Monolingual Document Retrieval for European Languages

Vera Hollink (vhollink@science.uva.nl)*,
Jaap Kamps (kamps@science.uva.nl),
Christof Monz (christof@science.uva.nl) and
Maarten de Rijke (mdr@science.uva.nl)
Language & Inference Technology Group, ILLC, U. of Amsterdam
Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands

Abstract. Recent years have witnessed considerable advances in information retrieval for European languages other than English. We give an overview of commonly used techniques and we analyze them with respect to their impact on retrieval effectiveness. The techniques considered range from linguistically motivated techniques, such as morphological normalization and compound splitting, to knowledge-free approaches, such as n-gram indexing. Evaluations are carried out against data from the CLEF campaign, covering eight European languages. Our results show that for many of these languages a modicum of linguistic techniques may lead to improvements in retrieval effectiveness, as can the use of language independent techniques.

1. Introduction

While information retrieval (IR) has been an active field of research for decades, for much of its history it has had a very strong bias towards English as the language of choice for research and evaluation purposes. Whatever they may have been, over the years, many of the motivations for an almost exclusive focus on English as the language of choice in IR have lost their validity. The Internet is no longer monolingual, and non-English content is growing rapidly. Today, less than a third of all domain names is registered in the US, and by 2005 two-thirds of all Internet users will be non-English speaking. Multilingual information access has become a key issue. The availability of cross-language retrieval systems that match information needs in one language against documents in multiple languages is recognized as a major contributing factor in the global sharing of information.

Multilingual IR implies a good understanding of the issues involved in monolingual retrieval. And there are other important factors that motivate monolingual European IR system development. Even in relatively multilingual countries such as Finland and The Netherlands,

* Currently at Social Science Informatics (SWI), Dept. of Psychology, U. of Amsterdam.

users continue to feel the need to access information and services in their native languages. For small European languages such as Dutch and Finnish, the costs of developing and maintaining a language technology infrastructure are relatively high. But languages with inferior computational tools are bound to suffer in an increasingly global society, for both cultural and economic reasons.

What are the issues involved in monolingual retrieval for European languages other than English? One common opinion is that the basic IR techniques are language-independent; only the auxiliary techniques, such as stopword lists, stemmers, lemmatizers, and other morphological normalization tools need to be language dependent (Harman, 1995a). But different languages present different problems. Methods that may be effective for certain languages may not be so for others; issues to be addressed include word order, morphology, diacritic characters, languages variants, etc.

Since its launch in 2000, the *Cross-Language Evaluation Forum* (CLEF) has been the main platform for experimenting with monolingual retrieval for European languages. The aim of this paper is to survey the current state of the art in monolingual retrieval for European languages. We do not aim at presenting an exhaustive overview of all known approaches to monolingual European IR: even if we had enough pages, we doubt whether an encyclopedic catalog would be very insightful. Instead, we focus on two types of approaches. The first concerns approaches that try to exploit language-specific features, such as inflectional morphology. The second type is geared specifically towards simplicity and language-independence. Thus, our focus will be on language-specific versus language-independent techniques for monolingual European IR, with special attention to the lessons learned in the course of the CLEF campaigns, using the CLEF test sets.

The remainder of the paper is organized as follows. In Section 2 we detail our experimental set-up; we refer to the editors' introduction for an overview of the test collections. In Section 3 we present a naive baseline against which more sophisticated approaches can be compared. Section 4 surveys linguistically informed approaches to monolingual European IR, and in Section 5 we consider language independent approaches. In Section 6 we provide a topic-wise analysis of our findings, and then make typological and other observations before concluding.

2. Experimental Setting

CLEF has to a large extent adopted the methodology of the Text REtrieval Conference (TREC), adapting the TREC ad hoc task to meet

the needs of cross-language retrieval (Peters and Braschler, 2001). In particular, multiple collections were made available, one for each of the participating languages. To create a balanced test collection, it is important that the corpus is comparable, meaning that the subcollections must be similar in content, genre, size, and time period. We refer to (Braschler and Peters, this volume) for details on the composition of the corpus at the time of writing (early 2003).

As explained by Braschler and Peters (this volume), CLEF uses the TREC conception of topics: structured statements of user needs from which queries are extracted, with title (T), description (D), and narrative (N) fields. Each topic consists of three fields: a brief title statement, a one-sentence description, and a more complex narrative specifying the relevance assessment criteria. To ensure maximal comparability across multiple languages, we restrict our attention to the 50 topics used at CLEF 2002 (topics 91–140). In all the runs on which we report in this paper we only use the T and D fields of the topics. From the topic descriptions we automatically removed stop phrases such as "Relevant documents report. . . ," "Find documents . . . ," for all eight languages.

All runs were created using the FlexIR system developed at the University of Amsterdam (Monz et al., 2002). FlexIR has been designed to facilitate experimentation with a wide variety of retrieval components and techniques. The retrieval model underlying FlexIR is the standard vector space model. All our runs use the Lnu.ltc weighting scheme (Buckley et al., 1995) to compute the similarity between a query and a document. For the experiments on which we report in this paper, we fixed slope at 0.2; the pivot was set to the average number of unique words per document.

Blind feedback was applied to expand the original query. Term weights were recomputed with the standard Rocchio method (Rocchio, 1971), where we considered the top 10 documents to be relevant and the bottom 500 documents to be non-relevant. In most runs we allowed at most 20 terms to be added to the original query; in some of the n-gram-based runs we allowed as many as 100 terms to be added.

We used stopword lists for each of the eight languages. To increase comparability, we used stopword lists from a single source. We use the stopword lists that come with the Snowball stemmer (more on this stemmer below). Unfortunately, the Finnish Snowball stemmer does not come with a stopword list. For Finnish, we resort to the stopword list created by Jacques Savoy (CLEF-Neuchâtel, 2003).

Table I summarizes the characteristics of the stopword lists for all the eight CLEF languages. Stopwords were removed at indexing time.

Unless indicated otherwise, we applied the same sanitizing operations for all our runs. All words were lowercased; for the runs in which

Table I. Stopword list lengths in number of words for eight European languages.

Dutch	English	Finnish	French	German	Italian	Spanish	Swedish
101	119	1,134	155	231	279	313	114

we used a lemmatizer, lowercasing took place after lemmatizing but before any other text operations took place. Diacritic characters are mapped to the unmarked characters.

To determine whether the observed differences between two retrieval approaches are statistically significant and not just caused by chance, we used the bootstrap method, a powerful non-parametric inference test (Efron, 1979). The method has previously been applied to retrieval evaluation by, e.g., Savoy (1997) and Wilbur (1994). The basic idea of the bootstrap is a simulation of the underlying distribution by randomly drawing (with replacement) a large number of samples of size N from the original sample of N observations. These new samples are called *bootstrap samples*; we set the number of bootstrap samples to the standard size of 1,000 (Davison and Hinkley, 1997). The mean and the standard error of the bootstrap samples allow computation of a confidence interval for different levels of confidence (typically 0.95 and higher). We compare two retrieval methods a and b by one-tailed significance testing. If the left limit of the confidence interval is greater than zero, we reject the null hypothesis, stating that method b is not better than a, and conclude that the improvement of b over a is statistically significant, for a given confidence level. Analogously, if the right limit of the confidence interval is less than zero, one concludes that method b performs significantly worse than a (Mooney and Duval, 1993).

