The impact on retrieval effectiveness of skewed frequency distributions

Mark Sanderson[1] and Keith van Rijsbergen
Department of Computing Science
University of Glasgow, Glasgow G12 8QQ, UK
(sanderson|keith)@dcs.gla.ac.uk

Abstract

We present an analysis of word senses that provides a fresh insight into the impact of word ambiguity on retrieval effectiveness with potential broader implications for other processes of information retrieval. Using a methodology of forming artificially ambiguous, words known as pseudo-words, and through reference to other researchers' work, the analysis illustrates that the distribution of the frequency of occurrence of the senses of a word plays a strong role in ambiguity's impact on effectiveness. Further investigation shows that this analysis may also be applicable to other processes of retrieval, such as Cross Language Information Retrieval, query expansion, retrieval of OCR'ed texts, and stemming. The analysis appears to provide a means of explaining, at least in part, reasons for the processes' impact (or lack of it) on effectiveness.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing - Linguistic processing; I.2.7 [Artificial Intelligence]: Natural Language Processing - Text analysis; I.6.4 [Simulation and Modeling]: Model Validation and Analysis.

General Terms: Experimentation, Measurement

Additional Key Words and Phrases: word sense ambiguity, word sense disambiguation, pseudo-words.

1 Introduction

A great many words in natural languages are ambiguous. The resolution of ambiguity is a task that has concerned a great many researchers in the field of Computational Linguistics. Over the years, many programs have been built that, given a word appearing in a certain context (with a definition of the word's possible senses), attempt to identify the word sense in the context. Such systems are known as *word sense disambiguators*. Early disambiguators were based on hand-built rule sets and only worked over a small number of words and senses *[Weiss 73]*, *[Small 82]*. This changed, however, with the availability of dictionaries and thesauri online. Using these reference works as a source of word sense definition and information, many disambiguation systems were built with the hope that they could be scaled up to work over a much wider vocabulary*[Lesk 86]*, *[Wilks 90]*, *[Sussna 93]*.

Such a possibility was of interest to researchers in the field of text based *Information Retrieval* (IR) systems where it was thought that word sense ambiguity was a cause of poor performance in the sys-

1. Now at the Department of Information Studies, University of Sheffield, Sheffield, UK.

tems. It was believed that if the words in a document collection were correctly disambiguated, IR effectiveness would improve. However, work where such disambiguation was performed failed to show any improvement *[Voorhees 93], [Wallis 93]*. From these results it becomes clear that research was needed to investigate the relationship between sense ambiguity, disambiguation, and IR. Investigations with similar aims but using different methods were conducted by Krovetz & Croft *[Krovetz 92]* and by one of the authors *[Sanderson 94]*.

1.1 Krovetz & Croft

As part of a wide-ranging paper on disambiguation and IR, Krovetz and Croft conducted a large-scale study on the relationship of relevance to sense matches and mismatches between query words and document words. Using the CACM and TIME test collections (*[Salton 83], [Sparck Jones 76]*), Krovetz & Croft performed a retrieval for each of the queries in these collections. For each retrieval, they examined the match between the intended sense of each query word and that word's sense as it occurred in the ten highest ranked documents. They counted the number of sense mismatches between query and document words and examined this figure in relation to the relevance or non-relevance of the document. They found that when the document was not relevant to the query, a sense mismatch was more likely to occur. From this analysis, it could be inferred that the level of sense match in the top ranked relevant documents was high. Krovetz and Croft speculated that this was due to the so-called *query word collocation effect*, which can be explained through an example.

If one were to enter the single word query ‘bank’ into an IR system, it is just as likely to retrieve economic documents as it is geographic ones. If, however, one entered a query containing many words, for example ‘bank economic financial monetary fiscal’ then, for top ranked documents, it is likely that many of the query words will collocate in those documents. It would be unlikely that an occurrence of ‘bank’ in such a document would refer to the margin of a river. Therefore, collocation can cause disambiguation to be unnecessary. Krovetz and Croft also described a second cause of the high degree of sense match in the top ranked documents, which is explained in Section 4.1.

Krovetz and Croft's study did not predict any significant improvements in retrieval effectiveness from the resolution of ambiguity in document collections. Instead, they described a number of situations where disambiguation may prove useful: where the effects of collocation were less prevalent such as high recall searches; and where query words were used in a less frequent sense.

1.2 Sanderson

Sanderson measured the effectiveness of an IR system retrieving from the Reuters 22,173 test collection *[Lewis 91]* and then measured it again after additional ambiguity was introduced into the collection using artificial ambiguous words known as *pseudo-words*[2]. The drop in effectiveness resulting from their introduction was a measure of the *impact* of that ambiguity. The results of the experiments showed that the introduced ambiguity did not reduce effectiveness as much as might have been expected. The published analysis of the results *[Sanderson 94]* concentrated on the length of queries showing that the effectiveness of retrievals based on a query of one or two words was greatly affected by the introduction

2. Simulated words that have multiple senses. The manner of their creation is explained in Section 2.

of ambiguity but much less so for longer queries. A confirmation of the co-location effect shown by Krovetz and Croft.

Although it was not stated in the paper, query term co-location within a document is also dependent on its length. If documents being retrieved were particularly short (e.g. just document titles) then co-location of query terms, regardless of query size, is likely to be low. Therefore, in a situation of retrieving from short documents, one would expect to see the same impact of ambiguity on retrieval effectiveness as was observed with short queries.

Sanderson also used pseudo-words to study the impact of automatic disambiguation on retrieval effectiveness, concentrating particularly on the impact of disambiguation errors. The number of mistakes made by disambiguators appears to vary depending on the subtlety of word sense to be discriminated between. A reasonable sense selection error for a disambiguator performing the discrimination task Sanderson assumed is around 20%-30% (taken from *[Ng 96]*[3]). He found that this level of error could cause effectiveness to be as bad or even lower than when ambiguity was left unresolved. Disambiguation error was thought to be a likely cause of the failures reported by Voorhees and by Wallis. To this end, Sanderson concluded that a disambiguator was only of use in a retrieval context if it disambiguated at a very high level of accuracy or if queries (or documents) were very short.

3. Ng & Lee stated that the error rate of their disambiguator was 30%, however, it was trained and tested on manually disambiguated corpora which themselves contained errors. This will have impacted on the disambiguator accuracy, therefore, a more generous error rate is suggested in this paper.

With reference to Sanderson's work, Schütze & Pedersen *[Schütze 95]* questioned the methodology of using pseudo-words to study ambiguity after examining the *pseudo-senses* that make up a pseudo-word. Looking at the distribution of these senses' frequency of occurrence within a collection, they found that one pseudo-sense of a pseudo-word typically accounts for the majority of occurrences of that pseudo-word. They suggested, though did not test, that this *skewed frequency distribution* of the pseudo-senses within a pseudo-word was an additional cause of the low impact of ambiguity on retrieval effectiveness. They further questioned whether this type of distribution correctly reflected that found in the senses of real ambiguous words. It is the two issues they raised that are addressed in this paper.

