Searching Strategies for the Hungarian Language

Jacques Savoy

Computer Science Dept., University of Neuchatel,
Rue Emile Argand 11, 2009 Neuchâtel, Switzerland
Jacques.Savoy@unine.ch

Abstract. This paper reports on the underlying IR problems encountered when dealing with the complex morphology and compound constructions found in the Hungarian language. It describes evaluations carried out on two general stemming strategies for this language, and also demonstrates that a light stemming approach could be quite effective. Based on searches done on the CLEF test collection, we find that a more aggressive suffix-stripping approach may produce better MAP. When compared to an IR scheme without stemming or one based on only a light stemmer, we find the differences to be statistically significant. When compared with probabilistic, vector-space and language models, we find that the Okapi model results in the best retrieval effectiveness. The resulting MAP is found to be about 35% better than the classical *tf idf* approach, particularly for very short requests. Finally, we demonstrate that applying an automatic decompounding procedure for both queries and documents significantly improves IR performance (+10%), compared to word-based indexing strategies.

1 Introduction

The majority of European languages belong to the Indo-European family and thus they share various syntactic features as well as words in their basic lexicon, as least from a phonological point of view. The Hungarian, Finnish and Basque languages however have fewer characteristics in common with these languages. The English lexicon for example has only a few words with Hungarian origins (e.g., saber, paprika, goulash), while the Hungarian lexicon contains many more words borrowed from the English language (e.g., modern, interview, sport, jury, pedigree, computer, internet).

During the first CLEF (www.clef-campaign.org) evaluation campaigns (Peters *et al.*, 2006), the emphasis was placed on the Roman (e.g., French, Italian, and Spanish) and Germanic (e.g, German, Dutch, and Swedish) family of languages (Sproat, 1992). From an IR point of view these languages are closer to the English while Hungarian represents a special case, especially given its more complex morphology and agglutinative aspects. Moreover, only a few IR experiments have been conducted with the Hungarian language. In fact, not until 2005 did the CLEF evaluation forum include this language in one of its tracks, when a real and reasonably large test collection respecting the required international standards was developed (Harman, 2005), (Buckley & Voorhees, 2005) (Gordon & Pathak, 1999). The main objective of our paper is therefore to carry out studies on the Hungarian language. This paper is divided as follows. Section 2 presents the context and related works, while Section 3 depicts the main characteristics of the test collection. Section 4 briefly describes the IR models used during our experiments. Section 5 evaluates three stemming approaches together with a comparison of the retrieval effectiveness of word-based schemes, and those where words are automatic decompounded. The main findings of this paper are summarized in Section 6.

2 Context and Related Work

In order to define pertinent matches between search keywords and documents, very frequently occurring terms in any given language are usually removed. These words tend not to have clear and important meanings (e.g., the, in, but, some). For the Hungarian language and following the guidelines suggested by Fox (1990), we first created a list of the top 200 most frequently occurring words found in the corpus, from which certain words were removed (e.g., police, minister, president, Magyar). To this list we manually added articles (e.g., the = “a”, “az”, this = “ez”, “e”, these = ”ezek”, …), pronouns (e.g., I = “én”, you = “te”, they = “ők”, etc.), possessive pronouns (e.g., my = “enyém”, “enyémek”, …), prepositions[1] (e.g., under = “alá”, “alatta”, “alóla” , …), conjunctions (e.g., and = “és”, but = “ám” , …), or very frequently occurring verb forms (e.g., to be = “lenni”, are = “vannak”, has = “neki van”, …). The final stopword list we suggest contained 761 Hungarian terms, a greater number than those usually proposed for the English language (e.g., Fox (1990) suggested 421 terms, while the SMART system included 571. Our list[2] is longer because a given pronouns or determinants may occur in numerous forms, reflecting the fact that Hungarian grammar comprises several grammatical cases.