In the following, we indicate improvements at a confidence level of 95% with "$^{\triangle}$" and at a confidence level of 99% with "$^{\blacktriangle}$". Analogously, decreases in performance at a confidence level of 95% are marked with "$^{\triangledown}$" and at a confidence level of 99% with "$^{\blacktriangledown}$". No markup is used if neither an increase nor a decrease in performance is significant at either of the 95% or 99% confidence levels.

3. A Naive Baseline

During the CLEF evaluation campaigns, a wide variety of approaches have been applied to monolingual retrieval in non-English European languages; consult (Peters, 2001; Peters et al., 2002; Peters, 2002) for overviews. One can organize the approaches in two camps. The first con-

sists of linguistically motivated approaches, which require knowledge of, and specific tools tailored to the language at hand. Some examples of these approaches are the use of stemming and lemmatizing. The second category are knowledge-poor approaches, which require little or no language-dependent knowledge. Examples are approaches like (character) n-grams of various lengths that may span word boundaries. Before exploring examples of both categories (in Sections 4 and 5) we present a simple baseline against which to compare later runs.

In our baseline runs we simply index the words as they are encountered in the collection. We do some cleaning up: diacritics are mapped to the unmarked character, and all characters are put in lower-case. Thus, a string like the German *Raststätte* (English: *road house*) is indexed as *raststatte*. Table II, column 3 lists mean average precision (MAP) scores for our baseline runs on the CLEF 2002 topics, for each of the eight languages, with stopword removal as detailed in Section 2.

A few observations are worth making. First, the scores vary considerably across the eight languages. Compared to current state of the art systems, for some languages (Dutch, English, Spanish) the word-based baseline performs very well. Second, the Finnish scores are very low; this may be due to the small size of the Finnish collection.

Table II. Mean average precision scores for the word-based baseline run, using the CLEF 2002 topics.

	Diacritic characters		
Language	kept	removed	% change
Dutch	0.4089	0.4482	+9.6%▲
English	0.4370	0.4460	+2.1%
Finnish	0.2061	0.2545	+23.4%
French	0.3627	0.4296	+18.5%▲
German	0.3812	0.3886	+1.9%
Italian	0.3764	0.4049	+7.6%△
Spanish	0.3944	0.4537	+15.0%
Swedish	0.2684	0.3203	+19.3%▲

In most (non-English) European languages, accents are used to indicate the precise pronunciation and to identify some homographs. The exact meaning of a phrase may be affected when accents are removed as, for example, in the French *un dossier critiqué* (English: *a criticized case*) and *un dossier critique* (English: *a critical case*). Intuitively, removing

accents may improve overall recall, but this might be counterbalanced by a loss of precision, due to false conflations.

To evaluate the relative importance of diacritic characters for retrieval purposes, we created baseline runs where marked characters are indexed as they occur in the collections. The mean average precision results are listed in Table II, column 2. For all of the languages, replacing marked characters by the unmarked characters leads to improvements, and for four of the languages (Dutch, French, Italian, and Swedish) the improvement is significant. Our results for French contradict Savoy's (1999) findings for the removal of diacritic characters; he found that ignoring accents in French does not significantly hurt precision, but it does not increase it either.

4. Using Morphological Normalization

It is widely held that the selection of index terms should exploit morphological features of the words occurring in the text collection (Frakes, 1992; Krovetz, 1993). Traditionally, inflectional and derivational morphology are distinguished (Matthews, 1991). *Inflection* is defined as the use of morphological methods to form inflectional word forms from a lexeme; inflectional word forms indicate grammatical relations between words. For example, the plural *books* is distinguished from the singular *book* by the inflection *-s*. Derivational morphology is concerned with the *derivation* of new words from other words using derivational affixes. For instance, *hanger* is derived from *hang*, and *countess* from *count*. Compounding, or composition, is another method to form new words. A *compound* is a word formed from two or more words written together; the component words themselves are independent words. For instance, the Dutch compound *zonnecel* (English: *solar cell*) is a combination of *zon* (English: *sun*) and *cel* (English: *cell*).

In theory, the three main morphological phenomena (inflection, derivation, and compounding) all affect retrieval effectiveness. Documents are not retrieved if the search key does not occur in the index. For effective retrieval morphological processing is needed in most languages to handle variant word forms. Morphological normalization — in the form of stemming, or otherwise — was originally performed for two principle reasons: the large reduction in storage required by a retrieval dictionary (Bell and Jones, 1979), and the increase in performance due to the use of word variants; in particular, *recall* can be expected to improve as a larger number of potentially relevant documents are retrieved (Hull, 1996). In the setting of non-English European languages with a complex morphology, such as Slovene or Finnish, a third rea-

son has been identified for performing morphological analysis: in such languages it may be difficult to formulate good queries without morphological programs (Popovic and Willett, 1992; Pirkola, 1999). Due to the common availability of computational resources, recent research has been more concerned with performance improvement than with storage reduction or support for query formulation.

In this section we consider the impact on retrieval effectiveness of three levels of morphological analysis: stemming, lemmatization, and compound splitting.

4.1. Stemming

We used stemmers implemented in the Snowball language (Snowball, 2003). Snowball, a string processing language, is specifically designed for creating stemming algorithms for use in IR. Partly based on the Porter stemmer for English (Porter, 1980), it aims to provide stemming algorithms for languages other than English. There are Snowball stemmers available for all the eight European languages we consider here. For our stemmed runs we perform the same sanitizing operations as for our earlier word-based runs; in particular, we removed stopwords before applying stemming. We made special efforts to make the runs as similar as possible across languages, but subtle differences between the runs remain. For instance, the Snowball stemmers are all based on the same stemming principles, but the specific rule sets may differ in quality between the languages.

Table III. Mean average precision scores for the word-based baseline runs, the stemmed runs, and the lemmatized runs. Best scores per language are in boldface.

Language	Word-based (baseline)	Stemmed	% change	Lemmatized	% change
Dutch	0.4482	**0.4535**	+1.2%	–	
English	0.4460	**0.4639**	+4.0%	0.4003	−10.2%
Finnish	0.2545	**0.3308**	+30.0%▲	–	
French	0.4296	**0.4348**	+1.2%	0.4116	−4.2%
German	0.3886	**0.4171**	+7.3%△	0.4118	+6.0%△
Italian	0.4049	**0.4248**	+4.9%	0.4146	+2.4%
Spanish	0.4537	**0.5013**	+10.5%▲	–	
Swedish	0.3203	**0.3256**	+1.7%	–	

Column 3 in Table III shows the mean average precision scores for the stemmed runs. A few things are worth noting. On top of the high-performing baseline runs for Dutch and English, there is little improvement. For Spanish stemming does yield a significant improvement. For the other two Romance languages (French and Italian) the baseline performance is improved, but not significantly. There are significant improvements for Finnish and German, but not for Swedish.