First a description of pseudo-words is presented followed by examples of their use in retrieval experiments. An analysis of the skewed distribution of the frequencies of occurrence of a pseudo-word's pseudo-senses is described next, along with an explanation of how this type of distribution impacts on retrieval effectiveness. Measurements are presented confirming that the senses of actual ambiguous words have the same skewed distribution as pseudo-senses and, therefore, pseudo-words are concluded to model well this aspect of ambiguous words. Experimental results showing the impact of pseudo-words on retrieval effectiveness are thus used to describe the impact of actual ambiguous words on effectiveness. Before concluding, the paper briefly examines additional applications of the analysis presented which indicate that other processes of retrieval may be better understood with such an analysis.

2 A methodology using pseudo-words

The aim of Sanderson's experiments was to gain a greater understanding of the relationship between word sense ambiguity, disambiguation accuracy, and IR effectiveness. In order to achieve this, it was necessary to be able to measure the amount of ambiguity in a test collection. This was achieved by using a technique of adding into the collection artificial ambiguous words called pseudo-words[4].

A pseudo-word is formed by concatenating a number of words chosen randomly[5] from a text collection. These words become the pseudo-senses of a newly formed pseudo-word and all of their occurrences within that collection are replaced by it: for example, randomly choosing the words 'banana', 'kalashnikov' & 'anecdote', and replacing all their occurrences in a collection by the pseudo-word 'banana/kalashnikov/anecdote'. Note that pseudo-words are mutually exclusive: a word can only be a member of one pseudo-word.

By adding pseudo-words into a test collection in this manner, a measurable amount of additional ambiguity is introduced into that collection and its impact on retrieval effectiveness can be determined. The size of pseudo-words can be varied by altering their number of pseudo-senses. A pseudo-word with *n* senses is referred to here as a *size n pseudo-word*.

2.1 What is meant by 'ambiguity'?

One aspect of ambiguity that was not addressed by Sanderson in his paper, was the type of ambiguity simulated by pseudo-words. This issue is described by Kilgarriff *[Kilgarriff 97]* who contended that the distinction between senses of a word could only be defined once the purpose for which they were intended was defined. Assuming that a dictionary provides a definitive and objective distinction between word senses is, according to Kilgarriff, unrealistic (Section 5 discusses this issue further in the context of IR). In Sanderson's work, pseudo-words were intended to mimic the senses used in the work of Voorhees who used the *WordNet* thesaurus *[WordNet]*, *[Miller 95]* for her sense definitions. Indeed, it will be shown later that an important quality of senses in this reference work, are simulated well by pseudo-words. Unless otherwise stated, references to senses and ambiguity in this paper should be taken as meaning senses as defined in WordNet. It is believed, however, that pseudo-words are a good simulation of senses defined in other reference works such as dictionaries and some evidence is presented to support this contention.

4. Pseudo-words were created by two groups in the same year working independently of each other: Gale et al. and Schütze. Gale, Church and Yarowksy introduced and tested a disambiguator using pseudo-words in a 1992 paper *[Gale 92c]*. (In the following year, Yarowsky *[Yarowsky 93]* incorrectly cited *[Gale 92a]* as being the original pseudo-word paper.) At the same time, Schütze introduced a means of testing a disambiguator using manually created ambiguous words *[Schütze 92]*, though he did not call them pseudo-words. Note, that inspired by Schütze, Grefenstette introduced the notion of artificial synonyms *[Grefenstette 94]*.

5. A pseudo random process was used based on a non-linear additive feedback random number generator: the *random* and *srandom* functions found in the *math.h* library of the *C* programming language.

2.2 The experiments

To illustrate the impact of the introduction of pseudo-words into a document collection, experiments on three conventional test collections, CACM, Cranfield 1400 *[Sparck Jones 76]*, and TREC-B, are now presented. In these experiments, size five pseudo-words were introduced into each collection and the effectiveness of an IR system retrieving from these additionally ambiguous collections was measured. All words in the collections and their respective queries were transformed into pseudo-words[6].

The CACM collection was composed of 3,204 documents and the Cranfield 1400 collection contained 1,400. The TREC-B collection used was that defined in the TREC-5 conference *[Harman 96]*, it contained ~70,000 documents; the queries used (known as topics in TREC) were numbers 1 to 300.

The retrieval system used was a conventional ranking system using a form of *tf*idf* term weighting scheme (1) which is an amalgam of

$$w_{ij} = \frac{\log(freq_{ij} + 1)}{\log(length_j)} \bullet \log\left(\frac{N}{n_i}\right) \qquad (1)$$

w_{ij} = tf•idf weight of term i in document j
$freq_{ij}$ = frequency of term i in document j
$length_j$ = number of terms in document j
N = number of documents in collection
n_i = number of documents in which term i occurs

Harman's *normalised within document frequency weight [Harman 92]* and a conventional *inverse document frequency* measure. Stop words were removed from the collection and the Porter stemmer *[Porter 80]* was applied to the collection text before pseudo-words were generated. A pessimistic interpolation technique (described in the Interpolation section of Chapter 7 of Van Rijsbergen's book *[Van Rijsbergen 79]*) was used to produce a set of precision values measured at ten standard recall levels.

As can be seen from the results in Figures 1, 2, & 3, the effectiveness resulting from this retrieval is little different from that resulting from a retrieval on the unmodified collection. Considering that

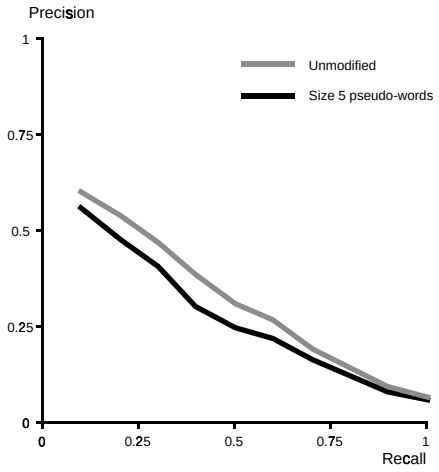

Figure 1. Introducing size five pseudo-words into the CACM collection.

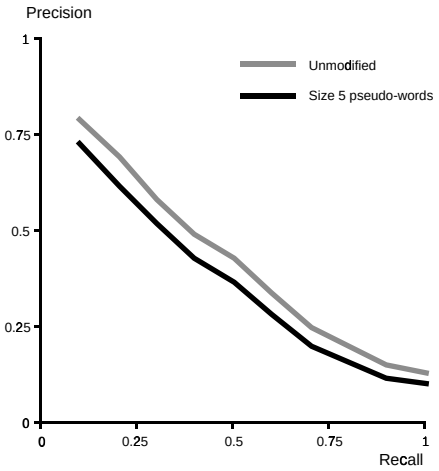

Figure 2. Introducing size five pseudo-words into the Cranfield 1400 collection.

6. It was possible that up to four words in each collection were left out of the transformation process due to the requirement that each pseudo-word had five senses.

