32 collections. We used the well-known Lnu.ltn weighting scheme [Singhal et al., 1996] with both one-word and phrasal features selected according the method described above (not grouped). We also employed query expansion using a similarity thesaurus ([Qiu, 1995]) constructed over the FBIS collection. Our second retrieval method for the combination used, for each topic, groups of features for each of the top 28 one-word features. We computed a ranked document list for each of the 28 groups using the above Lnu.ltn method and then used these lists for each topic as further input to the combination of methods. The third method for the combination was a generalisation of the Binary Independence Retrieval (BIR) model in which we used linear combinations of feature co-occurrence matrices on a per query basis using logistic regression. This method was applied to re-rank initial ranked lists of 2,000 documents which were generated for each query using the Lnu.ltn method.

These three method were then combined on a per-query basis by summing the rank positions of documents in selected ranked lists from those returned by the three different methods (1 ranked list from Lnu.ltn, 28 lists from the feature groupings, and 1 from the BIR method). Selection of ranked lists for each query is based on the training data.

In summary, we use an improved feature selection strategy by adding query title words to features selected by the U-measure. We then group semantically related query features and, treating each group as a query, generate ranked document lists for each group. These ranked lists, and a further ranked list generated by applying the Lnu.ltn weighting method with query expansion, may then be combined with a ranked document list generated from a generalisation of the BIR model using feature co-occurrence matrices. The combination of ranked lists from different methods produces the final ranked list of documents for the query. This combination of ranked lists generated by different methods has been shown to lead to improvements in performance when compared to the performance of the individual methods.

3 Chinese Retrieval

Our work on Chinese text retrieval has been underway since the TREC-5 evaluation, though we did not achieve results in time for the TREC-5 submission deadline. We had simply expanded the SPIDER indexing module to handle Chinese characters and to index Chinese texts using character bi-grams. We then applied this indexing without any further attention to Chinese language considerations - with the sole exception of removing the Chinese two-byte blank-space character when we found that the same document was being retrieved in the top rank position for multiple queries, since matches were taking place on the blank spaces in queries and documents!

Our main extension to our TREC-5 Chinese work is the use of a new manually generated stopword list which consists of almost 1,000 entries. Example entries in the stoplist are illustrated in Figure 1. Apart from using the stoplist to directly eliminate stopword features from the index, we also took advantage of the stopwords to establish obvious word boundaries in the running Chinese text. Our hope is that the text strings between stopwords are in many cases equivalent to phrasal units. In any case, the reduction in number of indexing features achieved through our stoplist has been substantial. Our TREC-5 index of the Chinese document collection included 2,089,778 individual bi-grams, which has been reduced to 1,148,010 features in our current index, a reduction of 45%. This reduction in the index resulted in no loss of retrieval performance when measured in average precision over the TREC-5 topics.

的	何
很	或
及	即
既	仅
并且	并为
不必	不断
出于	出自
处于	此处
此刻	此致
从而	从来
从未	大半
大抵	这些

Figure 1: Example entries in our Chinese stoplist

In addition to the removal of stopwords, our Chinese indexing module now also recognises both English words and Chinese numeric strings within Chinese texts and treats these as units for indexing instead of breaking them into bi-grams. The recognition of English words was included in our official TREC-6 submissions whereas the recognition of numbers was developed after our submission. Our automatic Chinese submission used the Lnu.ltn retrieval method with the full topic descriptions as queries and an automatic pseudo relevance feedback loop using the top 10 0 ranked documents to expand the query by the top 50 features.

We also submitted a manual retrieval run in the Chinese retrieval track using two Chinese researchers to search using their own query formulations based on the topic descriptions and their own relevance judgements to be used for relevance feedback. Their task was to find and mark as many relevant documents as possible. The submitted results for each topic were then based on a final retrieval run using the manually identified relevant documents. In searching for relevant documents our first searcher generally read 15 of the retrieved documents, spending approximately 40-50 minutes on each topic. The second searcher generally read about 40 documents for each topic and spent about an hour on each topic. Our searchers identified an average of 11 relevant documents for each topic, though one searcher was noticeably more reluctant to judge documents as definitely relevant than the other (we used more "maybe" relevant documents from that searcher).

