21

Morphological Processing in Mono- and Cross-Lingual Information Retrieval

KALERVO JÄRVELIN AND ARI PIRKOLA

21.1 Introduction

Text-based information retrieval (IR) matches text-based representations of information needs to text-based representations of documents. Since both representations deal with natural language and have different sources and characteristics, their match rarely is perfect. IR therefore has to deal with several difficult problems: First, the request representing the information need often is vague and short thereby providing little evidence for the IR system about desired document features - suggesting document relevance. Secondly, the request wording may be different from that of relevant documents due to many natural language features, e.g., synonymy and inflection. Thirdly, even useless documents may contain many request words.

In this paper we shall focus on text-based representations of documents and requests, and their matching. Our specific focus will be morphological processing of documents and requests in order to derive representations that better support document - request matching. This study is motivated by the morphological variability of natural languages. While much of IR research deals with English, English is morphologically fairly simple. Therefore findings in the English IR context do not necessarily apply IR in other languages with different morphological characteristics. We shall therefore contrast findings in English IR with findings in Finnish IR. Finnish is morphologically much more complex than English. For example, English nouns have singular and plural and two cases while Finnish nouns in principle may have over 2,000 different inflectional forms (Koskenniemi 1983).

Inquiries into Words, Constraints and Contexts.
Antti Arppe et al. (Eds.)

In IR multiple approaches have been adopted in the handling of morphological variation of words. The baseline is token-based indexing and retrieval - i.e. plain text words are used as such for the representation of documents and requests - with obvious problems when the request word tokens do not match the document word tokens. A simple way to alleviate the problems is to leave the document representation intact, but use a *truncation operation* on the request words to match in the index all document words having the same initial characters. In large text databases truncation tends to match too many words turning queries unmanageably long. Linguistic morphological processing can be alleviated also by approximate string matching, e.g., by *n-gramming* (McNamee and Mayfield 2004).

Among linguistically informed approaches, one possibility is to apply stemming on both request and document words thereby removing much of inflectional variation (Porter 1980). Stemming may however conflate fairly remote words to common stems turning them unspecific or fail to identify the common stem of some words in complex cases. An elaborated approach combining stemming and truncation is *stem generation* (Kettunen et al. 2005). Here several distinct inflectional stems are generated for one lemma before matching the token-based index - yielding shorter and more specific queries. A further possibility consists of the production of all inflectional word forms (Arppe 1996) for request words. However, in morphologically complex languages this tends to lead to excessively long queries. The final approach is *lemmatization* where the lemma of each document and request word is automatically identified and request word lemmas are compared to the lemma-based index.

Lemmatization would be the ideal approach for handling morphology in IR if not for two problems: word form ambiguity and out-of-vocabulary (OOV) words. Words are often ambiguous but may be disambiguated. However, most IR studies on disambiguation have reported no or minor improvements in retrieval performance (Krovetz and Croft 1992; Sanderson 1994). Lemmatizers often cannot handle OOV words (correctly). Often such words are proper names, which tend to be significant words in requests. Their incorrect treatment thus leads to severe loss in IR performance.

In this paper we accept lemmatization as the gold standard for morphological processing in IR and compare the plain words baseline and the morphologically simpler approaches to lemmatization w.r.t. IR performance. A prominent approach in lemmatization is the Two-level Morphology developed by Kimmo Koskenniemi (1983) and implemented in several lemmatization programs for several languages - e.g., FINTWOL, GERTWOL, ENGTWOL, SWETWOL. Riitta Alkula (2000; 2001) conducted the seminal experiments with Finnish morphological processing in IR and the TWOL software (among others). Her studies were performed in the Boolean exact-match retrieval environment. In this paper we shall focus however solely on experiments in best-match IR environments.

We shall look at the following research questions:

- Monolingual IR test condition:
    - What is the relative IR performance in Finnish IR of plain words, stemming, and inflectional stem generation w.r.t. lemmatization by FINTWOL.
    - Regarding FINTWOL: what is the relative IR performance of FINTWOL when compounds are split and they are kept intact.
    - What is the relative IR performance in English IR of plain words and stemming w.r.t. lemmatization by ENGTWOL.
- Cross-lingual IR test condition, keeping English as a source language, Finnish, German, and Swedish as target languages:
    - What is the relative retrieval performance in cross-language IR of stemming w.r.t. lemmatization by TWOL.
    - Regarding TWOL: what is the relative IR performance of TWOL when compounds are split and they are kept intact.