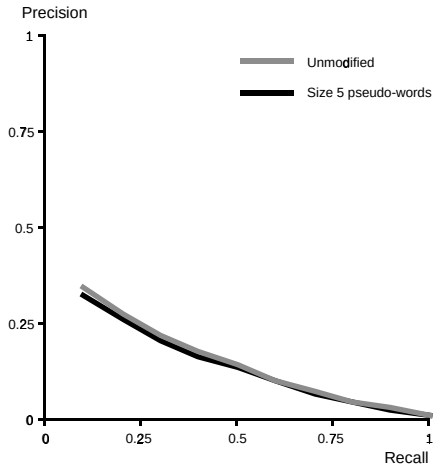

Figure 3. Introducing size five pseudo-words into the TREC-B collection.

the introduction of size five pseudo-words reduced the number of distinct terms in the collections to a fifth, the relatively small decrease in retrieval effectiveness is perhaps striking. The differences in reductions across the collections is most likely due to the differences in query length. TREC-B queries are, on average, 41 non-stop words in length as opposed to 12 for the CACM collection and 10 for the Cranfield 1400.

2.3 Summary

This section has presented a methodology that uses pseudo-words to explore the relationship between ambiguity and retrieval effectiveness. Experimental results showed the impact of introduced word sense ambiguity on retrieval effectiveness was not as significant as might have been thought. In addition, the results showed a link between the length of query submitted to a retrieval system and the impact of that ambiguity on effectiveness.

2.4 Postscript

Since conducting this work, it has come to light that experiments with a similar methodology but with a different purpose were carried out by Burnett et al. *[Burnett 79]* who were performing experiments using document signatures. They were investigating how best to generate small but representative signatures from a document. One of their experiments involved randomly pairing together words in the same way that size two pseudo-words are formed. They noted that retrieval effectiveness was not affected greatly by this pairing, a result that is in agreement with those presented here.

3 Analysis of frequency distribution

Although the experiments of Section 2.2, and those presented in *[Sanderson 94]*, showed query (and by implication document) length to be an important factor in the relationship between ambiguity and retrieval effectiveness, further analysis, by Schütze & Pedersen *[Schütze 95]* revealed that the skewed frequency distribution of pseudo-senses could also be causing the relatively small drops in effectiveness observed in those experiments. In this section, this factor is analysed and experiments are conducted to reveal its impact on retrieval effectiveness.

3.1 Examining the make up of pseudo-words

Words have very different frequencies of occurrence within a document collection, as shown by Zipf *[Zipf 49]*. This can be demonstrated by examining the text of the CACM collection which contains approximately 7,500 distinct words occurring 100,000 times. Figure 4 shows the distribution of the frequency of occurrence of this set of words. The graph shows their distribution is skewed. Such a distribution is often referred to as a *Zipfian distribution*. Therefore, creating pseudo-words by random selection from these words is likely to result in pseudo-words composed of pseudo-senses with a similar (Zipfian) skew. This becomes apparent after examining the frequency of occurrence of the senses of

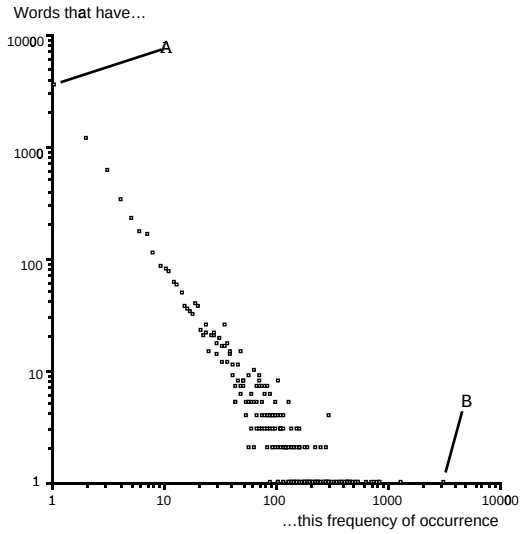

Figure 4. Distribution of the frequency of occurrence of words in the CACM collection. Graph plotted on a logarithmic scale. Point A shows that around 3,600 of the words (about half of all words in the collection) occur in the collection only once. Point B shows that one word occurs around 3,000 times in the collection, accounting for 3% of all occurrences in the collection.

four randomly selected pseudo-words generated from the CACM collection:

- the senses of the size five pseudo-word '12/span/prospect/preoccupi/nonprogram'[7] occurred 218, 18, 3, 2, and 1 times in the CACM collection respectively;
- the senses of 'assist/prohibit/minicomput/ness/inferior' occurred 27, 5, 5, 2, and 1 times;
- the senses of 'taken/multic/purdu/beginn/pavlidi' occurred 28, 8, 4, 2, and 1 times;
- and the senses of 'note/makinson/disappear/gilchrist/xrm' occurred 97, 3, 2, 2, and 1 times.

7. The unusual ending of some of the words is due to the application of a stemmer *[Porter 80]* to the words of the CACM collection before formation of pseudo-words.

The extent to which this skewed distribution existed in pseudo-words was more fully investigated: sets of size two, three, four, five, and ten pseudo-words were created from the words of the CACM collection, and the distribution of the frequency of occurrence of their pseudo-senses, was examined. For each of these pseudo-words, it was found that one sense accounted for the majority of occurrences of the pseudo-word of which it was a part. The results of this analysis are shown in Table 1 which displays the percentage of occurrences accounted for by a pseudo-word's *commonest sense*. From these figures, it was concluded that the distribution of the frequency of occurrence of the pseudo-senses was skewed.

No. of senses	Commonest sense (%)	
2	**92**	{50}
3	**85**	{33}
4	**80**	{25}
5	**78**	{20}
10	**65**	{10}

Table 1. Percentage of occurrences accounted for by commonest pseudo-sense of a pseudo-word (computed by micro averaging). The figures in brackets (shown for comparison) are the percentages that would result if pseudo-senses occurred in equal amounts. Measurements made on the CACM collection.

3.2 Why do skewed pseudo-words not affect effectiveness?

An examination was undertaken to discover if the skewed frequency distribution of pseudo-senses was in part responsible for the retrieval results presented in Section 2.2. Initially, the frequency of occurrence of test collection query words was examined. It was found that the majority of these words had a relatively high frequency of occurrence in their respective collections. This was significant as, if a high frequency query word was made part of a

pseudo-word, there was a high probability that the other pseudo-senses of that pseudo-word would have low frequencies of occurrence (because of skewed frequency distributions). Therefore, the pseudo-word's commonest sense would account for the majority of its occurrences and would, in effect, be little different from the high frequency query word that was its main component. Consequently, there would be little change in the retrieval effectiveness of an IR system retrieving on that word.

To illustrate, query fourteen of the CACM collection contains the word 'algorithm', which occurs in 1,333 documents. After pseudo-words were introduced into the collection, this query word became the commonest pseudo-sense of the word 'algorithm/telescop/pomental/lanzano/mccalla', which occurred in 1,343 documents, only ten more than the original query word. Therefore, turning this relatively high frequency query word into a pseudo-word had little impact on the word's frequency of occurrence and therefore, little impact on its use in retrieval.