4 Cross-Language Information Retrieval

Our participation in the new cross-language retrieval track has covered each of the three individual languages, though not all of the potential language pairings for the cross-language experiments. The main thrust of our work was directed toward the evaluation of our approach to cross-language retrieval which uses multilingual similarity thesauri constructed automatically over comparable corpora. We were also interested however in investigating the potential role of machine translation technology for cross-language retrieval.

Each of our monolingual baseline runs in English, German and French essentially used the same approach. We have constructed stopword lists for each language and also employ stemming or word normalisation for each language. We use Porter's stemming algorithm for English [Porter, 1980], an equivalent rule-based stemmer for French, and a lexicon-based word normalisation module for German which includes compound word analysis. Our German module uses the CELEX lexicon [Baayen et al., 1993]. More details of our indexing for multiple languages can be found in [Wechsler et al., 1997]. Retrieval used the Lnu.ltn retrieval method with the full topic descriptions as queries and an automatic pseudo relevance feedback loop using the top 10 initially ranked documents to expand the query by the top 50 features. Our monolingual runs were denoted ETHee1, ETHff1, and ETHdd1.

4.1 German/French Retrieval

Our German-French cross-language experiments were based on similarity thesauri ([Schäuble, 1997](pp 29), [Sheridan and Ballerini, 1996]). The similarity thesauri were constructed based on the comparable documents of the SDA (Swiss news agency) collections in French and German used in this evaluation. The SDA collection consists of 141,046 French documents and 185,099 German documents. The French and German collections are composed independently and the differing numbers of documents in each language reflect the fact that the German collection reflects much more news local to northern Switzerland and southern German while the French document contain many stories local to southwest Switzerland and southeast France. Despite these differences however, there are many news items common to the two collections, especially relating to international events. It is this commonality which lends the collection its comparability, which we exploit for building similarity thesauri. Comparable documents were identified based on the document dates, co-occurring cognates (words which are identical in each language, especially places and names), and the manually assigned news classification codes assigned by the SDA. Document pairs which scored above a given threshold on matching these features were considered comparable and paired off. This resulted in a collection of 83,698 document pairs considered to be comparable. This resource was then made available to all participants of the cross-language retrieval track.

In constructing German-French similarity thesauri, the 83,698 document pairs were turned into 83,698 "surrogate" bilingual documents by simply concatenating each French and German document of each pair. The surrogate documents were then treated as a single collection and indexed in the SPIDER system, allowing us to construct German-French and French-German similarity thesauri as described in [Sheridan et al., 1997]. For cross-language retrieval in each direction the full topic description was expanded from the topic language to the target language terms most similar to the topic concept using the similarity thesaurus. The target language terms from the similarity thesaurus were then submitted as a query, followed by an automatic pseudo relevance feedback loop using the top 10 initially ranked documents to expand the query by the top 50 features. Our runs ETHfd1 and ETHdf1 use the 25 most similar terms from the similarity thesaurus as the query pseudo-translation and runs ETHfd2 and ETHdf2 take the 50 most similar terms.

4.2 English/German Retrieval

We decided that the Associated Press documents of the English collection were not so likely to be comparable to the SDA documents in German, which therefore ruled out a similar approach to building similarity thesauri as that used for the German-French experiments. We were however able to make use of the resource provided by the University of Maryland and the LOGOS Corporation, who translated the documents of the SDA German collection into English. This presented us with a bilingual parallel corpus on which to construct similarity thesauri. For our experiments however, we decided to take the stance which claims that full translations of all documents is not a practical approach in general for cross-language retrieval. We therefore did not rely on the complete document translations in any of our experiments. Instead, we compromised by extracting from each document and each translation the title and lead fields, representing a summary of abstract of each document. While assuming that full document translation for large collections is impractical, we verified that translation of titles and abstracts was a viable alternative by also automatically translating the titles and leads of the German NZZ collection using an off-the-shelf PC-based system called T1. Our German-English similarity thesauri for both German-to-English and English-to-German retrieval were based on the surrogate bilingual document collection constructed out of the German *document summaries* of the SDA and NZZ (except for the month of March which caused translation problems) collections together with their MT-translated equivalents (LOGOS for SDA and T1 for NZZ).