We shall review recent empirical findings produced at the University of Tampere (Airio 2005; Kettunen et al. 2005; Kettunen 2005).

This paper is organized as follows. Chapter 2 first discusses morphological differences between Finnish and English, then presents findings on morphological normalization in IR and then discusses morphological processing in cross-language IR. Chapter 3 presents the test settings and Chapter 4 the results. Chapter 5 contains discussion and conclusions.

21.2 Morphological Processing in IR

21.2.1 Morphological Differences between Finnish and English

Morphology studies word structure and formation and consists of inflectional morphology and derivational morphology. The former focuses on the formation of inflectional forms from lexemes. The latter is concerned with the derivation of new words from other words or roots. English and Chinese have a simple morphology whereas many other languages, e.g., Germanic languages or such languages as Finnish are morphologically more complex.

The *Finnish language* is a very inflectional and compound rich language. If Finnish text words are stored in their inflected forms in the database index, this results in clearly greater space requirements for Finnish text compared to that of English texts of corresponding length. For example, Finnish has more case endings than is usual in Indo-European languages. Finnish case endings correspond to prepositions or postpositions in other languages (cf. Finnish *auto/ssa, auto/sta, auto/on, auto/lla* and English in the car, out of the car, into the car, by car). There are 15 cases, while English has only two (Karlsson 1987).

In Finnish, several layers of endings may be affixed to word stems, indicating number, case, possession, modality, tense, person, and other morphological characteristics. This results in an enormous number of possible distinct word forms: a noun may have some 2,000 forms, an adjective 6,000, and a verb 12,000 forms. Moreover, these figures do not include the effect of derivation, which increases the figures roughly by a factor of 10 (Koskenniemi 1985). Consonant gradation makes the inflection even more complicated, as the stem of a word may alter when certain types of endings are attached to it. For example, the word *laki* (law) has in practice four inflected stems: *laki*-, *lake*-, *lai*-, and *lae*-. The common root of the stems consists of only two characters, which renders it inappropriate as a search key.

Several languages, Germanic and Finno-Ugrian languages included, are rich in compounds in contrast to English, which is phrase-oriented. For example, in Finnish, The Dictionary of Modern Standard Finnish contains some 200,000 entries, of which two-thirds are compound words (Koskenniemi 1983). For example the English phrase 'Turnover Tax Bureau' is *liike|vaihto|vero|toimisto* in Finnish (word boundaries here marked by '|'). In Finnish, compounding results in a problem of retrieving the second or later elements of compounds, for example *verotoimisto* (tax bureau), if the searcher is not able to recall all possible first components.

The fairly simple morphology of English suggests that the costs of morphological processing in IR are low. One may dispense with the morphological processing and still achieve good results. However, stemming has been shown to be useful in English IR (Section 2.2). In contrast to English, the complex morphology of Finnish suggests that simple morphological methods may not be sufficient, but lemmatization or some other sophisticated method is required to achieve the best possible results.

21.2.2 Previous Research

Stemming has been the most widely applied morphological technique in IR. With stemming, the searcher does not need to worry about the correct truncation point of search keys. Stemming also reduces the total number of distinct index entries. Further, stemming causes query expansion by bringing word variants, derivations included, together (see, e.g. Alkula 2001; Krovetz 1993; Pirkola 2001). Some early research results with English collections questioned the effectiveness of stemming (Harman 1991). Later results by, e.g. Krovetz (1993) and Hull (1996) found stemming useful especially when long enough retrieved sets of documents are analyzed. Hull also found out that stemming is always useful with short queries. With short queries and short documents, a derivational stemmer is most useful, but with longer ones the derivational stemmer brings in more non-relevant documents. Stemming increases search key ambiguity and greedy stemming may be counterproductive. With long queries and documents, relevant material can be identified with conservative stemming. In languages other than English, stemmers have been even more successful than in English text retrieval - e.g., in Slovenian (Popovic and Willett 1992), French (Savoy 1999), Modern Greek (Kalamboukis 1995), and Arabic (Abu-Salem et al. 1999).