How do these results compare to findings in the literature on the effect of stemming on retrieval performance? (Kraaij and Pohlmann, 1996) report that for Dutch the effect of stemming is limited; it tends to help as many queries as it hurts. For English, previous retrieval experimentation did not show consistent significant improvements by applying rule-based stemming (Frakes, 1992; Harman, 1991). Likewise, for German and French, there are reports indicating results similar to those for English (Moulinier et al., 2001). Our results for German indicate a significant improvement, in line with (Tomlinson, 2002a; Braschler and Ripplinger, 2003). The significant improvement for Spanish differs from earlier findings on the impact of stemming on retrieval effectiveness; for instance, (Figuerola et al., 2002) report a minor positive impact of inflectional stemming over a word-based baseline, and a negative impact of derivational normalization.

For Italian, improvements similar to ours have been reported for a similar stemming algorithm (Tomlinson, 2002b; Tomlinson, 2002a). An interesting experiment on deriving a stemming algorithm purely based on corpus statistics is reported in (Bacchin et al., 2002). Their affix removal procedure improves retrieval effectiveness, be it slightly less than a Porter-style stemming using linguistically informed rules.

For Swedish and Finnish, morphological normalization tools are few and far between. The results for Swedish reported by (Tomlinson, 2002a) agree with our findings. Finally, for Finnish, we realize our highest improvement of retrieval effectiveness. This agrees with other reports in the literature. The experiments of (Tomlinson, 2002a, p.208) also show the biggest improvement for Finnish. A new Finnish stemming algorithm trying to "conflate various word declinations to the same stem" is reported in (Savoy, 2002b, p.33). The use of a commercial morphological normalization tool is reported in (Hedlund et al., 2002; Airio et al., 2002).

Finally, the scores in columns 2 and 3 do not indicate any cross-lingual phenomena. Stemming significantly helps retrieval effectiveness for some languages, from different language families (both Germanic and Romance), but it hardly affects the performance for other languages from the very same language families.

4.2. Lemmatization

Porter-like stemmers in the Snowball family can be very aggressive, and may produce non-words. For example, the description field of the Dutch version of Topic 95 reads

> *Zoek artikelen over gewapende conflicten in de Palestijnse gebieden en de betrokkenheid van een deel van de bevolking bij dit geweld.*

(English: Find articles dealing with armed conflicts in the Palestinian territories and the involvement of a part of the civil population.) After case folding, stopping, and stemming, this yields

> *palestijn conflict artikel* *gewap conflict palestijn gebied betrok* *del bevolk geweld*

where non-words are marked with an asterisk. In this section we report on the use of a lexical-based stemmer, or lemmatizer, instead of a stemming algorithm.

As before, to increase comparability across languages, we tried to use a single (family of) lemmatizer(s) for as many of the eight languages as possible. The lemmatizers that we ended up using are part of TreeTagger (Schmid, 1994), a probabilistic part-of-speech tagger based on decision trees; unfortunately, this tagger is only available for English, French, German, and Italian. For our retrieval purposes we did not use the part-of-speech information provided by TreeTagger, but only the lemmas it produces. To each word TreeTagger assigns its syntactic root by lexical look-up. Mainly number, case, and tense information is removed, leaving other morphological processes intact.

The results of using a lemmatizer instead of a stemmer for English, French, German and Italian are listed in the fourth column of Table III. The lemmatized run yields significant improvements over the baseline for German only. For Italian, there is an improvement, while for English and French, there are drops in retrieval effectiveness, although none of these are significant.

4.3. Compound Splitting

So far, we have considered three types of retrieval runs, exploiting increasingly deep levels of morphological analysis: word-based, stemmed, and lemmatized. In this subsection we go one step further. The eight European languages considered here differ widely in the amount of *compound* formation they admit. Compounds in English are typically joined by a space or hyphen, think of *computer science* or *son-in-law*, but there are exceptions such as *iceberg*, *database* or *bookshelf*. Compounds of the latter kind are common in Dutch, Finnish, German, and Swedish. Compounds can simply be a concatenation of several words,

but sometimes a linking element is used.[1] Examples of this phenomenon in English are rare, although compounds like *spokesman* use a linking element -s- (Krott et al., 2001). Finnish does not use linking elements, but they occur frequently in Dutch, German, and Swedish. Linking elements in Dutch include -s-, -e-, and -en- (Krott et al., 2001). In German, linking elements include -s-, -n-, -e-, and -en- (Demske, 1995). Finally, Swedish linking elements include -s-, -e-, -u-, and -o- (Josefsson, 1997). Hedlund (2002) lists even more linking elements for German and Swedish. Note that the linking elements are by no means obligatory for compound formation; in Dutch, German, and Swedish compounds without linking elements are abundant.

In some cases, compound splitting can give awkward results. The German *Bahnhof* (English: *train station*), for instance, is split into *Bahn* (English: *rail*) and *Hof* (English: *court/yard*). While 'rail' is semantically related to 'train station,' this is less obvious for 'court' or 'yard.' A more dramatic example is provided by the Dutch *brandstof* (English: *fuel*), which is split into *brand* (English: *fire*) and *stof* (English: *dust/matter*), two words only loosely related to the compound. Hence, it may happen that compound splitting adds unrelated words to a document, thus causing a topic drift. A safeguard against such topic drift is to add compound parts while retaining the original compound word. In our experiments, we only retain the minimal parts of a compound and the compound itself. If a compound is more complex, i.e., contains more than two compound parts, intermediate compound parts could in principle also be considered.

Most authors use corpus or lexicon based approaches for identifying and splitting compounds. A notable exception is Savoy (2002b) who uses a rule-based approach for German compounds. This obviously requires an in-depth knowledge of the language at hand. A common technique is to use a standard lexicon or dictionary as a source of words that may occur in a compound, see e.g., (Kraaij and Pohlmann, 1996; Chen, 2002). This approach may suffer from the fact that plurals are usually not included in a lexicon or dictionary, yet are frequently used in compound formation. An alternative is to consider the words of the corpus as potential base words. This is the approach used for the experiments in this section. There have been many experiments with refinements of this approach, e.g., by considering syntactic categories of words (Monz and de Rijke, 2002; Braschler and Ripplinger, 2003), or by considering translation resources (Hedlund, 2002; Koehn and Knight, 2003).

[1] Linking elements are also referred to as connectives, interfixes, linkers, linking morphemes, or fogemorphemes.