It was hypothesised that if the majority of query words were, like 'algorithm', the commonest sense of a pseudo-word, this would help to explain the relatively small drop in retrieval effectiveness resulting from the introduction of such words. To test this hypothesis, size five pseudo-words were introduced into the CACM, Cranfield 1400, and TREC-B collections and the number of query words that were the commonest sense of a pseudo-word was counted. As can be seen from the results in Table 2, the majority of query words had this 'commonest sense' property. These results, certainly suggested that the skewed frequency distribution of a pseudo-word's pseudo-senses was an additional cause of the relatively small drop in retrieval effectiveness found in the experiments presented in Section 2.2.

Number of	CACM	Cranfield	TREC-B
queries	52	225	285
query words	645	2159	11848
query words in collection	618	2151	11781
commonest sense query words	474	1827	11048
commonest sense(%)	**77**	**85**	**94**

Table 2. Percentage of query words that were the commonest sense of a pseudo-word.

To further confirm this, the experiments of Section 2.2 were repeated exactly as before, except that the pseudo-words introduced into the collection were of a different type: their pseudo-senses had an equal frequency of occurrence[8]. The graphs in Figures 5, 6, & 7 show the difference in retrieval effectiveness

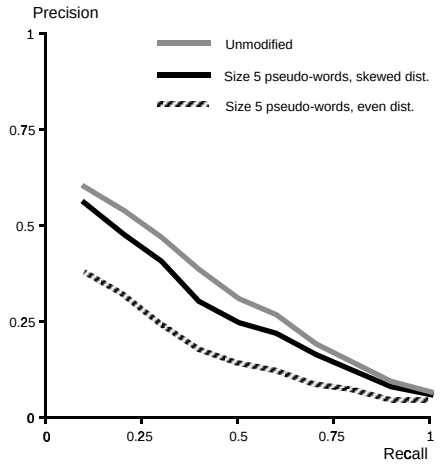

Figure 5. Comparison of pseudo-word types on the CACM collection.

8. This type of pseudo-word was formed for a particular document collection, by sorting that collection's word index by (word) frequency of occurrence and then grouping contiguous sets of sorted words into pseudo-words. This means of grouping ensured that words with equal or almost equal frequency of occurrence were joined together.

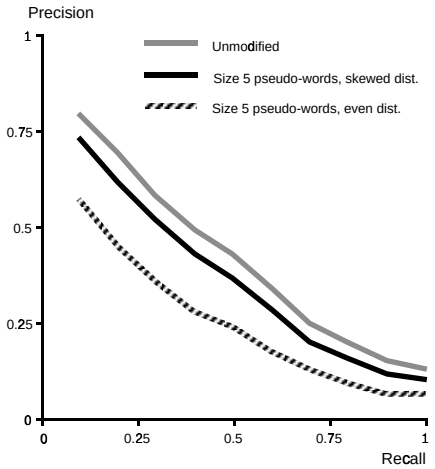

Figure 6. Comparison of pseudo-word types on the Cranfield 1400 collection.

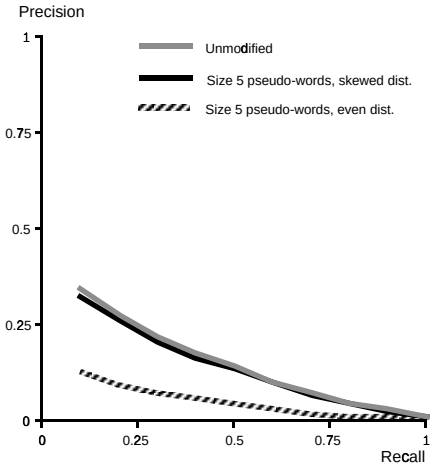

Figure 7. Comparison of pseudo-word types on the TREC-B collection.

when size five pseudo-words, whose senses have even distributions, are introduced into a collection. As can be seen, across all three collections, the impact on effectiveness of introducing pseudo-words with even distributions is significantly greater than the introduction of pseudo-words with a skewed distribution.

From these results and the analysis of query words, it was concluded that the relatively low impact of ambiguity reported in the experiments of Section 2 was not only due to the relatively long queries of the collections but also due to the skewed frequency distribution of the pseudo-senses used in the experiments.

4 Do pseudo-words model ambiguous words well?

Given the discussion so far, it would not be unreasonable to wonder how well pseudo-words model ambiguous words. In other words, do the senses of ambiguous words have the same skewed distribution as pseudo-senses? There is a well known result from a number of disambiguation researchers that suggests this is the case. In their research on establishing a lower bound baseline for measuring the significance of a disambiguator's accuracy, Gale et al. *[Gale 92a]* found that if a disambiguator used a strategy of selecting the commonest sense of a word, it would be correct 75% of the time. More recently, Ng & Lee *[Ng 96]* reported on the creation of a sense tagged corpus containing 191,000 word occurrences. They found the commonest sense of the words they tagged (which had on average nine senses per word) accounted for 64% of all occurrences.

It is possible to measure the frequency distribution of word senses using the SEMCOR sense tagged corpus which is publicly released with WordNet *[WordNet]*, *[Miller 95]*. It is a 100,000 word corpus consisting of around 15,000 distinct words. All word occurrences were manually tagged with senses as defined in the Wordnet thesaurus (version 1.4). Using this corpus, the distribution of the frequency of occurrence of ambiguous word senses can be plotted (Figure 8). Examining the graph in Figure 8 reveals that the *senses* in the SEMCOR corpus have a skewed frequency distribution similar to that of the *words* in the CACM collection as shown in Figure 4.

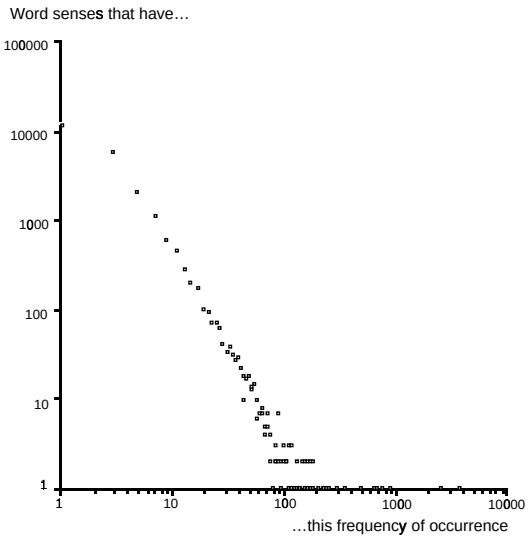

Figure 8. Distribution of the frequency of occurrence of senses in the SEMCOR corpus. Graph plotted on a logarithmic scale.