The benefits of *lemmatization* are the same as in stemming. In addition, when basic word forms are used, the searcher may match an exact search key to an exact index key. Such accuracy is not possible with truncated, ambiguous stems. Homographic word forms cause ambiguity (and precision) problems - this may also occur with inflectional word forms (Alkula 2001). Another problem is owing to words that cannot be lemmatized, e.g. foreign proper names, because the lemmatizer's dictionary does not contain them. Such problem words need special handling.

Compound words may be split into their components in lemmatization. When indexing a text collection, both compounds and their components may be recorded in the database index thus enabling retrieval through all combinations of compound components. Recent findings suggest that lemmatization with compound splitting improves retrieval performance in Boolean (Alkula 2001) and best-match retrieval (Kunttu 2003). Their most important effects, however, may be the cognitive simplification of query formulation. The searcher is greatly relieved if she need not consider potential expressions like "Verkehrswegeplanungsbeschleunigungsgesetzveränderungsentwurf"[1] when interested in legislation on road planning.

21.2.3 Morphological Processing in Cross-Language Retrieval

Cross-language information retrieval (CLIR) refers to an information retrieval task where the language of queries is other than that of the retrieved documents. Different approaches to cross-language retrieval are discussed in Oard and Diekema (1998). In *dictionary-based CLIR* a standard method is to replace each source language key by all of its target language equivalents included in a translation dictionary (Pirkola 1998; Pirkola et al. 2001). The main problems associated with dictionary-based CLIR are (1) OOV words, (2) morphological processing of keys, (3) phrase identification and translation, and (4) lexical ambiguity in source and target languages. Here our focus is on the problem (2).

Morphological processing is needed in three situations in dictionary- based CLIR: processing of source language search keys for dictionary look- up, processing of inflected dictionary output words, and processing of database

[1] In German - a proposal for changing the law on speeding up the planning of roads - here no compounds.

index keys. Lemmatization is often used in the first stage to facilitate matching of source keys with dictionary headwords (in base forms) also in the case of inflected search keys. Alternatively, source keys and headwords can be conflated into the same form by a stemmer (Davis and Ogden 1997). One problem related to stemming is that different headwords may be conflated into the same form. In our experiments source language (English) keys were lemmatized by ENGTWOL for dictionary look-up (Section 3).

If index keys are stemmed, dictionary output words also have to be stemmed (Davis and Ogden 1997). In the case of the lemmatized index keys, the lemmatization of the output words does not seem necessary, but might be useful since some dictionary output words may be in inflected forms, e.g., some phrase component words (Hedlund et al. 2001).

Regarding word inflection CLIR effectiveness depends to a great extent on the morphological processing of index keys. This issue is the focus of our cross-language IR experiments. We explore the matching of target language queries against different types of indexes as described in Section 3.

21.3 Test Data and Settings

The tests of this study were conducted in the Information Retrieval Laboratory of the Department of Information Studies, University of Tampere. Actual searches were conducted with a probabilistic partial match system, InQuery, version 3.1 (Callan et al. 1992, Broglio et al. 1995) in two different testing environments called *Environment One* and *Environment Two*. In Environment One we studied monolingual Finnish IR, and in Environment Two monolingual Finnish and English IR and cross-lingual English to Finnish, German, and Swedish IR. Next we describe the two environments.

The test collection of *Environment One*, TUTK, contains a full text database of newspaper articles published in three Finnish newspapers in 1988 - 1992 (Sormunen 2000). The database consists of 53,893 articles. The articles of the database are fairly short on average. Typical text paragraphs are two or three sentences in length. The topic set consists of 30 topics. Topics are long: the mean length of the original topics is 17.4 words. The relevance of documents is assessed on a four-level scale. In this study we used a binary relevance scale and combined the documents on the levels 2 and 3 into a class of relevant documents, and the documents on the levels 0 and 1 into a class of irrelevant documents.

We used the following morphological programs: FINTWOL (for lemmatization), MaxStemma (for stem generation), and Finnish Snowball stemmer which is freely available on the Web (http://snowball.tartarus.org). MaxStemma was implemented by Kimmo Kettunen in early 1990’s. Its original version is described in more detail in (Kettunen 1991a, 1991b).

The MaxStemma stem generator works in the following fashion: given the base word form (nominative singular for nouns), it produces all the differing inflectional stems of the words. Depending on the input noun, 1 - 5 different stems (including the base form) are produced for a noun. For instance, if the input word is *kissa* ('cat'), the program would generate the following inflectional stems for the word: *kissa, kissoi, kissoj*.