We use the algorithm reported in (Monz and de Rijke, 2002) for identifying compounds and decompounding; Figure 1 shows the pseudo-code for the recursive compound splitting function `split`.

```
1  string split(string s)
2   {
3     int length = strlen(s);
4     string r;
5     for(int char_pos=1; char_pos<=length; char_pos++)
6     {
7       if(substr(1,char_pos,s)∈lexicon
8        && !strcmp(split(substr(char_pos+1,length,s)),''))
9       {
10        r = split(substr(char_pos+1,length,s));
11        return concat(substr(1,char_pos,s),+,r);
12      } else if(substr(1,char_pos,s)∈lexicon
13        && strcmp(substr(char_pos+1,char_pos+1,s),'s')
14         && !strcmp(split(substr(char_pos+2,length,s)),''))
15      {
16        r = split(substr(char_pos+2,length,s));
17        return concat(substr(1,char_pos,s),+,r);
18      };
19    };
20    if(s∈lexicon) return s else return '';
21  }
```

Figure 1. Pseudo-code for the algorithm underlying the compound splitter; note that we have only included the case for the linking element -s- (lines 13, 14); the full list of linking elements considered is given in the main text.

The function `split` takes a string, i.e., a potentially complex noun, as argument and it returns a string where the compound boundaries are indicated by a plus sign. For instance, `split(bahnhof)` returns `bahn+hof`. If it cannot split a string into smaller components it returns the same string, and if it fails to analyze a string at all, it returns the empty string.

We used the words in the collection as our lexicon, plus their associated collection frequencies. We ignore words of length less than four as potential compound parts, thus a compound must have at least length eight. As a safeguard against oversplitting, we only regard compound parts that have a higher collection frequency than the compound itself. We consider linking elements -s-, -e-, and -en- for Dutch; -s-, -n-, -e-, and -en- for German; and -s-, -e-, -u-, and -o- for Swedish. We prefer a split with no linking element over a split with a linking element, and a split with a single character linker, over a two character linker.

For retrieval purposes, we decompound both the documents and queries by keeping the compound word and adding its minimal compound parts. This approach has some side effects whose impact is not clear yet. For example, what is an appropriate matching strategy for compounds? In our implementation, compounds and their parts are treated independently of each other, i.e., the term weight (tf.idf score) is computed independently for the compound and its parts. While this approach seems overly simplistic (since the compound parts may be conceptually related to the compound), it rewards compound matching in contrast to simple term matching, which seems appropriate since compounds are more specific than their compound parts. This issue of compound matching and assigning weights to compounds is similar to the problem of phrase matching and phrase weighting in English (Fagan, 1987; Strzalkowski, 1995).

Table IV. Mean average precision scores for the four compound-rich languages Dutch, Finnish, German, and Swedish, using the CLEF 2002 topics. Best scores are in boldface.

	Word-based			Stemmed		
Language	(baseline)	Split	% change	(baseline)	Split+Stem	% change
Dutch	0.4482	0.4662	+4.0%	0.4535	**0.4698**	+3.6%
Finnish	0.2545	0.3020	+18.7%$^{\triangle}$	0.3308	**0.3633**	+9.8%
German	0.3886	0.4360	+12.2%$^{\triangle}$	0.4171	**0.4816**	+15.5%$^{\blacktriangle}$
Swedish	0.3203	0.3395	+6.0%	0.3256	**0.4080**	+25.3%$^{\blacktriangle}$

Table IV lists the mean average precision scores for compound-splitting of the word-based run; and for compound splitting of the stemmed run where the plain words are split first, and then processed by the stemming algorithm. The improvement of compound splitting over the word-based run, in column 4, ranges from 4% to 18.7%. The improvement of compound splitting over stemming, indicated in column 7, ranges from 3.6% to 25.3%. The combined improvement of splitting and stemming over the word-based runs ranges from 5% for Dutch to 43% for Finnish.

Kraaij and Pohlmann (1996), Monz and de Rijke (2002), and Chen (2002) show that compound splitting leads to improvements in monolingual retrieval performance for Dutch, and Moulinier et al. (2001), Monz and de Rijke (2002), Chen (2002), Savoy (2002b), and Braschler and Ripplinger (2003) obtain similar results for German. Chen (2002) conducts decompounding experiments for German and Dutch, and reports a similar impact on the effectiveness when combined with stem-

ming. Hedlund (2002) reports on the effectiveness of compound splitting for Swedish. The study of compounding and the combinatorial behavior of compounds in the setting of cross-lingual retrieval has also received a fair amount of attention; see, e.g., (Hedlund, 2002; Koehn and Knight, 2003).

5. Using n-Grams

The wish to retrieve documents in arbitrary languages and over arbitrary domains has led a various authors to avoid language-dependent resources such as stopword lists, lexicons, decompounders, stemmers, lemmatizers, phrase lists, and manually-built thesauri. Instead, many such teams have considered retrieval approaches based on n-grams.

Word and character n-grams have a long history. The area of speech recognition has seen much work in n-grams. An especially big boost in their use came from Jelinek, Mercer, Bahl, and colleagues at the IBM Thomas J. Watson Center, and Baker at CMU. These two labs independently used word n-grams in their speech recognition systems; (Jelinek, 1990) summarizes many early language modeling innovations. Much recent work on language modeling has focused on ways to build more sophisticated n-grams; see (Jurafsky and Martin, 2000). Over the years, there have been various attempts at using word-based n-grams to improve retrieval in several European languages; see, e.g., (Kraaij and Pohlmann, 1998; Amati et al., 2002).

Character n-grams also have a long history, beginning, perhaps, with the work of (Shannon, 1951). They have been used for text compression (see, for instance, (Wisniewski, 1987)), spelling-related applications (see, for instance, (Ullman, 1977)), and general string searching (Kotamarti and Tharp, 1990). In information retrieval (character) n-grams have been used since the late 1970s, by (Burnett et al., 1979; Willet, 1979; De Heer, 1982), amongst others. Their use of n-grams was aimed mainly at developing language independent indexing and retrieval techniques. In a series of papers, starting at TREC-7, Mayfield and McNamee (1999) and McNamee and Mayfield (2002b) have advocated the use of n-grams in both monolingual and multilingual retrieval.

5.1. n-Grams Based on Words

Retrieval approaches based on n-grams use the following simple scheme. An n-long window is slid along the text, moving one character at a time; at each position of the window the sequence of characters in the window is recorded. The document (and the topics) are then represented by

the n-grams thus recorded. Despite its simplicity, this scheme allows for many variations. Some authors allow the sliding window to cross word boundaries; inspired by ideas in (Damashek, 1995), Mayfield and McNamee (1999) seem to have been the first authors to implement this technique. Many authors do not allow their n-grams to cross word boundaries. Some authors apply a stopword list before gathering, and some mix n-gram-based approaches with linguistically motivated ideas.

In this section we present results on the use of (character) n-grams, of varying length, for retrieval purposes. What is the most appropriate length of n-grams to be used? One rule of thumb found in the literature is to let n be the largest integer that is less than the average word length in the collection (Savoy, 2002a). Table V gives the average word

Table V. Average word lengths for all collections.