For our English-to-German retrieval experiments, each run included as a first step the expansion of the source-language topic using a retrieval and pseudo relevance feedback loop over the English document collection, using the top ten documents as relevant and expanding by the top 50 features. The expanded query was then submitted to whatever query translation method was being used in the particular run. This pre-translation query expansion was only used for English-to-German as a post-experiment analysis showed that the query expansion tended to broaden the query topic too much, whereas we expected that more specific query topics would do better through translation (especially since similarity thesauri introduce a further query expansion and broadening effect). We therefore expect our English-to-German runs which used the pre-translation expansion to underperform the similarity-configured equivalent runs in other language pairs which did not use pre-translation expansion.

Submitted runs denoted by ETHed2 and ETHed3 used the similarity thesaurus trained over the document summaries and translations. ETHed2 consisted of pre-translation pseudo relevance feedback query expansion to 50 features on the AP collection, followed by pseudo-translation to the 25 most similar German terms which were then submitted as a German query, followed by a pseudo relevance feedback loop in German, assuming the top 10 documents relevant and expanding by the top 50 features. The same process held for ETHed3, except the similarity thesaurus returned the most similar 50 German terms. One observation worth making here is that the cross-language runs based on similarity thesaurus constructed using MT produced parallel corpora are still susceptible to errors introduced during the MT process. This is illustrated by the fact that a lookup of the similarity thesaurus for English features similar to the German *Waldheim* will return *forest* as a similar features because *Waldheim* has been consistently translated as *"forest home"* in the underlying corpus. Building similarity thesauri over artificial corpora is therefore less reliable than using manually created corpora.

Our run ETHed1 was used to explore if machine translation of documents summaries was enough in itself to achieve acceptable cross-language performance. The English topic was expanded using pseudo relevance feedback over the AP collection as described above and the resulting query was submitted to the *English* translations of the German SDA and NZZ documents. The ranked list of returned translations (whose DOCNO id still identifies the original German document) was submitted as the result. On the other hand, run ETHed4 explored the usefulness of machine translation for straight query translation. Query topics were directly translated from English to German using the off-the-shelf T1 MT system and the German translations submitted, followed by a pseudo relevance feedback loop in German, assuming the top 10 documents relevant and expanding by the top 50 features. Note that run ETHed4 did *not* use pre-translation query expansion.

Our experiments in German-to-English retrieval were in many ways similar to the English-to-German runs described above. Since the English documents were not translated to German however, we could not attempt an equivalent of run ETHed1 which submitted the queries against translated versions of document summaries. We also dropped the pre-translation query expansion step having decided that it seemed likely to lead to a loss of effectiveness.

Runs denoted ETHde1 and ETHde2 used the similarity thesaurus trained over the document summaries and translations. ETHde1 consisted of pseudo-translation of the German topic to the 25 most similar English terms which were then submitted as a query, followed by a pseudo relevance feedback loop, assuming the top 10 documents relevant and expanding by the top 50 features. The same process held for ETHde2, except the similarity thesaurus returned the most similar 50 German terms. ETHde3 was the mirror of run ETHed4, directly translating the German topic to English using the T1 MT system and submitting the translation as query, followed by pseudo relevance feedback.

5 Spoken Document Retrieval

The Spoken Document Retrieval (SDR) track provided the perfect opportunity for evaluating our probabilistic weighting approach to indexing and retrieval of audio recordings [Wechsler and Schäuble, 1995]. Unfortunately our previous work had focused on German recordings, so the first step in our SDR participation was the development of an equivalent speech recognition system for English. We therefore built a speaker-independent phoneme recogniser for English speech using the HTK Toolkit [Young et al., 1993]. In comparison with the transcriptions provided as part of the track, our system achieved a phoneme recognition rate of 62.47%. For translating textual queries, reference transcriptions, and IBM's word-based recognition output to phonemic transcriptions we use an adapted version of the Carnegie Mellon Pronouncing Dictionary [CMU, 1995] with an additional rule-based translation program [Wasser, 1985] for translating out-of-vocabulary words.