As was done with pseudo-words, the distribution of the frequency of occurrence of word senses within ambiguous words was examined, Table 3 displays the percentage of occurrences accounted for by a word's commonest sense. The percentage was computed for separate sets of words, the set a word belongs to is defined by the number of senses that word has. As can be seen, a word's commonest sense accounts for the majority of that word's occurrences, thus confirming the results of Gale et al. and of Ng & Lee.

No. of senses	Size of set	Commonest sense (%)	
2	3145	**92**	{50}
3	1697	**85**	{33}
4	1046	**79**	{25}
5	640	**72**	{20}
6	448	**68**	{17}
7	275	**63**	{14}
8	200	**60**	{13}
9	141	**60**	{11}
10	93	**53**	{10}

Table 3. Percentage of occurrences accounted for by the commonest sense of a word (computed by micro averaging). The figures in brackets (shown for comparison) are the percentages that would result if senses occurred in equal amounts. Measurements made on the SEMCOR corpus.

Tables 1 & 3 show a similarity. It was unexpected that the frequency of occurrence of an ambiguous word's senses would be modelled so well by the random word selection process used to form pseudo-words. Although both senses and words were found to have a similar skew within a collection, one might have anticipated that the distribution of the senses of an ambiguous word would be affected by other factors that would not be captured by the pseudo-word creation process. This, however, did not seem to be the case and therefore, it was concluded that the skewed distribution of a pseudo-word's pseudo-senses is a good model of this aspect of an ambiguous word[9].

4.1 Does 'real' ambiguity impact on retrieval effectiveness?

Given that the senses of ambiguous words have the same skewed distribution as the pseudo-senses of pseudo-words, and that such a distribution is an identified cause of the small drop in retrieval

9. The strength of similarity shown in Tables 1 & 3 may be co-incidental, only more extensive tests on manually tagged corpora (when available) will be able to confirm or deny this. Further, it should be noted that a typical feature of most manually disambiguated corpora, currently available, is the lack of multiple manual assessments of their words. The observations of skew described and referenced in this paper are largely based on corpora having this property. It is well documented that levels of inter assessor agreement in sense tagging tasks can be low *[Gale 92a]*. Therefore, it is possible that the skewed distribution of senses will have different properties on multiply assessed corpora. Once such resources are available, it may be prudent to re-examine this issue.

effectiveness found in the experiments of Section 2.2, it might seem reasonable to question if this small drop will also be found in the case of real ambiguity. As was shown in the analysis of pseudo-senses in Section 3.2, when the senses of a large percentage of ambiguous query words have a skewed distribution and are used in the commonest sense present in the collection (being retrieved from), ambiguity will not greatly impact on effectiveness.

A short study of the senses of the query words of one hundred TREC queries (No. 201-300) was undertaken to determine how many words were used in their commonest sense. The short title section of these queries were manually disambiguated (by one person) with respect to the sense definitions of WordNet. This thesaurus was chosen as it contains relative frequency of occurrence information for senses, which allow one to determine the commonest sense of any word in WordNet. Examining the nouns of these queries, it was found that 158 out of 207 (76%) were used in their commonest sense. (There were 11 nouns not found in WordNet or whose intended sense was not defined.) For these queries at least, the commonest sense predominates.

There is one caveat to the use of WordNet described here: the measurement of the commonest sense is calculated from corpora other than the collections the TREC queries were intended for. Therefore, it is possible that a word's commonest sense in the TREC collections is somewhat different from that defined in WordNet. In order to allow for this, it would be necessary to analyse the frequency distributions of senses in both query and collection. Such a study was performed in Krovetz & Croft's wide ranging analysis of ambiguity and retrieval effectiveness *[Krovetz 92]*. They examined the patterns of usage of the senses of query words in the CACM collection. Comparing their usage in the collection as a whole to that found in the queries, they found a similar pattern between the two: i.e. the commonest sense of a query word as used in a collection was the commonest sense of that word in a query. Krovetz & Croft used this study to conclude that the skewed distributions of ambiguous words were an important factor in the overall low impact of ambiguity on retrieval effectiveness.

From their results and those of the short study described above, it was concluded that pseudo-words accurately simulate the skewed distribution of ambiguous words and the results drawn from pseudo-word based retrieval experiments provide a good indication of what will happen in the 'real case'.

4.2 Other aspects of ambiguity and pseudo-words

Before drawing the discussion on pseudo-word simulation to a close, it is necessary to address one other aspect of senses and its simulation by pseudo-words, namely the relatedness of a word's senses. Although there are some words, known as *homographs* or *homonyms*, whose senses have no relationship (the geographic and economic senses of 'bank' being one, oft-cited, example), the majority of ambiguous words have senses that are related in some manner. Such words are said to be *polysemous*.

Because of their random selection from a corpus, the pseudo-senses of a pseudo-word have no relationships between them and it is necessary to examine how important this deficiency of pseudo-words is. The relationships between a word's senses can take one of two forms. For one, the lack of pseudo-sense relatedness may be important, for the other, it may be less so. The two are now discussed.

4.2.1 Related but not important?

There are some words which have senses that are related but in a manner that is of little importance when considering pseudo-words. Two examples of this relationship are now described.

- Etymological relationships - senses that are related through some historical link. The word 'cardinal' as a religious figure and 'cardinal' as a mathematical concept are etymologically related.
- Metaphorical relationships - 'surfing' as a water sport or 'surfing' as a pursuit of browsing web pages.

Although the senses of these words are related, it is questionable how important the relatedness is in relation to the accuracy of pseudo-word simulation. The random formation of pseudo-words brings together words that appear in different contexts and it could well be that this is an accurate simulation of etymologically and metaphorically related senses. One can imagine, that the two senses of the word 'surf', will appear in quite different contexts and so in that respect will be similar to the randomly selected pseudo-senses of a pseudo-word.

This behaviour of ambiguous words has been observed in two analyses. Yarowsky *[Yarowsky 93]* tested the hypothesis that a word's different senses appear in different contexts. For the words he examined, Yarowsky found the hypothesis to be "90-99% accurate for binary ambiguities". Gale et al. *[Gale 92b]* examined the broader context of senses showing that if a word is used in a particular sense in a discourse, all other occurrences of the word in that discourse were used in the same sense (this quality of senses was mimicked by the pseudo-words used in the experiments of this paper). From the two studies, it would appear that for the classes of polysemous word described so far, the relatedness of senses may not be as important as imagined.

4.2.2 Related and important

The applicability of the Yarowsky and Gale et al. studies may be limited, however, as both works examined words which had a small number of broadly defined and distinct senses. Many words have a larger number of more tightly related senses and for these, the different context and same discourse rules may not apply as consistently. For example, it is not hard to imagine that the word 'French' could refer to the language and the people within the same discourse surrounded by similar contexts. For words of this kind, the lack of pseudo-sense relatedness may be significant. It is not an issue addressed by the experiments presented here. An examination of this area is left for future work.

5 Applications of ambiguity analysis

From the work presented by the author in *[Sanderson 94]* and that presented in this paper, it has been shown that sense ambiguity impacts on retrieval effectiveness far less than was originally thought. This does not mean, however, that ambiguity should be ignored. It is believed that the factors of query/document length, and of skewed distribution of senses can be used as a means of assessing when ambiguity will be a significant problem to a retrieval system and, therefore, suggest when some form of disambiguation (on document or query) should be investigated.