Dutch	English	Finnish	French	German	Italian	Spanish	Swedish
5.4	5.8	7.3	4.8	5.8	5.1	5.1	5.4

lengths for the eight European languages. The use of character n-grams increases the size of both dictionaries and inverted files, typically by a factor of five or six, over those of comparable word-based indices. This may be a disadvantage in less memory-rich environments (McNamee and Mayfield, 2002a).

Table VI. Mean average precision scores for runs using 4-grams and 5-grams (within word boundaries) and 6-grams (across word boundaries), using the CLEF 2002 topics. The baseline is a simple word-based run; all comparisons are against this baseline. Best scores are in boldface.

Language	Word-based (baseline)	4-gram (within)	5-gram (within)	6-gram (across)
Dutch	0.4482	0.4495 (+0.3%)	0.4401 (−1.8%)	**0.4522** (+0.9%)
English	0.4460	**0.4793** (+7.3%)	0.4341 (−2.7%)	0.4261 (−4.5%)
Finnish	0.2545	0.3536 (+37.4%)▲	**0.3762** (+47.8%)▲	0.3560 (+39.9%)△
French	0.4296	**0.4583** (+6.7%)	0.4348 (+1.2%)	0.4427 (+3.1%)
German	0.3886	0.4679 (+20.3%)▲	**0.4699** (+20.9%)▲	0.4574 (+17.7%)▲
Italian	0.4049	**0.4355** (+7.6%)▲	0.4140 (+2.3%)	0.3980 (−1.7%)
Spanish	0.4537	0.4605 (+1.5%)	0.4648 (+2.5%)	**0.4671** (+3.0%)
Swedish	0.3203	**0.4080** (+27.4%)▲	0.3854 (+20.3%)△	0.3942 (+23.1%)△

We generated our n-gram runs after performing the basic sanitizing operations described in Section 2, on both the topics and the documents, and after removing stopwords. We used (sliding) n-grams of length 4 and 5 without crossing word boundaries, and n-grams of length 6 that do cross word boundaries. To give an example, the Dutch version of Topic 108 contains the phrase *maatschappelijke gevolgen* (English: *societal consequences*); using 6-grams sliding across word boundaries, this becomes:

> *... maatsc aatsch atscha tschaa schapp chappe happel appeli ppelij pelijk elijke lijke_ ijke_g jke_ge ke_gev e_gevo _gevol gevolg ...*

In Table VI we summarize the scores for the n-gram-based runs; n-grams provide large (and significant) increases in mean average precision scores over the word-based baseline in 4 of the 8 languages. There does not seem to be an obvious correlation between average word length and the n-gram-length of the best performing run per language. The best settings vary from one language to another; even within a single language family (such as West Germanic, to which Dutch, English, and German belong), we get different optimal n-gram-length settings.

5.2. n-Grams Based on Morphologically Normalized Terms

What is the effect of first carrying out language-dependent morphological normalization steps, and then creating index terms by means of n-grams? Do we gain anything if n-grams are formed using stems or lemmas, instead of words? In Tables VII and VIII we consider the

Table VII. Mean average precision scores for runs for which 4-grams and 5-grams (within word boundaries) and 6-grams (across word boundaries) were formed after stemming, using the CLEF 2002 topics. The baseline is formed by the stemmed runs discussed in Subsection 4.1; all comparisons are against this baseline. Best scores are in boldface.

Language	Stemmed (baseline)	4-gram-stem (within)	5-gram-stem (within)	6-gram-stem (across)
Dutch	**0.4535**	0.4372 (−3.6%)	0.4462 (−1.6%)	0.4524 (−0.2%)
English	**0.4639**	0.4075 (−12.2%)	0.3795 (−18.2%)$^{\blacktriangledown}$	0.4245 (−8.5%)
Finnish	0.3308	0.3644 (+10.2%)	**0.3935** (+20.8%)$^{\triangle}$	0.3898 (+17.8%)
French	0.4348	0.4058 (−6.4%)	0.3876 (−10.9%)$^{\triangledown}$	**0.4364** (+0.4%)
German	0.4171	0.4539 (+8.8%)	0.4271 (+2.4%)	**0.4702** (+12.7%)$^{\blacktriangle}$
Italian	**0.4248**	0.3881 (−8.6%)	0.3605 (−15.1%)$^{\blacktriangledown}$	0.3808 (−10.4%)$^{\triangledown}$
Spanish	**0.5013**	0.4468 (−10.9%)$^{\blacktriangledown}$	0.4226 (−15.7%)$^{\blacktriangledown}$	0.4586 (−8.5%)$^{\triangledown}$
Swedish	0.3256	**0.4010** (+23.2%)$^{\blacktriangle}$	0.3857 (+18.5%)$^{\triangle}$	0.3876 (+19.0%)$^{\triangle}$

combinations across all eight languages. Column 2 in Table VII repeats the mean average precision results for the stemmed runs described previously. Columns 3, 5 and 7 give the results of running the Snowball stemmer first, and then generating 4-grams, 5-grams, and 6-grams, respectively, from Snowball's output. The application of n-gramming after stemming does not yield uniform gains or losses in scores across languages, or language families. As with n-gramming plain words (Table VI), we see significant gains for Finnish, German, and Swedish. Unlike n-gramming plain words, we now also have significant drops in scores (for English, French, Italian, and Spanish) for some settings.

Table VIII. Mean average precision scores for runs for which 4-grams and 5-grams (within word boundaries) and 6-grams (across word boundaries) were formed after lemmatization, using the CLEF 2002 topics. The baseline is formed by the lemmatized runs discussed in Subsection 4.2; all comparisons are against this baseline. Best scores are in boldface.

Language	Lemmatized (baseline)	4-gram-lemma (within)	5-gram-lemma (within)	6-gram-lemma (across)
English	0.4003	0.4133 (+3.3%)	0.3845 (−4.0%)	**0.4273** (+6.8%)
French	0.4116	**0.4454** (+8.2%)$^{\triangle}$	0.4318 (+4.9%)	0.4381 (+6.4%)
German	0.4118	**0.4869** (+18.2%)$^{\blacktriangle}$	0.4548 (+10.4%)$^{\triangle}$	0.4759 (+15.6%)$^{\blacktriangle}$
Italian	**0.4146**	0.4068 (−1.9%)	0.3877 (−6.5%)	0.3924 (−5.4%)

In Table VIII, column 2 recalls the results for the earlier lemmatized runs, and columns 3, 5 and 7 give the results of creating lemmas first, and then generating 4-grams, 5-grams, and 6-grams, respectively, from the lemmas. As before, the picture that emerges is mixed: significant gains for German (as with n-gramming words or stems) and now also for French, (non-significant) drops in scores for Italian, and (non-significant) drops and gains for English.

6. Discussion and Conclusions

To conclude this paper we start with a brief topic-wise analysis of the CLEF 2002 test suites. We then discuss some typological issues, and wrap up with general observations.