On the other hand it must be recognized these lists were established on the basis of certain arbitrary decisions (Savoy, 1999), even though commercial information systems tend to adopt a very conservative approach with only a few stopwords. The DIALOG system for example uses only 9 items (namely “an,” “and,” “by,” “for,” “from,” “of,” “the,” “to,” and “with”) (Harter, 1986). Another example is the WIN™ system, which ignores the single word (“the”) when indexing documents, but a larger stopword list may be used when analyzing the request (Moulinier, 2004).

Once high-frequency words were removed, an indexing procedure generally applied a stemming algorithm in order to conflate word variants into the same stem or root. In developing such a procedure, we may define a light stemming approach whereby the stemmer removes only inflectional suffixes related to number (singular vs. plural), gender (masculine, feminine) or representing grammatical cases (e.g. in Latin “rosae” and “rosarum” are related to the nominative form “rosa”). We could also remove derivational suffixes, usually those used to form new words belonging to another part of speech (e.g., power, powerful, powerlessly).

In the rest of this section, we report on the main morphological difficulties characteristic of the Hungarian language (Section 2.1) and describe how we could generally

[1] More precisely postpositions because they appear after the nouns they qualify.

[2] The stopword list and stemmers are freely available at http://www.unine.ch/info/clef/

derive a stemmer for those languages having more complex morphologies (Section 2.2). Finally, we will explain how compound words have a significant impact on retrieval effectiveness.

2.1 Main Aspects of Hungarian Morphology

The Hungarian language shares certain similarities with the Finnish language. Although both languages do not strictly belong to the same family, they can be viewed as cousins. Comparable to the Latin or the German languages, Hungarian is characterized by many grammatical cases (23 in total, although some are limited to a set of nouns or appear only in fixed and predefined forms). Each Hungarian case has its own unambiguous suffix. For example, the noun "house" ("ház") or fire ("tűz") may appear as "házat" or "tüzet" (the accusative case, as in "(I see) the house / fire", with suffixes underlined), "házakat" or "tüzeket" (accusative plural, as in "(I see) the houses / the fires"). Grammatical cases are often denoted through adding a suffix to nouns, and also to names. The Hungarian name for the city of Paris is "Párizs", and thus we may encounter variant forms such as "Párizsban" ((to stay in) Paris), "Párizsba" ((to go into) Paris) or "Párizsból" ((to come from inside) Paris), with these forms corresponding to the English preposition "in", respectively "from". Three other grammatical cases correspond to the English preposition "over", and three other forms are related to the meaning of "near". From these examples, we usually find that English prepositions do not have a direct translation, but rather their meaning appears in a grammatical case and therefore in the corresponding suffix. The attachment of suffixes is not limited to geographic names. For example, with the proper Hungarian name "Péter", we may also found the form "Péteré" (Peter's), "Pétert" ((I see) Peter), "Péterrel" (with Peter), or "Erdősné" (the Erdos' wife).

The Hungarian suffixes may also be used in conjunction with possessive pronouns (my, their) as in "házamat" ("(I see) my house"), with the suffix '-(a)m' used to indicate the English pronoun "my". Thus, a suffix could represent four types of information; namely case, possessive pronoun, number (singular/plural), and the fact that a given noun possesses something (with the suffix "-é/éi-" as in "házéi" ("more things of the house")). Combining these suffixes may produce forms such as "házaimat" ("(I see) my houses") where the plural form is indicated by the letter "-i-", or "házaméban" ("in something of my house"), or "házakéiban" ("with things of the houses").

Finally, the morphological rules are not too strict and the inclusion of vowels is sometimes allowed in order to facilitate the pronunciation (e.g., in "házamat" = "ház" (house) + '-m' (my) + '-t' (accusative)). Similar agglutinative aspects may be found in other languages such as Turkish, where the noun "ev" (house) may take on the form "evler" (the houses), "evlerim" (my houses) and "evlerimde" (in my houses).

From an IR point of view, certain Hungarian linguistic aspects are easier to process. For example, a gender distinction (feminine / masculine / neutral) is not attached to a noun (as in English with she/he/it = "ő" or with the noun "ship"). Moreover, adjectives are mainly invariable as in "a szép virág" (the pretty flower) or "a szép virágok" (the beautiful flowers). The only exception is the plural form used with a copulative verb (e.g., the flowers are beautiful = "a virágok szépek").