Our retrieval method for spoken documents consists of four steps: slot detection, probability estimation, weighting and document ranking. The slot detection module locates possible occurrences of a query word in a spoken document. It also returns partial matches in order to cope with phoneme recognition errors. A similar method was employed in [Mittendorf et al., 1995] for retrieval of OCR-corrupted documents. For each identified slot we then determine the probability that the query word was actually spoken in that slot. The estimation employs a phoneme string similarity function enhanced with a phoneme confusion matrix which we derived from the SDR training set. Unlikely slot occurrences are then pruned using a threshold function. The aim is to filter out slots with lower probabilities as these are often "false alarms". The threshold is also sensitive to the length of query features, therefore favouring longer features which then to provide better clues of document relevance.

Estimated slot probabilities then contribute to the computation of an *expected feature frequency* for features in documents [Schäuble, 1997]. We adjust the computed expected feature frequencies however, compensating for systematic errors observed under training conditions. This is a similar compensation as was applied in the TREC-5 confusion track for OCR data [Ballerini et al., 1996]. Given values for expected feature frequencies and inverse expected collection frequencies, we then turned our attention to the retrieval function. In particular, we examined the issue of document length normalisation and developed a new document weighting scheme that uses what we call *logarithmic document length normalisation*. This scheme gave better results than the standard methods (cosine, pivoted document length normalisation etc.) when applied on the training collection. We also introduced a *logarithmic feature length normalisation* with the aim of adding weight to longer than average indexing features in a given query. This follows our belief that longer query features have more information value and, furthermore, are more likely to be detected in spoken documents. The logarithmic feature length normalisation alone was seen to accomplish an 18% improvement in average precision when performing known-item searches on the training collection with 13 queries (the SDR training queries plus some we devised ourselves). These compensation and normalisation methods were included in our supplementary submissions, ETHS2 based on our own speech recognition and ETHB2 based on IBM output.

For our obligatory runs ETHS1, ETHB1 and ETHR1 we added a compensation of documents weights to smooth the differences between the number of slots in a given document for a particular query feature compared to the average number of slots for that feature over the whole document collection. This compensation may be called *pivoted number-of-slots normalisation* since it resembles the pivoted document length normalisation of [Singhal et al., 1996]. A further new compensation measure addresses the variability in the ratio of Expected RSV to actual RSV for a given document depending on the number of query features in that document. This variability can lead to corruption of the ranked document list. We can prove theoretically the existence of such a variability and have confirmed it in our experimental experience. Using an analysis of the training data to tune the parameters of our compensation and normalise the variability in the ERSV/RSV ratio we have found that this correction delivers a 19% improvement in average precision on our training collection.

Note that in respect to the estimation of slot probabilities, our submitted runs on the reference data (perfect textual documents) and baseline data (word-based recognition) used slots which had been positively identified (probability 1) since in these cases we only want exact feature matching. We used the full probabilistic slot matching for the runs submitted based on our own phoneme-level speech recogniser (ETHS1, ETHS2).

Although we are aware that our own speech recognition system needs much further improvement, we consider the probabilistic retrieval approach described above as the major contribution of our work in this track. Initial investigations of the performance of this probabilistic approach compared to the SPIDER system used in the TREC-5 adhoc retrieval task show no significant difference in performance. We therefore aim to achieve comparable performance on spoken document retrieval tasks using current state-of-the-art speech recognition systems as we achieve on perfect text documents with the SPIDER system.

6 Conclusion

We have described here our participation in the TREC-6 evaluation with submission in the main routing task and in the tracks concerned with Chinese text retrieval, cross-language information retrieval, and spoken document retrieval. The main themes of this work can be summarised as:

Routing - We have improved our U-measure for feature selection by including features from query titles. We group semantically related features on a per query basis and generate queries of multiple semantic groups for each topic. The results of the feature group retrieval are combined, also on a per query basis, with results of retrieval using the Lnu.ltn retrieval method and a retrieval method using feature co-occurrence matrices. The combination of ranked lists generated by different methods has been shown to lead to improvements in performance when compared to the performance of the individual methods.

Chinese - We have built upon our earlier efforts by including a new manually composed stoplist which has helped to reduce our index space by 45% without adversely affecting in retrieval performance. Chinese text is indexed using character-bigrams, although English words and numeric strings are identified and indexed whole.