6.1. Topic-Wise Analysis

The CLEF 2001 and 2002 evaluation campaigns use a single set of 50 topics each; CLEF 2000 used 40 topics. The set of topics is translated

into all the collection languages by native speakers of the respective languages. The use of a single set of topics in all eight CLEF 2003 languages creates a unique opportunity for comparing the relative performance of topics across languages. Additionally, by looking at the mean score per topic over the eight languages, we can abstract from accidental features caused by the particular choice of words in the topic formulation. This may lead to a better understanding of the types of topics that are hard or easy for information retrieval systems. A full-fledged exposition of this type of analysis requires a full paper in its own right. Here, we discuss our five best and five worst scoring CLEF 2002 topics in detail. For an analysis of the English and German CLEF 2001 topics, consult (Womser-Hacker, 2002).

Table IX. Average precision scores per topic for the word-based runs, restricted to the five best performing CLEF 2002 topics (marked with •) and the five worst performing topics (marked with °). The last column lists the mean of the average precision scores per topic.

Topic	Dutch	English	Finnish	French	German	Italian	Spanish	Swedish	Mean
94•	0.8199	0.8324	0.0158	0.7778	0.9315	0.5237	0.8595	0.9500	0.7138
98•	0.9444	1.0000	0.5020	0.8166	0.5957	0.6057	0.4645	0.8333	0.7203
107°	0.0304	0.1322	0.0000	0.1349	0.1165	0.1117	0.0604	0.0694	0.0819
109°	0.0333	0.2100	0.0266	0.0960	0.0037	0.1621	0.5243	0.0629	0.1399
111°	0.0001	0.5453	0.0000	0.0164	0.0368	0.0360	0.0236	0.0143	0.0841
115°	0.0281	0.1774	0.0000	0.0420	0.0104	0.5366	0.2018	0.0065	0.1253
119•	0.5204	0.7693	0.1520	0.8952	0.7486	0.7993	0.7203	0.6820	0.6609
123•	0.8434	0.5471	0.5588	1.0000	1.0000	0.8498	0.8783	0.9096	0.8234
128°	0.0298	0.1083	0.0193	0.2656	0.1430	0.0618	0.3420	0.1548	0.1406
130•	0.6231	0.5042	0.6000	0.6106	0.7617	0.3279	0.7913	0.7196	0.6173

The Finnish collection covers only 30 of the 50 CLEF 2002 topics, and for two additional topics the English collection contains no relevant documents. We restrict our attention to the five best and five worst scoring topics amongst the remaining set of 28; the average precision scores for these ten topics are shown in Table IX. The average precision scores can radically differ over topics (Harman, 1994). But the scores (of the same topics) across multiple languages tend to be more robust.

Our worst scoring topic is Topic 107, about the effect of genetic engineering on the food chain. The mean score over the eight languages is 0.0819. Our best scoring topic is Topic 123, about the Jackson-Presley marriage, with a mean score of 0.8234 over the eight languages. Why is Topic 123 'easier' than Topic 107? Topic 123 contains proper

names (Michael Jackson and Lisa Mary Presley). All our top 5 scoring topics (94, 98, 119, 123, 130) somehow involve proper names (94: return of Solzhenitsyn; 98: films by the Kaurismäkis; 119: destruction of Ukrainian nuclear weapons; 123: marriage Jackson-Presley; 130: death of Nirvana leader). This confirms the intuition that proper names are good discriminatory terms, even for non-English. Note, however, that in a multilingual setting such as CLEF, spelling variants may undo the positive impact of using proper names. Although Topic 94 is the third best scoring topic with a mean score of 0.7138, the score for Finnish is remarkably low (0.0158). This may be caused by the failure to relate different forms of the name (the Finnish corpus contains words like *Solzhenitsyjen*, *Solzhenitsynia*, *Solzhenitsynin*, *Solzhenitsyn*, and *Solzhenitsyneille*). This is a clear case where techniques like stemming or n-gramming help retrieval effectiveness.

Our 5 worst topics (107, 109, 111, 115, 128) do not deal with proper names but with very general terms that have relatively high collection frequencies (107: genetic engineering; 109: computer security; 111: computer animation; 115: divorce statistics; 128: sex in advertisements). This explains part of the poor performance on these topics; in some cases an additional cause may be the somewhat awkward back-translations of some of these general phrases. For example, in Topic 107, 'genetic engineering' is translated in Dutch as *genetische manipulatie*. While this is a valid translation, in Dutch it is common to use the original English phrase.

In summary, the scores (of the same topic) tend to be fairly robust across the CLEF languages, but the special multilingual setting provided by CLEF may affect the scores for reasons that are absent in the traditional monolingual English setting. We tried to identify some of the differences between the best and worst scoring topics. This is still a far cry from understanding which topics are easy or hard for information retrieval systems: predicting the difficulty of topics is notoriously hard (Voorhees and Harman, 1998).

6.2. Typological Observations

With the advent, and continuing expansion, of the CLEF evaluation campaign, we have evaluation test suites for eight European languages, from various language families. To what extent can we draw typological conclusions from the work presented here? Many of the traditional classifications of European languages use the following families:

- *West Germanic languages* (e.g., Dutch, English, and German)
- *Scandinavian languages* (e.g., Swedish)

– *Romance languages* (e.g., French, Italian, and Spanish)

– *Finno-Ugric languages* (e.g., Finnish);

see e.g., (Whaley, 1997). It is not clear how this traditional classification is useful for retrieval purposes. As we have seen in Subsections 4.1, 4.2, and 4.3, the effectiveness of a particular morphological normalization method such as stemming or compound splitting does not correlate with this classification. We seem to need a more fine-grained classification at the level of language features before we can draw cross-language conclusions. The kind of features that we need to take into account include, for instance, the extent to which a language has compounding.

Recent morphological typology offers promising possibilities here. Based on traditional typology, it operates with two independent variables, *index of synthesis* and *index of fusion*. The first refers to the amount of affixation in a language, the second to the ease with which morphemes can be separated from other morphemes in a word. Pirkola (2001) argues that in languages of low inflectional index of synthesis and low inflectional index of fusion, inflection does not interfere with term matching to the same degree as in languages of high indexes. Within one language these indices could be used to predict the effectiveness of morphological processing, and between languages they could be used to compare the results of monolingual experiments.

6.3. Concluding Remarks

Although many of the traditional boundaries are disappearing in today's global information society, linguistic barriers are still omnipresent. CLEF and other evaluation campaigns can make an important contribution to breaking down these barriers. In this paper our focus has been on the basic task in a multilanguage information retrieval setting: monolingual retrieval for a variety of European languages. Arguably, an effective monolingual retrieval system is the core component of a bilingual, or multilingual, retrieval system.

We have given an overview of commonly used, reasonably generic techniques and we have analyzed them with respect to their impact on retrieval effectiveness. The techniques considered range from linguistically motivated techniques, such as morphological normalization and compound splitting, to knowledge-free approaches, such as n-gram indexing. Evaluations were carried out against data from the CLEF campaign, covering eight European languages. In our experiments we found that the following approaches consistently improve mean average precision (although not always significantly): removing diacritics; stem-

ming; compound splitting (for Dutch, Finnish, German, and Swedish); and n-grams (for 4-grams generated from words).