2.2 Stemming Strategies

In the IR domain we usually assume that stemming is an effective means of enhancing retrieval efficiency by conflating several different word variants into a common form. Most stemming approaches achieve this through applying morphological rules for the language involved (e.g., see (Lovins, 1968) and (Porter, 1980) for the English language). In such cases suffix removal is also controlled through the adjunct of quantitative restrictions (e.g., '-ing' would be removed if the resulting stem had more than three letters as in "running", but not in "king") or qualitative restrictions (e.g., '-ize' would be removed if the resulting stem did not end with 'e' as in "seize"). Moreover, certain ad hoc spelling correction rules are applied to improve conflation accuracy (e.g., "running" gives "run" and not "runn"), due to certain irregular grammar rules, usually applied to facilitate easier pronunciation.

Such simple stemming procedures (algorithmic stemming) ignore word meanings and tend to make errors, usually due to over-stemming (e.g., "general" becomes "gener", and "organization" is reduced to "organ") or to under-stemming (e.g., with Porter's stemmer, the words "create" and "creation" do not conflate to the same root). For this reason the use of an on-line dictionary has been suggested as a means of obtaining better conflation (Krovetz, 1993).

Compared to other languages having more complex morphologies (Sproat, 1992), English is considered quite simple and the use of a dictionary to correct stemming procedures could be more helpful for those other languages such as French (Savoy, 1993). When a language has an even more complex morphology, deeper analysis could be required (e.g., for Finnish (Korenius *et al.*, 2004), or for Hungarian (Halácsy, 2006)), where lexical stemmers are clearly more elaborate and not always freely available (e.g., Xelda system at Xerox). They are more labor intensive and their implementation is complex. Moreover their use depends on a large lexicon and a complete grammar for the language involved. These application also requires more processing time and could thus be problematic, especially when document collections are very large and dynamic (e.g., within a commercial search engine on the Web). Additionally, lexical stemmers must be capable of handling unknown words such as geographical names, products, proper names or acronyms (out-of-vocabulary problems). Lexical stemmers thus cannot be viewed as error-free approaches. Finally, it must be recognized that when inspecting language usage and real corpora, the observed morphological variations are less extreme than those that might be imagined when inspecting the grammar. Kettunen & Airo (2006) indicate for example that in theory Finnish nouns have around 2,000 different forms, yet in actual collections the occurrence of most of these forms is rare. As a matter of fact in Finnish, 84 to 88% of the occurrences of inflected nouns are generated by only six out of a possible 14 cases.

While stemming schemes are normally designed to work with general texts, some may also be especially designed for a specific domain (e.g., in medicine) or a given document collection, such as that developed by Xu & Croft (1998), which used a corpus-based approach. This more closely reflects language usage (including word frequencies and other co-occurrence statistics), instead of a set of morphological rules in which the frequency of each rule (and therefore its underlying importance) is not precisely known.

In analyzing the IR stemming performance, Harman (1991) demonstrated that no statistically significant improvements could be obtained from applying any of three different stemming strategies, namely those of Lovins (1968), Porter (1980) as well as a basic stemming technique conflating singular and plural English word forms (and based on three rules). A query-by-query analysis revealed that stemming did indeed affect performance, even though the number of queries showing improvements was nearly equal to the number of queries resulting in decreased performance. Other studies (Hull, 1996), usually limited to one language (English), show that modest improvements can result from applying a stemmer. When compared with approaches that ignored stemming, differences were not always statistically significant.