Already the statements on the utility of disambiguation for short queries have been supported through experimentation by Sanderson *[Sanderson 96]* who showed a small improvement in effectiveness for retrievals based on single word queries when documents and queries were represented by word senses, identified by an automatic disambiguator. Similarly, results of experiments investigating manual disambiguation of short documents (image captions)

by Smeaton & Quigley *[Smeaton 96]* has also provided evidence showing effectiveness improving for this type of retrieval.

The analysis of sense frequency distributions presented in this paper provides an explanation for the results of Schütze & Pedersen *[Schütze 95]* whose use of a disambiguator on large queries and documents resulted in a 7-14% improvement in retrieval effectiveness, the first published results showing an automatic disambiguator working successfully with an IR system.

To understand the reasons for their results, which apparently contradict those presented here, it is necessary to first explain how Schütze & Pedersen's disambiguator worked. Unlike a 'classic' disambiguator, it did not use a dictionary or thesaurus as a source of word sense definitions, instead it used only the corpus to be disambiguated. Its disambiguation method was as follows. For each word in the corpus, the context of every occurrence of that word within the corpus was examined and *common contexts* were clustered. For example, given the word 'ball', one might find that within a corpus of newspaper articles, this word appears in a number of common contexts: a social gathering; and perhaps a number of different sports (tennis, football, cricket, etc.). For Schütze & Pedersen's disambiguator, each one of these common contexts constituted an individual sense of the word. This is what is unusual about their disambiguator: the senses are quite unlike those found in a dictionary. It is unlikely for instance, that a dictionary would distinguish between different types of the sporting sense of 'ball'. A further difference is that the disambiguator only attempted to identify the commonest senses of a word: Schütze and Pedersen stated that a common context was only identified as a sense if it occurred more than fifty times in the corpus. So different are Schütze & Pedersen's senses from the classic definition of the word, that they are referred to here as *word uses* instead.

The differences between uses and senses identified here causes the frequency distribution of word uses to be different from those of word senses. The requirement that uses must occur at least fifty times eliminates the very infrequent and therefore makes the frequency distribution of uses less skewed. In addition, it is likely that the commonest senses of a word will be employed in a number of distinct uses: e.g. the sporting sense of 'ball', mentioned above, written in tennis, football, and cricket contexts. The breaking up of a word's commonest senses would have the effect of causing the frequency distribution of word uses to be less skewed than those of word senses.

From the results in Section 3.2 comparing the detrimental impact on retrieval effectiveness caused by even and skewed distributions, it was shown that even distributions impact more on effectiveness than skewed. As it is believed that word uses have a less skewed, and therefore more even, frequency distribution when compared to word senses, it is concluded that the improvement in retrieval effectiveness reported by Schütze & Pedersen is due to this difference in the frequency distributions[10].

5.1 Other skewed frequency analyses

In this section, other processes of IR are examined using the methodology of analysing distributions of frequencies of occurrence in relation to retrieval effectiveness. The examinations are brief and

10. As stated in the introduction, this explanation is suggested by Schütze & Pedersen though not directly tested through experimentation.

are intended only to show the potential of the analysis rather than to provide a thorough study. Five areas are examined: document signatures, *stemming*, *Optical Character Recognition* (OCR), *Cross Language Information Retrieval* (CLIR), and query expansion.

5.2 Document signatures

As was already shown in Section 2.4, pseudo-words have a potential utility reducing the size of document signatures. Further investigation of the relationship between signature size and retrieval effectiveness, mediated through different forms of pseudo-words may prove useful.

5.3 Stemming

The positive impact of stemming on retrieval effectiveness is at best regarded as minimal. Harman *[Harman 87]* examined three stemming techniques on a set of test collections and concluded that stemming did not result in improvements in retrieval effectiveness. In contrast, Krovetz *[Krovetz 93]* with his more sophisticated stemmer showed a small but consistent and significant, improvement over a number of test collections, in particular, those having short documents. More recently, Xu and Croft *[Xu 98]*, using a corpus based enhancement to the Porter stemmer *[Porter 80]*, have shown further, but again small, improvements over Krovetz's stemmer.

One possible explanation for the relatively small improvements brought about by stemming may lie in the skewed frequency of occurrence of word stem variants. The process of stemming word variants to a morphological root has similarities to the formation of pseudo-words with the difference that stemming is intended to improve retrieval effectiveness. Even with a cursory examination of word variants as shown in Table 4, one can see that their frequency of occurrence appears to follow a similar skew to that found in the analysis of pseudo-words, although with the amount of skew varying.

Variant	Occs.	Variant	Occs.
water	**121**	wonderful	**36**
waters	**31**	wonder	**28**
water's	**1**	wondering	**16**
watered	**1**	wondered	**12**
watering	**1**	wonders	**3**

Table 4. Frequency of occurrence of Porter stem variants of the words 'water' and 'wonder' as measured in a small document collection.

Stemming was also studied by Church *[Church 95]*, who examined the correlation (in terms of document co-occurrence) between the variants of a word stem. Church presented his correlation measure as a way of predicting the worth of stemming a particular word, though it was not actually tested in retrieval experiments. By examining the relative frequency of occurrence of stem variants, it may be possible to complement Church's work by producing an enhanced predictor of the worth of stemming.

5.4 OCR

Smeaton and Spitz *[Smeaton 97]* have examined a type of pseudo-word in OCR called a *Word Shape Token* (WST). These are words composed of a seven letter alphabet, known as *Character Shape Codes* (CSCs), into which the English 52 letter alphabet (capital and lower case) is mapped based on characteristics of letter shape. Smeaton and Spitz state that the advantages of recognising CSCs over letters is an order of magnitude speed increase in recognition along with greater accuracy. The disadvantage is that many words are mapped to the same WST, making WSTs similar to pseudo-words. The amount of concatenation varies depending on the let-

ters of a word and on its length: longer words are, for example, less likely to map to the same WST.

Smeaton and Spitz examined the impact on retrieval effectiveness of retrieving from documents represented by WSTs instead of words. Initial experiments conducted by them showed very large reductions, however, they stated this was due to certain query words mapping to the same WSTs with as many as a thousand other words. Through a process of eliminating these massively concatenated words from queries, the average number of query words mapping to the same WST was around twenty and the reduction in retrieval effectiveness compared to using just words was approximately a half.

In the light of the work presented in this paper, it is anticipated that an analysis of frequency distributions of the component words of WSTs may provide indications of how better to choose which WSTs should be eliminated from a query in order to maximise retrieval effectiveness.