The Snowball stemmer returns stems out of inflected word forms. Snowball is a Lovins' style stemmer that strips off suffixes from the input word according to a suffix list and set of rules and returns stems for the words (Frakes 1992, Porter 2001).

The experiments conducted in *Environment Two* used CLEF (Cross Language Evaluation Forum; http://clef.isti.cnr.it/) data and the UTACLIR query translation system of the University of Tampere. We used CLEF 2003 Finnish, German, Swedish, and English test collections, test topics and relevance assessments. There are 60 CLEF 2003 topics, translated into all the CLEF languages, including the present test languages.

The UTACLIR system utilizes several external language resources (translation dictionaries, stemmers and lemmatizers, and stop word lists) in processing queries for retrieval (Airio et al. 2003). Word processing in UTACLIR proceeds as follows. First the topic words are lemmatized. The existence of a lemmatizer for the source language is vital, because stemmed words are not translatable. The lemmatizer produces one or more basic forms for a token. After normalization, stop-words are removed, and non-stop words are translated. If translation equivalents are found, they are normalized utilizing a lemmatizer or a stemmer, depending on the target index. If translation equivalents are not found, they are identified in the target index by n-gramming the source word. Queries are structured utilizing InQuery's synonym operator: the target words derived from the same source word are grouped into the same synonym group (Pirkola 1998).

For comparing performance of different word normalization tools and decompounding in monolingual and cross-lingual IR different kinds of indexes were created (inflected, stemmed, lemmatized with decompounding, and lemmatized without decompounding). As normalization tools we used TWOLs and Snowball stemmers for Finnish, German, Swedish, and English. Altogether 16 test runs were performed, out of which 7 were monolingual and 9 cross- lingual.

The approach in the *monolingual stemmed runs* was to stem the topic words, and perform retrieval in the stemmed index. In the *monolingual lemmatized runs*, the topic words were lemmatized, and retrieval was performed in the lemmatized indexes. For Finnish there were two lemmatized indexes (compounds were and were not split) and for English one (compounds were not split). In the *inflected word form* runs, topic words were added as such into the query, and retrieval was performed in the inflected word form index.

For the cross-language IR tests two lemmatized runs (one in the decompounded index and one in the index without decompounding) and one stemmed run were performed for all the language pairs.

21.4 Results

The results of the monolingual Finnish IR experiments in Environment One are presented in Table 1. The results of the monolingual Finnish and English IR experiments conducted in Environment Two are presented in Table 2. The results of the cross-lingual IR experiments are shown in Table 3.

From Table 1 one may see that FINTWOL (lemmatization) performs slightly better than MaxStemma (stem generation). The performance of Snowball (stemming) is clearly below the former. The worst performance was achieved for Plain Words. On the average Plain Words achieved 54.0 % of FINTWOL's performance.

In Environment Two for Finnish monolingual runs the best result was achieved with the decompounded lemmatized index, the next best with the stemmed index, and the worst with the inflected index (Table 2). The results of English monolingual runs are in line with the majority of the earlier results: no statistically significant differences could be found between the inflected run and the normalized runs.

Table 3 shows average precision for the cross-lingual runs. Retrieval in the lemmatized indexes where compounds were split performed best in all the cross-lingual runs. In English-Finnish and English-German, the next best was the run in the lemmatized index without decompounding, and the stemmed run achieved the worst result. In English-Finnish, the stemmed run performed clearly worse than both of the lemmatized runs: the result was 41.4 % worse than that of the run in the lemmatized decompounded index.

In English-Swedish and in English-German, the differences between the two lemmatized runs were statistically significant by the Wilcoxon signed ranks test at the 0.01 level, but differences between the run in the lemmatized index without decompounding and stemmed run were not significant. In English-Finnish the situation is opposite: the differences between the two lemmatized runs were not statistically significant, but between the run in the lemmatized index without decompounding and stemmed run they were significant. All the differences between the cross-lingual stemmed runs and the runs in the lemmatized decompounded indexes were statistically significant at the 0.01 level.