It was also surprising to note that during the last CLEF evaluation campaigns (Peters *et al.*, 2006), participants suggested a limited number of stemmers and only attempted to compare a few of them. For example, when evaluating the two statistical stemmers used for five languages, Di Nunzio *et al.* (2004) showed that relative retrieval performances would vary for each of these languages. This means that any given stemming approach may work well for one language yet not for another. When compared to statistical stemmers, Porter's stemmers seem to work slightly better. For German, Braschler & Ripplinger (2004) showed that for short queries stemming may enhance mean average precision by 23%, compared to 11% for longer queries. Finally, Tomlinson (2004) evaluated the differences between Porter's stemmer and the lexical stemmer (in which stemming is based on a dictionary of the corresponding language and a more complex morphological analysis). Moreover, Tomlinson (2004) found that for the Finnish and German languages, the lexical stemmer tended to produce better results statistically, while for the Dutch, Russian, Spanish, French and English languages performance differences were small and insignificant. For the Swedish language, the algorithmic stemmer produced mean average precision that was statistically better than a lexical stemming approach.

2.3 Compound Words

Compound word construction (e.g., handgun, viewfinder) is another morphological characteristic that may have an impact on retrieval effectiveness. Most European languages involve some form of compound construction, indicated by a hyphen in some cases (e.g, in Hungarian "Közép-Európa" (Central Europe) or in French "porte-clefs" (key ring)) or by the suffix attached to the genitive case (e.g., in German with the "-s" suffix in "Lebensversicherungsgesellschaftsangestellter" = "Leben" (life) + "-s" + "Versicherung" (insurance) + "-s" + "Gesellschaft" (company) + "-s" + "Angestellter" (employee)).

In general however no "glue" is used to build compound forms from two or more words, such as in the English (viewpoint) or the German ("Bankangestelltenlohn" = "Bank" + "Angestellter" + "Lohn" (salary)). Such word composition is not limited to the Germanic family, however, for similar compound constructions are also found in Finnish, such as "rakkauskirje" = "rakkaus" (love) and "kirje" (letter) or "työviikko" = "työ" (work) and "viikko" (week); and in Italian, such as "capoufficio" (chief of the office), or "capofamiglia" (chief of the family).

In Hungarian, typical compound word formations would be “Magyarország” = “Magyar” (hungarian) + “ország” (country), or in “hétvégé” = “hét” (week / seven) + “vég” (end). The decompounding process may of course introduce errors by decompounding terms into semantically unrelated forms or into forms having other meanings. The word “breakfast” fox example may be split into the existing words “break” and “fast”, thus introducing other unrelated meanings (fracture, gap, escape, luck/speedy, dissipated, firmly, etc.).

However, the real underlying difficulty is not the presence of such compound forms but the fact that such forms may vary between the request and the relevant documents. Recently, Braschler & Ripplinger (2004) showed that decompounding German words could significantly improve retrieval performance. To automatically break up compound words into their different components, Chen (2003) suggested using a word list, and then obtaining their frequencies directly from the trained corpus. Savoy (2003) proposed looking at impossible or improbable letter sequences as a means of defining breaking point(s).

3 Test Collection

The corpus used in our experiments is composed of articles extracted from the newspaper *Magyar Hírlap,* published in 2002. This corpus was made available for the CLEF evaluation campaigns in 2005 and 2006, and contains 49,530 documents or around 105 MB of data, encoded in UTF-8 format. On average, each article contains about 142 indexing terms (or 108 distinct indexing terms) with a standard deviation of 140 (minimum: 2, maximum 4,984). A typical document in this collection begins with a short title (<TITLE> tag), usually followed by the first paragraph under the <LEAD> tag, and finally the body (<TEXT> and <P> tags). Table 1 lists an example covering a news about hurricanes in Cuba. Except for the two terms and names “Mexico” and “Yucatán”, the rest of the words in the document differ radically from our lexicon. As such it is almost impossible to get a general idea of Hungarian document contents.

This test collection contains 98 topic descriptions (see examples listed in Table 2a for English[3] and in Table 2b for Hungarian). Each description is subdivided into four different fields, namely a unique identifier (<NUM>), a brief title (<TITLE>), a full statement of the user's information need (<DESC>), and some background information that helps in assessing the topic (<NARR>). The available topics cover various subjects (e.g., “Consumer Boycotts”, “Football Refereeing Disputes”, or “Lottery Winnings”), and include both regional (“Swiss Referendums”, “Trial of Paul Touvier”) and international coverage (“Theft of The Scream”). In order to work within more realistic conditions, we will build our queries using only the title section of the topic description (or T). Additional information about the elaboration of this document collection or topics can be found in (Peters & Braschler, 2004) (Peters *et al.*, 2006).