5.5 CLIR

Given the evidence, already presented, on the skewed frequency distribution of senses defined in dictionaries and thesauri, it would seem reasonable to wonder if CLIR systems using translation dictionaries will be good candidates for an analysis of the frequency distribution of the possible translations of words. If for example, the distribution of a word's translations were mostly skewed, and in general its commonest translation was the correct one, then it may be possible that translating a word into all its possible translations would not harm retrieval effectiveness by much. In their experiments, Hull and Grefenstette *[Hull 96]* used a translation dictionary and reported that using a strategy of concatenating a word's possible dictionary translations produces retrieval effectiveness that "performs quite well given its simplicity". However, they also state that introducing incorrect translations "seriously hurts performance". From such statements, it is not really possible to determine much about frequency distribution and ambiguity.

In order to gain an additional understanding about the use of translation dictionaries in CLIR, an analysis of one dictionary, the Collins English-Spanish bilingual machine readable dictionary (as described in *[Ballesteros 97]*), was conducted by measuring the frequency of occurrence of the English translations of Spanish words. Using a method similar to that performed in Section 4 on the sense tagged SEMCOR corpus, the translations were grouped into sets based on the number of translations there were for a Spanish word. The frequency of occurrence of the English translations (after being stemmed) was measured in the TREC-B collection, 1991-93. The results are shown in Table 5. As can be seen, for each of the sets, on average, the commonest translation accounts

Number of translations	Size of set	Commonest translation (%)	
2	5076	**82**	{50}
3	2751	**72**	{33}
4	1557	**64**	{25}
5	961	**57**	{20}
6	625	**53**	{17}
7	447	**49**	{14}
8	268	**44**	{13}
9	190	**40**	{11}
10	126	**36**	{10}

Table 5. Percentage of occurrences accounted for by the commonest translation of a Spanish word (computed by micro averaging). The figures in brackets (shown for comparison) are the percentages that would result if translations occurred in equal amounts. Measurements made using the Collins Spanish to English dictionary and the TREC-B collection.

for the vast majority of occurrences. The dominance of the commonest sense is less strong than that shown in pseudo-words and word senses, but, nevertheless, it is present. This short analysis, seems to suggest that the skewed frequency of occurrence of the possible translations of a word in some part accounts for the relative success of the simplistic translation strategy reported by Hull and Grefenstette.

5.6 Query expansion

The automatic expansion of query words with words chosen from a thesaurus or dictionary has not been successful. Voorhees *[Voorhees 94]* tried automatically expanding the words of TREC queries with related words taken from the WordNet thesaurus *[WordNet]* without success. One of the unusual qualities of TREC queries is their great length (on average 41 non-stop words per query) and one might speculate that the reasons for Voorhees lack of success can be attributed to this feature: perhaps the expansion terms were poor quality and added little to the already large number of terms. At TREC-6 a new task was introduced: the very short query task, ad hoc retrieval based on the title section of TREC queries which were on average 2.5 non-stop words in length.

It was hypothesised that because the queries were shorter, expansion techniques like that tried by Voorhees may be more successful. Therefore, such a method was attempted. The one chosen was a semi-automatic form that required the manual identification of the sense of each query word followed by the automatic expansion of the identified senses with synonyms taken from the WordNet thesaurus. The motivations and results of the experiment are described in detail in the report to TREC-6 *[Crestani 97]*. The main conclusion was that even with the short queries, the expansion method did not improve retrieval effectiveness over a strategy of leaving the query alone. Over the 45 queries tested[11] (queries 251-300), this strategy was found to leave 14 unchanged, improve 8 queries, and degrade 23. In the light of the frequency distribution work reported in this paper, however, a possible improvement to the expansion process was hypothesised.

Perhaps query words used in their commonest sense did not need expansion as their sense would be so prevalent in the collection anyway. If, however, a query word was used in one of its less common senses, expansion might prove useful in ensuring that documents containing that sense were ranked highly. (Assuming of course that the expansion words were used in those documents in their ‘correct’ sense.)

A repeat of the TREC experiment to test this hypothesis was conducted on the TREC-B collection using the same 45 queries described above. Expansion was conducted in the same manner: manually identifying the sense of query words and expanding less common senses with synonyms taken from WordNet. Information on the frequency of occurrence of word senses was gained from WordNet. Although this may not reflect the frequencies of occurrence found in the document collection, it was hoped that this information would be accurate enough for the purposes of this experiment.

Using the strategy of only expanding the less common senses of query words on the TREC queries resulted in 36 queries being left unchanged, 4 improved, and 5 degraded. The increased number of unchanged queries was not surprising given that fewer expansions

11. Five of the fifty queries had no relevant documents and were ignored in this experiment.

took place. The ratio of improved to degraded queries changed from around 1:3 to almost 1:1, although the degradation from the five queries was worse than the improvement from the four. Nevertheless, the study appeared to indicate that the strategy of targeting query words using a less common sense was promising, though obviously one that required improvement before it could be employed in any retrieval system.

6 Conclusions

In this paper, a series of experiments were presented that measured and analysed the impact on retrieval effectiveness of word sense ambiguity. All of these experiments applied a methodology that used a form of artificial ambiguity known as pseudo-words. The skewed frequency distribution of pseudo-senses was described and shown to be one important factor in the impact of ambiguity on effectiveness. Experiments and analyses were presented that showed the skew to be a good model of the frequency distribution of the senses of actual ambiguous words. This provided further evidence that conclusions drawn from experiments based on pseudo-words are applicable to cases of real ambiguity.

Through further experimentation and reference to previous work it was confirmed that the self-disambiguating nature of long queries and the skewed frequency distribution of the word senses are both factors in the low impact on retrieval effectiveness of word sense ambiguity. Recent research on disambiguation and IR was analysed and the two factors were shown to play an important role in explaining the success of these approaches.

Finally, a number of additional processes of retrieval were examined in the light of knowledge about skewed frequency distributions. These analyses appeared to provided insight into the reasons for these processes' impacts on retrieval effectiveness.

7 Acknowledgements

The authors wish to thank Ian Ruthven, Mounia Lalmas, Bob Krovetz and the reviewers for their comments on earlier drafts of this paper. This work was supported by the VIPIR project which was funded by the University of Glasgow.

8 References

Ballesteros 97
L. Ballesteros and W.B. Croft (1997). *Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval*, in Proceedings of ACM SIGIR Conference, 20: 84-91.

Burnett 79
J.E. Burnett, D. Cooper, M.F. Lynch, P. Willett, M. Wycherley (1979). *Document retrieval experiments using indexing vocabularies of varying size. - 1. Variety generation symbols assigned to the fronts of index terms*, in Journal of Documentation, 35(3): 197-206.

Church 95
K.W. Church (1995). *One Term or Two?*, in the Proceedings of the 18th ACM SIGIR Conference: 310-318.

Crestani 97
F. Crestani, M. Sanderson, M. Theophylactou, M. Lalmas (1997). *Short Queries, Natural Language and Spoken Documents Retrieval: Experiments at Glasgow University*, In the proceedings of the 6th TREC conference (TREC-6) published by NIST.

Gale 92a
W. Gale, K.W. Church, D. Yarowsky (1992). *Estimating upper and lower bounds on the performance of word-sense disambiguation programs*, in Proceedings of the 30th ACL, 249-256.