Table 1. The performance of monolingual Finnish runs in Environment One

Morphological tool	Average Precision %	Change % w.r.t FINTWOL
1. FINTWOL	35.0	
2. MaxStemma	34.2	-2.3
3. Snowball	27.7	-20.9
4. Plain Words	18.9	-46.0

Table 2. The performance of monolingual Finnish and English runs in Environment Two

Language	Index type	Average precision %	Change % w.r.t 1a or 2a
1a. Finnish	Lemmatized, split	50.5	
1b. Finnish	Lemmatized, no split	47.0	-7.0
1c. Finnish	Stemmed	48.5	-4.0
1d. Finnish	Inflected	31.0	-38.6
2a. English	Lemmatized, no split	45.6	
2b. English	Stemmed	46.3	+1.5
2c. English	Inflected	43.4	-4.8

Table 3. The performance of cross-language runs with English as the source language

Target Language	Index type	Average precision %	Change % w.r.t 1a, 2a or 3a
1a. Finnish	Lemmatized, split	35.5	
1b. Finnish	Lemmatized, no split	29.0	-18.3
1c. Finnish	Stemmed	20.8	-41.4
2a. Swedish	Lemmatized, split	27.1	
2b. Swedish	Lemmatized, no split	17.4	-35.8
2c. Swedish	Stemmed	19.0	-29.9
3a. German	Lemmatized, split	31.0	
3b. German	Lemmatized, no split	26.4	-14.8
3c. German	Stemmed	25.7	-17.1

21.5 Discussion and conclusions

In this paper we focused on the question of the effectiveness of morphological processing in mono and cross-lingual IR. In our Monolingual IR tests we found out that, in Finnish IR, lemmatization by FINTWOL outperforms other approaches, in particular plain words and stemming, while inflectional stem generation approaches the performance of lemmatization. Their difference in performance is not significant. However, in the latter approach, the index must be harvested for full words matching the generated stems. Thus queries tend to become unmanageably long. Kettunen (2005) has however found that by extending the inflectional stems by regular expressions, query length can be reduced dramatically with only a minor penalty in performance.

In the second set of monolingual tests we found that the performance of lemmatization by FINTWOL when compounds were split vs. kept intact, splitting compounds clearly improved performance. Interestingly, in the test collection used, stemming by Snowball approached lemmatization in performance. In the English monolingual tests, stemming was found better than lemmatization by ENGTWOL. Simpler morphology and the lack of compound words in English compared to Finnish seem to explain the finding. However, another test collection might yield slightly different results.

In our cross-lingual IR tests, *English* was the source language, and *Finnish*, *German*, and *Swedish* served as target languages. In all findings, lemmatization and splitting compounds by TWOL clearly outperforms other approaches. This further confirms the importance of handling compound words properly in compound-rich languages. The relative performance of lemmatization without splitting compounds vs. stemming gave mixed results, which may in part be explained by the quality of stemmers.

In summary, lemmatization and splitting compounds in morphologically complex languages seem to consistently provide the best performance. The down sides are that this approach requires large dictionaries, which need to be updated, and techniques for handling the unavoidable and important out-of-vocabulary words. Automatic stem generation seems to be a much lighter-weight approach delivering competitive performance, at least in the case of Finnish. However, in this approach, after harvesting full index words matching the generated stems, queries tend to become long. This may be critical for efficiency in some IR environments. Further research in morphological processing for IR is therefore in order.

References

Abu-Salem, H., Al-Omari, M. and Evens, M.W. 1999. Stemming methodologies over individual query words for an Arabic information retrieval system. *Journal of the American Society for Information Science* 50(6): 524-529.

Airio, E., Keskustalo, H., Hedlund T. and Pirkola, A. 2003. Multilingual experiments of UTA at CLEF 2003: The impact of different merging strategies and word normalizing tools. Peters, C. and Borri, F. (eds.) *Results of the CLEF 2003 Evaluation Campaign, Cross-Language Evaluation Forum, Italy*, pp. 13 - 18.

Airio, E. 2005. Word normalization and decompounding in mono- and cross- lingual IR. *Information Retrieval*. To appear.

Alkula, R. 2000. *Merkkijonoista suomen kielen sanoiksi*. Acta Universitatis Tamperensis 763, Available at: http://acta.uta.fi/pdf/951-44-4886-3.pdf.

Alkula, R. 2001. From plain character strings to meaningful words: producing better full text databases for inflectional and compounding languages with morphological analysis software. *Information Retrieval* 4: 195 - 208.