[3] We first reported topics written in English because we believe most readers can more easily understand them. Of course, the same topic descriptions are available in the Hungarian language as shown in Table 2b and, in our experiments we only used the Hungarian topic descriptions.

```
<DOC>
<DOCNO>  MH-20020923-071
<COLUMN>  World
<TITLE>  KUBA
<SOURCE>  (MTI/CNN)
<TEXT>
<P>  Izidor, a kíméletlen hurrikán – Hurrikán söpört végig a hét végén Kuba nyugati
részén.  Az országban 290 ezer embert kitelepítettek.  A jelentések szerint egyelőre
nem követelt áldozatokat a szélvihar.  A mezőgazdaságban azonban jelentős károkat
okozott a helyenként 160 km/h sebességű szélvihar és az ezzel járó igen heves
esőzések.  Az Izidor hurrikán most Mexikó felé tart, ahol riadókészültséget
rendeltek el a várható heves szél és esőzések miatt.  A legmagasabb fokú
készültséget Yucatán szövetségi államban van érvényben.  A helyi hatóságok
felkészültek, hogy több mint 50 ezer embert átmenetileg ki kell költöztetni ottho-
naikból.
<DOC>
<DOCNO>  MH-20020923-072
…
```

Table 1. Example of an article written in Hungarian

```
<NUM>  255
<TITLE>  Internet Junkies
<DESC>  Does frequent use of the Internet cause addiction?
<NARR>  Relevant documents discuss whether regular use of the Internet is
habit-forming and can lead to physiological or psychological dependence

<NUM>  294
<TITLE>  Hurricane Force
<DESC>  What is the speed of winds in a hurricane?
<NARR>  The strength or force of a hurricane is evidenced by the wind speed.
Relevant documents must provide specific figures for hurricane storm force or wind
speed.

<NUM>  320
<TITLE>  Energy Crises
<DESC>  Find information on any kind of energy or fuel shortage.
<NARR>  Relevant documents must mention where the energy crisis occurred and
state the causes.
```

Table 2a. Examples of three topic descriptions

In this Hungarian collection, both documents are provided without any additional or specific editorial control or verification. Some documents may therefore be only partially available (some parts could have been removed) and spelling errors may occur in documents or in topic descriptions, without being explicitly introduced. This could happen for example when examining the performance of an IR system being used within more difficult contexts.

The relevance judgments were made by human assessors during the CLEF 2005 evaluation campaign for Topics #251 to #300, and in year 2006 for Topic #301 to 325 and Topic #351 to #375. Two topics (#307 “Films Set in Scotland”, and #370 “The Harry Potter Phenomenon”) were removed because no relevant information on them was found in the corpus. From an inspection of these relevance assessments, the average number of relevant articles per topic was 22.93 (median: 16; standard deviation: 21.96). Topic #272 ("Czech President's Background") had only one pertinent document while Topic #311 ("Unemployment in Europe") had the greatest number of relevant articles (134).

<NUM> <TITLE> Internetfüggők <DESC> Okoz-e függőséget az internet gyakori használata? <NARR> A megfelelő cikkek arról írnak, hogy vajon az internet rendszeres használata szokásformáló hatású-e, és vezethet-e fiziológiai vagy pszichológiai függőséghez
<NUM> 294 <TITLE> A hurrikánok ereje <DESC> Mekkora a szél sebessége egy hurrikán belsejében? <NARR> A hurrikánok erejét a szél sebessége jellemzi. A megfelelő cikkek konkrét adatokat szolgáltatnak a hurrikán erejére vagy a szél sebességére vonatkozóan
<NUM> 320 <TITLE> Energiaválságok <DESC> Keressünk cikkeket, melyek bármiféle energia- vagy üzemanyaghiányról szólnak! <NARR> A releváns cikkekből ki kell, hogy derüljön, hol és miért jelentkezett energiaválság**.**