Table X. Summary of the top scoring runs per language, using the CLEF 2002 topics. (The two runs listed for Swedish achieved the same top score.)

Language	Type of run
Dutch	Split, and then stemmed
English	Words, 4-grammed
Finnish	Stemmed, and then 5-grammed
French	Words, 4-grammed
German	Lemmatized, and then 4-grammed
Italian	Words, 4-grammed
Spanish	Stemmed
Swedish	Words, 4-grammed/Split, and then stemmed

We have summarized our top scoring runs per language in Table X. For Finnish, German, Italian, Spanish, and Swedish, the top scoring run is significantly better than a naive baseline where words are taken as index terms (with diacritics removed). For all languages except English the top scoring run significantly improves over a run in which words are indexed as they occur in the collections, i.e., with marked characters.

What, if any, is the general conclusion resulting from our findings? As we have just seen, there is no uniform best combination of settings. The top scoring runs for five languages (Dutch, Finnish, German, Spanish, Swedish) employ a modicum of linguistic techniques. However, six of the top scoring runs employ n-gram-based indexing, either directly, or after performing a linguistically informed preprocessing step. The best linguistically informed technique, over all eight languages, is "splitting, and then stemming" (only splitting the compound-rich languages, Dutch, German, Finnish and Swedish). The best language independent technique, again over all languages, is "4-gramming of words."

These observations give rise to two hypotheses, viz. that the uniform strategies "splitting, and then stemming" and "4-gramming of words" are indeed the best strategy for all languages. If we test whether the top scoring runs in Table X are significantly better than the respective uniform strategies, we find the following. The hypothesis "4-gramming of words is best" is contradicted for Spanish. The Spanish stemmed run does significantly improve over the Spanish 4-gram run (with confidence 99%). For all the other languages, however, there is no run that significantly improves over the 4-gram run. The hypothesis "splitting,

and then stemming is best" (treating splitting as a no-op for English, French, Italian, and Spanish), is not contradicted. Put differently, there is no language for which the best performing run significantly improves over the "split, and stem" run. In conclusion, the hypothesis that 4-gramming is the best strategy is refuted for Spanish, but the hypothesis that splitting and then stemming is the best strategy is not refuted by our experiments.

Acknowledgements

We are extremely grateful to three anonymous referees for their extensive and insightful comments. We want to thank Carol Peters for editorial help and patience.

Vera Hollink was supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001. Jaap Kamps was supported by NWO under project number 400-20-036. Christof Monz was supported by the Physical Sciences Council with financial support from NWO, project 612-13-001. Maarten de Rijke was supported by grants from NWO, under project numbers 612-13-001, 365-20-005, 612.069.006, 612.000.106, 220-80-001, and 612.000.207.

References

Airio, E., H. Keskustalo, T. Hedlund, and A. Pirkola (2002), 'Utaclir @ CLEF 2002 – Towards a Uniform Translation Process Model'. In (Peters, 2002), pp. 51–58.

Amati, G., C. Carpinetto, and G. Romano (2002), 'Italian monolingual retrieval with PROSIT'. In (Peters, 2002), pp. 145–152.

Bacchin, M., N. Ferro, and M. Melucci (2002), 'University of Padua at CLEF-2002: Experiments to Evaluate a Statistical Stemming Procedure'. In (Peters, 2002), pp. 161–168.

Bell, C. and K. P. Jones (1979), 'Toward everyday language information retrieval systems via minicomputers'. *Journal of the American Society for Information Science* **30**, 334–338.

Braschler, M. and B. Ripplinger (2003), 'Stemming and Decompounding for German Text Retrieval'. In: *Advances in Information Retrieval, 25th BCS-IRSG European Colloquium on IR Research (ECIR)*. pp. 177–192.

Buckley, C., A. Singhal, and M. Mitra (1995), 'New retrieval approaches using SMART: TREC-4'. In (Harman, 1995b), pp. 25–48. NIST Special Publication 500-225.

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

Chen, A. (2002), 'Cross-language retrieval experiments at CLEF-2002'. In (Peters, 2002), pp. 5–20.

CLEF-Neuchâtel (2003), 'CLEF Resources at the University of Neuchâtel'. `http://www.unine.ch/info/clef` (visited February 1, 2003).

Damashek, M. (1995), 'Gauging Similarity via N-Grams: Language Independent Categorization of Text'. *Science* **267**, 843–848.

Davison, A. and D. Hinkley (1997), *Bootstrap Methods and Their Application*. Cambridge University Press.

De Heer, T. (1982), 'The application of the concept of homeosemy to natural language information retrieval'. *Information Processing & Management* **18**(5), 229–236.

Demske, U. (1995), 'Word vs. Phrase Structure: The Rise of Genitive Compounds in German'. *ZAS Papers in Linguistics* **3**, 1–28.

Efron, B. (1979), 'Bootstrap methods: Another look at the jackknife'. *Annals of Statistics* **7**(1), 1–26.

Fagan, J. (1987), 'Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods'. Ph.D. thesis, Department of Computer Science, Cornell University.

Figuerola, C. G., R. Gómez, A. F. Z. Rodríguez, and J. L. A. Berrocal (2002), 'Spanish monolingual track: The impact of stemming on retrieval'. In (Peters et al., 2002), pp. 253–261, Springer.

Frakes, W. B. (1992), 'Stemming algorithms'. In: W. B. Frakes and R. Baeza-Yates (eds.): *Information Retrieval, Data Structures and Algorithms*. Prentice-Hall, pp. 131–160.

Harman, D. K. (1991), 'How effective is suffixing'. *Journal of the American Society for Information Science* **42**(1), 7–15.

Harman, D. K. (1994), 'Overview of the Second Text REtrieval Conference (TREC-2)'. In: D. K. Harman (ed.): *Proceedings of the Second Text REtrieval Conference (TREC-2)*. pp. 1–20. NIST Special Publication 500-215.

Harman, D. K. (1995a), 'Overview of the Third Text REtrieval Conference (TREC-3)'. In (Harman, 1995b), pp. 1–20. NIST Special Publication 500-225.

Harman, D. K. (ed.) (1995b), 'Proceedings of the Third Text REtrieval Conference (TREC-3)'. NIST Special Publication 500-225.

Hedlund, T. (2002), 'Compounds in dictionary-based cross-language information retrieval'. *Information Research* **7**(2). Available at `http://InformationR.net/ir/7-2/paper128.html` (visited February 1, 2003).

Hedlund, T., H. Keskustalo, A. Pirkola, E. Airio, and K. Järvelin (2002), 'Utaclir @ CLEF 2001 – Effects of Compound Splitting and N-Gram Techniques'. In (Peters et al., 2002), pp. 118–136, Springer.

Hull, D. (1996), 'Stemming algorithms: a case study for detailed evaluation'. *Journal of the American Society for Information Sience* **47**(1), 70–84.

Jelinek, F. (1990), 'Self-organized language modeling for speech recognition'. In: A. Waibel and K.-F. Lee (eds.): *Readings in Speech Recognition*. Morgan Kaufmann, pp. 450–506.