Cross-Language - The SPIDER system indexing engine includes stoplists and stemming or word normalisation modules for each of the languages covered. We have further evaluated our similarity thesaurus approach for cross-language retrieval, constructing similarity thesauri over comparable corpora when available and investigating the use of parallel collections created by applying machine translation to documents or document abstracts when no comparable corpus is available. We have also tested the effectiveness of retrieval directly using machine translation output, both with document abstract translations or query translations.

Speech Retrieval - Apart from developing our own speech recognition system for English, we have made major innovations in our probabilistic approach to retrieval from errorful information. These include new document length and feature length normalisations in the weighting stage and a new compensation measure in the retrieval method which normalises the variability in Expected RSV compared to the actual RSV which arises across documents which contain different numbers of query features.

Our work on each of these themes is ongoing and we hope that several aspects of the work presented here can be further developed and reported much more formally in the information retrieval literature.

References

[CMU, 1995] (1995). *Carnegie Mellon Pronouncing Dictionary (cmudict.0.4, 1995).* http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[Baayen et al., 1993] Baayen, R., Piepenbrock, R., and van Rijn, H. (1993). The CELEX Lexical Database. Technical report, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

[Ballerini et al., 1996] Ballerini, J., Büchel, M., Domenig, R., Knaus, D., Mateev, B., Mittendorf, E., Schäuble, P., Sheridan, P., and Wechsler, M. (1996). SPIDER Retrieval System at TREC5. In *TREC-5 Proceedings*.

[Mittendorf et al., 1995] Mittendorf, E., Schäuble, P., and Sheridan, P. (1995). Applying Probabilistic Term Weighting to OCR Text in the case of a Large Alphabetic Library Catalogue. In *Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA*, pages 328–335.

[Porter, 1980] Porter, M. F. (1980). An Algorithm for Suffix Stripping. *Program*, 14(3):130–137.

[Qiu, 1995] Qiu, Y. (1995). *Automatic Query Expansion Based on a Similarity Thesaurus*. PhD thesis, Swiss Federal Institute of Technology.

[Schäuble, 1997] Schäuble, P. (1997). *Multimedia Information Retrieval—Content-Based Information Retrieval from Large Text and Audio Databases*. Kluwer Academic Publishers, Boston/London/Dordrecht.

[Sheridan and Ballerini, 1996] Sheridan, P. and Ballerini, J. P. (1996). Experiments in Multilingual Information Retrieval using the SPIDER System. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland*, pages 58–65.

[Sheridan et al., 1997] Sheridan, P., Braschler, M., and Schäuble, P. (1997). Cross-Language Information Retrieval in a Multilingual Legal Domain. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, Pisa, Italy*, pages 253–268.

[Singhal et al., 1996] Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted Document Length Normalization . In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland*, pages 21–29.

[Wasser, 1985] Wasser, J. A. (1985). *English to Phoneme Translation.* Program in public domain, ftp://ftp.doc.ic.ac.uk/packages/unix-c/utils/phoneme.c.gz.

[Wechsler and Schäuble, 1995] Wechsler, M. and Schäuble, P. (1995). Speech retrieval based on automatic indexing. In Ruthven, I., editor, *Proceedings of the Final Workshop on Multimedia Information Retrieval (MIRO'95)*, Electronic Workshops in Computing, Glasgow. Springer.

[Wechsler et al., 1997] Wechsler, M., Sheridan, P., and Schäuble, P. (1997). Multi-language text indexing for internet retrieval. In *Proceedings of the 5th RIAO Conference, Computer-Assisted Information Searching on the Internet, Montreal, Canada*, pages 217–232.

[Young et al., 1993] Young, S., Woodland, P., and Byrne, W. (1993). *HTK Version 1.5: User, Reference & Programmer Manual.* Entropic Cambridge Research Laboratory, Sheraton House, Castle Park, Cambridge CB3 OAX, England.

[Zavrel and Veenstra, 1995] Zavrel, J. and Veenstra, J. (1995). The Language Environment and Syntactic Word-Class Acquisition. In *Proceedings of the Groningen Assembly on Language Acquisition (GALA95)*, pages 365–374.