Table 2b. Examples of three topic descriptions (in Hungarian)

During the indexing process in our automatic runs, we retained only the following logical sections from the original documents: <TITLE>, <LEAD>, <TEXT>, and <P>. From the topic descriptions we automatically removed certain phrases such as "Relevant document report …" or "Keressünk olyan cikkeket, amelyek …". Finally, diacritic characters (namely, á, é, í, ó, ö, ő, ú, ü, and ű) usually not present in English documents (with certain exceptions, such as "résumé" or "cliché") were replaced by their corresponding non-accentuated letter. Removing accents from Hungarian words may however generate additional semantic ambiguity (e.g., between "kor" (age), "kór" (illness), "kör" (circle), and "kőr" (heart, in card games) or "ver" (hurt) and "vér" (blood)). In our evaluations, we investigated the effective impact of removing accents, a practice applied successfully by several of the best-performing approaches in several CLEF evaluation campaigns involving various languages (Peters *et al.*, 2004; 2006).

4 IR Models

In order to obtain a broader view of the relative merit of the various retrieval models and stemming approaches, we used two vector-space schemes and three probabilistic models. First we adopted the classical *tf idf* model. In this case the weight attached to each indexing term was the product of its term occurrence frequency (or tf_{ij} for indexing term t_j in document d_i) and its inverse document frequency (or idf_j). To measure similarities between documents and requests, we computed the inner product after normalizing (cosine) the indexing weights.

During the first TREC evaluation campaigns better weighting schemes were suggested, especially schemes assigning more importance to the first occurrence of a term, compared to any successive and repeated occurrences. Therefore, the *tf* component was computed as the ln(*tf*)+1. Moreover, we might assume that a term's presence in a shorter document would provide stronger evidence than in a longer document, leading to more complex IR models; for example the IR model denoted by "Lnu" (Buckley *et al.*, 1996).

In addition to these two vector-space schemes, we also considered probabilistic models such as that of Okapi (Robertson *et al.*, 2000). As a second probabilistic approach we implemented the Geometric-Laplace (GL2) model, taken from the *Divergence from Randomness* (DFR) framework (Amati & van Rijsbergen, 2002) whereby the two information measures formulated below are combined:

$$w_{ij} = \mathrm{Inf}^1_{ij} \cdot \mathrm{Inf}^2_{ij} = -\log_2[\mathrm{Prob}^1_{ij}] \cdot (1–\mathrm{Prob}^2_{ij}) \qquad (1)$$

in which Prob^1_{ij} is the pure chance probability of finding tf_{ij} occurrences of the term t_j in a document. On the other hand, Prob^2_{ij} is the probability of encountering a new occurrence of term t_j in the document given, tf_{ij} occurrences of this term had already been found.

Within this framework, the GL2 model was based on the following formulae:

$$\mathrm{Prob}^1_{ij} = [1/(1+\lambda_j)] \cdot [\lambda_j /(1+\lambda_j)]^{tf} \quad \text{with } \lambda_j = tc_j/n \qquad (2)$$

$$\mathrm{Prob}^2_{ij} = tfn_{ij}/(tfn_{ij} + 1) \quad \text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl)/l_i)] \qquad (3)$$

where tc_j is the number of occurrences of term t_j in the collection, n the number of documents in the corpus, l_i the length of document d_i, *mean dl* (= 150) the average document length, and *c* a constant (fixed at 1.75).

Finally, we also considered an approach based on a language model (LM) (Hiemstra, 2000), known as a non-parametric probabilistic model (the Okapi and GL2 are viewed as parametric models). Probability estimates would thus not be based on any known distribution (as in Equation 2) but rather estimated directly, based on occurrence frequencies in document d_i or corpus C. Within this language model paradigm, various implementations and smoothing methods might also be considered, and in this study we adopted a model proposed by Hiemstra (2000) as described in Equation 4, which combines an estimate based on document ($P[t_j \mid d_i]$) and corpus ($P[t_j \mid C]$).

$$P[d_i \mid q] = P[d_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot P[t_j \mid d_i] + (1-\lambda_j) \cdot P[t_j \mid C]]$$
$$\text{with } P[t_j \mid d_i] = tf_{ij}/l_i \quad \text{and } P[t_j \mid C] = df_j/lc \quad \text{with } lc = \textstyle\sum_k df_k \qquad (4)$$

where λ_j is a smoothing factor (fixed at 0.35 for all indexing terms t_j), df_j indicates the number of documents indexed with the term t_j, and *lc* the size of the corpus C.