Josefsson, G. (1997), *On the principles of word formation in Swedish*. Lund University Press, Lund.

Jurafsky, D. and J. H. Martin (2000), *Speech and Language Processing*. Prentice-Hall.

Koehn, P. and K. Knight (2003), 'Empirical Methods for Compound Splitting'. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Kotamarti, U. and A. L. Tharp (1990), 'Accelerated text searching through signature trees'. *Journal of the American Society for Information Science* **41**, 79–86.

Kraaij, W. and R. Pohlmann (1996), 'Viewing stemming as recall enhancement'. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 40–48.

Kraaij, W. and R. Pohlmann (1998), 'Comparing the effect of syntactic vs. statistical phrase index strategies for Dutch'. In: *Proceedings ECDL'98*. pp. 605–617.

Krott, A., R. H. Baayen, and R. Schreuder (2001), 'Analogy in morphology: modelling the choice of linking morpheme in Dutch'. *Linguistics* **39**, 51–93.

Krovetz, R. (1993), 'Viewing morphology as an inference process'. In: *Proceedings SIGIR'93*. pp. 191–202.

Matthews, P. H. (1991), *Morphology*. Cambridge University Press.

Mayfield, J. and P. McNamee (1999), 'Indexing using both n-grams and words'. In (Voorhees and Harman, 1999), pp. 419–424. NIST Special Publication 500-242.

McNamee, P. and J. Mayfield (2002a), 'JHU/APL Experiments at CLEF: Translation resources and score normalization'. In (Peters et al., 2002), pp. 193–208, Springer.

McNamee, P. and J. Mayfield (2002b), 'Scalable Multilingual Information Access'. In (Peters, 2002), pp. 133–140.

Monz, C. and M. de Rijke (2002), 'Shallow morphological analysis in monolingual retrieval for Dutch, German,and Italian'. In (Peters et al., 2002), pp. 262–277, Springer.

Monz, C., M. de Rijke, J. Kamps, W. van Hage, and V. Hollink (2002), 'The FlexIR information retrieval system'. Manual, Language & Inference Technology Group, ILLC, U. of Amsterdam.

Mooney, C. and R. Duval (1993), *Bootstrapping: A Nonparametric Approach to Statistical Inference*, Sage Quantitative Applications in the Social Science Series No. 95. Sage Publications.

Moulinier, I., J. McCulloh, and E. Lund (2001), 'West Group at CLEF 2000: Non-English Monolingual retrieval'. In (Peters, 2001), pp. 253–260, Springer.

Peters, C. (2001), ed., 'Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000', Vol. 2069 of *LNCS*. Springer.

Peters, C. (2002), ed., 'Results of the CLEF 2002 Cross-Language System Evaluation Campaign'.

Peters, C. and M. Braschler (2001), 'Cross-language system evaluation: The CLEF campaigns'. *Journal of the American Society for Information Science and Technology* **52**(12), 1067–1072.

Peters, C., M. Braschler, J. Gonzalo, and M. Kluck (2002), eds., 'Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001', Vol. 2406 of *LNCS*. Springer.

Pirkola, A. (1999), 'Studies on Linguistic Problems and Methods in Text Retrieval'. Ph.D. thesis, University of Tampere.

Pirkola, A. (2001), 'Morphological typology of languages for IR'. *Journal of Documentation* **57**(3), 330–348.

Popovic, M. and P. Willett (1992), 'The effectiveness of stemming for natural-language to Slovene textual data'. *Journal of the American Society for Information Sience* **43**(5), 384–390.

Porter, M. (1980), 'An Algorithm for Suffix Stripping'. *Program* **14**(3), 130–137.

Rocchio, Jr., J. J. (1971), 'Relevance feedback in information retrieval'. In: G. Salton (ed.): *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs NJ, Chapt. 14, pp. 313–323.

Savoy, J. (1997), 'Statistical inference in retrieval effectiveness evaluation'. *Information Processing and Management* **33**(4), 495–512.
Savoy, J. (1999), 'A stemming procedure and stopword list for general French corpora'. *Journal of the American Society for Information Science* **50**(10), 944–952.
Savoy, J. (2002a), 'Report on CLEF-2001 Experiments: Effective combined query-translation approach'. In (Peters et al., 2002), pp. 27–43, Springer.
Savoy, J. (2002b), 'Report on CLEF-2002 Experiments: Combining multiple sources of evidence'. In (Peters, 2002), pp. 31–46.
Schmid, H. (1994), 'Probabilistic Part-of-Speech Tagging Using Decision Trees'. In: *Proceedings of International Conference on New Methods in Language Processing*.
Shannon, C. E. (1951), 'Prediction and entropy of printed English'. *The Bell System Technical Journal* **30**, 50–64.
Snowball Stemmers, `http://snowball.tartarus.org/` (visited February 1, 2003).
Strzalkowski, T. (1995), 'Natural Language Information Retrieval'. *Information Processing & Management* **31**(3), 397–417.
Tomlinson, S. (2002a), 'Experiments in 8 European Languages with Hummingbird SearchServer^{TM} at CLEF2002'. In (Peters, 2002), pp. 203–214.
Tomlinson, S. (2002b), 'Stemming Evaluated in 6 Languages by Hummingbird SearchServer^{TM} at CLEF2001'. In (Peters et al., 2002), pp. 278–287, Springer.
Ullman, J. R. (1977), 'Binary n-gram technique for automatic correction of substitution, deletion, insertion, and reversal errors in words'. *Computer Journal* **20**, 141–147.
Voorhees, E. M. and D. K. Harman (1998), 'Overview of the sixth Text REtrieval Conference (TREC-6)'. In: E. M. Voorhees and D. K. Harman (eds.): *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*. pp. 1–28. NIST Special Publication 500-240.
Voorhees, E. M. and D. K. Harman (1999), eds.,'Proceedings of the Seventh Text REtrieval Conference (TREC-7)'. NIST Special Publication 500-242.
Whaley, L. J. (1997), *Introduction to Typology: The Unity and Diversity of Language*. Sage Publications.
Wilbur, J. (1994), 'Non-parametric significance tests of retrieval performance comparisons'. *Journal of Information Science* **20**(4), 270–284.
Willet, P. (1979), 'Document retrieval experiments using indexing vocabularies of varying size. II. Hashing, truncation, digram and trigram encoding of index terms'. *Journal of Documentation* **35**, 296–305.
Wisniewski, J. L. (1987), 'Effective text compression with simultaneous digram and trigram encoding'. *Journal of Information Science* **13**, 159–164.
Womser-Hacker, C. (2002), 'Multilingual Topic Generation within the CLEF 2001 Experiments'. In (Peters et al., 2002), pp. 389–393, Springer.

Address for Offprints:
Maarten de Rijke
Language & Inference Technology Group
ILLC, U. of Amsterdam, Nieuwe Achtergracht 166
1018 WV Amsterdam, The Netherlands