Gale 92b
W. Gale, K.W. Church, D. Yarowsky (1992). *One sense per discourse*, in the Proceedings of (DARPA) Speech and Natural Language Workshop.

Gale 92c
W. Gale, K.W. Church, D. Yarowsky (1992). *Work on Statistical Methods for Word sense Disambiguation*, in Intelligent Probabilistic Approaches to Natural Language Papers from the 1992 Fall Symposium, AAAI Press, Technical Report FS-92-04, ISBN 0-929280-42-3, 54-60

Grefenstette 94
G. Grefenstette (1994). *Explorations in automatic thesaurus discovery*, Boston: Kluwer Academic Publishers.

Harman 87
D. Harman (1987). *A failure analysis on the limitations of suffixing in an online environment*, in Proceedings of the 10th ACM SIGIR Conference: 102-107.

Harman 92
D. Harman (1992). *Ranking Algorithms*, in Information Retrieval: data structures & algorithms, W. B. Frakes & R. Baeza-Yates eds, Prentice Hall: 363-392.

Harman 96
D. Harman (1996). *Overview of the Fifth Text REtrieval Conference (TREC-5)*, in Proceedings of the Text Retrieval Conference (TREC-5), National Institute of Standards and Technology, Gaithersburg, MD 20899, USA. Also available at 'trec.nist.gov/'.

Hull 96
D.A. Hull & G. Grefenstette (1996). *Querying across languages: a dictionary-based approach to multilingual information retrieval*, in Proceedings of the 19th ACM SIGIR Conference, 19: 49-57.

Kilgarriff 97
A. Kilgarriff (1997). *I don't believe in word senses*, in Computers and the Humanities, 31(2): 91-113.

Krovetz 92
R. Krovetz & W.B. Croft (1992). *Lexical Ambiguity and Information Retrieval*, in ACM Transactions on Information Systems, 10(2): 115-141.

Krovetz 93
R. Krovetz (1993). Viewing morphology as an inference process, in Proceedings of the 16th ACM SIGIR Conference, 191-202.

Lesk 86
M. Lesk (1986). *Automatic sense disambiguation: how to tell a pine cone from an ice cream cone*., in Proceedings of the 1986 SIGDOC Conference: 24-26

Lewis 91
D.D. Lewis (1991). Representation and learning in information retrieval, in PhD Thesis, COINS Technical Report 91-93: Department of Computer and Information Science, University of Massachusetts, Amherst, MA 01003.

Miller 95
G.A. Miller (1995). WordNet: A lexical database for English, in Communications of the ACM, 38(11): 39-41.

Ng 96
Hwee Tou Ng & Hian Beng Lee (1996). *Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach*, in Proceedings of the ACL, 34: 40-47.

Porter 80
M.F. Porter (1980). *An algorithm for suffix stripping*, in Program - automated library and information systems, 14(3): 130-137.

Van Rijsbergen 79
C.J. van Rijsbergen (1979). *Information retrieval* (second edition), in London: Butterworths. Also available at 'www.dcs.gla.ac.uk/Keith/Preface.html'

Salton 83
G. Salton, E.A. Fox, H. Wu (1983). *Extended Boolean Information Retrieval*, in Communications of the ACM, 26(11): 1022-1036.

Sanderson 94
M. Sanderson (1994). *Word sense disambiguation and information retrieval*, in Proceedings of the 17th ACM SIGIR Conference, 17: 142-151.

Sanderson 96
M. Sanderson (1996). *Word Sense Disambiguation and Information Retrieval*. PhD Thesis, Technical Report (TR-1997-7) of the Department of Computing Science at the University of Glasgow, Glasgow G12 8QQ, UK.

Schütze 92
H. Schütze (1992). *Context Space*, in Intelligent Probabilistic Approaches to Natural Language Papers from the 1992 Fall Symposium, AAAI Press, Technical Report FS-92-04, ISBN 0-929280-42-3: 113-120.

Schütze 95
H. Schütze & J.O. Pedersen (1995). *Information retrieval based on word senses*, in Proceedings of the Symposium on Document Analysis and Information Retrieval, 4: 161-175.

Small 82
S. Small & C. Rieger (1982). *Parsing and comprehending with word experts (a theory and its realisation)*, in Strategies for Natural Language Processing, W.G. Lehnert & M. H. Ringle, Eds., LEA,: 89-148.

Smeaton 96
A.F. Smeaton & I. Quigley (1996). *Experiments on Using Semantic Distances Between Words in Image Caption Retrieval*, in Proceedings of ACM SIGIR Conference, 19: 174-180.

Smeaton 97
A.F. Smeaton & A.L. Spitz (1997). *Using Character Shape Coding for Information Retrieval*, in Proceedings of the International Conference on Document Analysis and Recognition, 4.

Sparck Jones 76
K. Sparck Jones & C.J. van Rijsbergen (1976). *Progress in documentation*, in Journal of Documentation, 32(1): 59-75.

Sussna 93
M. Sussna (1993). *Word sense disambiguation for free-text indexing using a massive semantic network*, in Proceedings of the International Conference on Information & Knowledge Management (CIKM), 2: 67-74.

Voorhees 93
E. M. Voorhees (1993). *Using WordNet™ to disambiguate word sense for text retrieval*, in Proceedings of the 16th ACM SIGIR Conference, 171-180.

Voorhees 94
E.M. Voorhees (1994). *Query expansion using lexical-semantic relations*, in Proceedings of the 17th ACM SIGIR Conference, 61-70.

Wallis 93
P. Wallis (1993). *Information retrieval based on paraphrase*, in Proceedings of PACLING Conference, 1.

Weiss 73
S.F. Weiss (1973). *Learning to disambiguate*, in Information Storage and Retrieval, 9: 33-41.

Wilks 90
Y. Wilks, D. Fass, C. Guo, J.E. Mcdonald, T. Plate, B.M. Slator (1990). *Providing Machine Tractable Dictionary Tools*, in Machine Translation, 5: 99-154.

WordNet
'www.cogsci.princeton.edu/~wn/'. WordNet was developed by the Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller (Principal Investigator). Ongoing development of WordNet is supported by DARPA/ITO (Information Technology Office).

Xu 98
J. Xu and W.B. Croft (1998). *Corpus-Based Stemming using Co-occurrence of Word Variants*, in ACM Transactions on Information Systems, 16(1): 61-81.

Yarowsky 93
D. Yarowsky (1993). *One sense per collocation*, in Proceedings of the ARPA human language technology workshop.

Zipf 49
G.K. Zipf (1949). *Human behaviour and the principle of least effort: an introduction to human ecology*, in Cambridge (Mass.): Addison-Wesley P.

9 Appendix - availability of resources

At the time of writing this paper, details on how to access the TREC collections could be found on the web at trec.nist.gov. The Cranfield 1400 and CACM collections were at www.dcs.gla.ac.uk/idom/. For further details on formation of pseudo-word transformed collections, please contact the first author.