Arppe, A. 1996. Information explosition and the use of linguistic tools in Finland. *Kieli ja Tietokone, AFinLAn Vuosikirja 1996*. Suomen Soveltavan Kielitieteen Yhdistyksen Julkaisuja, 54 (= AFinLA Series, 54), pp. 7-32.

Broglio, J., Callan, J., Croft, B. and Nachbar, D. 1995. Document retrieval and routing using the INQUERY system. *Proceedings of the Third Text Retrieval Conference (TREC-3), Gaithesburg*, MD: National Institute of Standards and Technology, special publication 500-225, pp. 29 - 38.

Callan, J., Croft, B. and Harding, S. 1992. The INQUERY retrieval system. *Proceedings of the Third International Conference on Databases and Expert Systems Applications*, Berlin: Springer Verlag, pp. 78 - 84.

Davis, M. and Ogden, W. 1997. QUILT: Implementing a large-scale cross- language text retrieval system. *Proceedings of the 20th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, Phidadelphia, PA, USA, pp. 92-98.

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

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

Hedlund, T., Keskustalo, H., Pirkola, A., Airio, E., and Järvelin, K. 2001. *UTA-CLIR @ CLEF 2001. Working Notes for CLEF 2001 Workshop*. Available at: http://www.ercim.org/publication/ws-proceedings/CLEF2/hedlund.pdf

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

Kalamboukis, T.Z. 1995. Suffix stripping with modern Greek. *Program* 29(3): 313-321.

Karlsson, F. 1987. *A Finnish grammar*. Porvoo: WSOY.

Kettunen, K., Kunttu, T. and Järvelin, K. 2005. To stem or lemmatize a highly inflectional language in probabilistic IR environment. *Journal of Documentation*. To appear.

Kettunen, K. 1991a. Doing the stem generation with Stemma. J. Niemi (ed.), *Papers from the Eighteenth Finnish Conference of Linguistics*. Kielitieteellisiä tutkimuksia, Joensuun yliopisto, N:o 24, pp. 80 - 97.

Kettunen, K. 1991b. Stemma, a robust noun stem generator for Finnish. *Humanistiske Data* 1: 26 - 31.

Kettunen, K. 2005. Developing an automatic linguistic truncation operatorfor best-match retrieval in inflected word form text database indices. Submitted to Information Retrieval.

Koskenniemi, K. 1983. *Two-level morphology: a general computational model for word-form recognition and production*. Publications of the Department of General linguistics, University of Helsinki. No. 11.

Koskenniemi, K. 1985. FINSTEMS: a module for information retrieval. Karlsson, F. (ed.), *Computational morphosyntax. Report on research 1981 - 84*. Publications of the Department of General linguistics, University of Helsinki. No. 13., pp. 81 - 92.

Krovetz, R. and Croft, W.B. 1992. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems* 10(2): 115-141.

Krovetz, R. 1993. Viewing morphology as an inference process. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Pittsburg, PA, pp. 191-202.

Kunttu, T. 2003. *Perus- ja taivutusmuotohakemiston tuloksellisuus todennäköisyyksiin perustuvassa tiedonhakujärjestelmässä*. Informaatiotutkimuksen pro gradu -tutkielma. Informaatiotutkimuksen laitos, Tampereen yliopisto. (M.Sc. Thesis, University of Tampere, Dept. of Information Studies).

McNamee, P. and Mayfield, J. 2004. Character n-gram tokenization for European language text retrieval. *Information Retrieval* 7(1-2): 73-97.

Oard, D. and Diekema, A. 1998. Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)* 33: 223-256.

Pirkola, A. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. *Proceedings of the 21st Annual International ACM Sigir Conference on Research and Development in Information Retrieval*. Melbourne, Australia, pp. 55-63.

Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K. 2001. Dictionary- based cross-language information retrieval: problems, methods, and research findings. *Information Retrieval* 4(3/4): 209-230.

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

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

Porter, M.F. 1980. An algorithm for suffix stripping. *Program* 14: 130-137.

Porter, M. 2001. *Snowball: A language for stemming algorithms*. Available at: http://snowball.tartarus.org/texts/introduction.html

Sanderson, M. 1994. Word sense disambiguation and information retrieval. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland*, pp. 142-151.

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.

Sormunen, E. 2000. *A method for measuring wide range performance of Boolean queries in full-text databases*. Acta Universitatis Tamperensis 748. Tampere: University of Tampere.