5 Evaluation

5.1 Evaluation Methodology

To evaluate our various IR schemes, we adopted the mean average precision (MAP) computed by the `trec_eval` software to measure retrieval performance (based on a maximum of 1,000 retrieved records). This performance measure has been used by all evaluation campaigns for more than 15 years in order to objectively compare various IR strategies, particularly regarding their ability to retrieve relevant items (ad hoc tasks) (Braschler & Peters, 2004), (Buckley & Voorhees, 2005).

Using the mean as a measure of the system's performance signifies that we attached an equal importance to all queries. Comparisons between two IR strategies will therefore not be based on a single query with respect to those available in the underlying test-collection or when specifically created in order to demonstrate that a given IR approach must be rejected. We also believe that it is important to conduct experiments involving the largest possible number of observations. To achieve this goal, we combined the topic descriptions from the CLEF 2005 and 2006 evaluation campaigns in order to base our findings on a relatively large number (98 observations).

To statistically determine whether or not a given search strategy would be better than another, we applied the bootstrap methodology (Savoy, 1997), (Abdou & Savoy, 2006). In our statistical tests, the null hypothesis H_0 stated that both retrieval schemes produce similar MAP performance. Such a null hypothesis would be accepted if two retrieval schemes returned statistically similar MAP, otherwise it must be rejected. Thus, in the experiments presented in this paper, statistically significant differences were detected by a two-sided non-parametric bootstrap test (significance level $\alpha = 5\%$).

In order to consider the best practices available, we implemented some of the most effective IR models based on the latest NTCIR (Noriko, 2005) or CLEF evaluation campaigns (Peters *et al.*, 2006). We are convinced when comparing IR models and strategies, it is not really appropriate to base our findings on IR models known for having relatively poor retrieval effectiveness. When working with really effective IR models in terms of relatively high MAP, it could be more difficult to identify statistically significant performance improvements.

Finally, it is also well known that the basis for comparisons between two (or more) IR strategies must be similar, using the same document collection and the same topics, as was mentioned by (Buckley & Voorhees, 2005).

> "The primary consequence of the noise is the fact that evaluation scores computed from a test collection are *relative* scores only. The only valid use for such scores is to compare them to scores computed for other runs using the exact same collection." (Buckley & Voorhees, 2005, p. 73).

Thus, it is clearly impossible to compare the performance obtained using an English test collection with that achieved based on another document collection written in the Hungarian language or directly performances obtained from the CLEF 2005 topics with those of CLEF 2006. Even if this were possible, we could not directly compare our performance measures with those available in the CLEF proceedings, due to the fact that the official evaluations were based on longer topic descriptions (TD) and also due to the clearly different contexts (participants had to meet strict deadlines and did not have access to the relevance judgments needed in determining the best parameter settings for their IR systems).

5.2 IR Models & Stemming Evaluation

Using the words as indexing units, Table 3 depicts the MAP achieved by our five different IR models under three different stemming strategies (None, Light, and Stemmer 2). Based on the article example shown in Table 1, we could conclude that an indexing strategy based on words is quite reasonable. In fact, unlike the Chinese language, the words are conventionally delimited and also relatively short, unlike some German words (e.g., "Friedensnobelpreis" = "Frieden" (peace) + "Nobel" + "Preis" (prize)).

In Table 3, the best performance under a given condition is depicted in bold. The first column indicates the tested IR model, the second (under the label "None") the retrieval performance when ignoring the stemming procedure. The third column (labeled "Light") lists the results of a light stemming approach adapted to remove only the number (plural form using two rules in the stemmer and depicted in Figure 1), the possessive markers (e.g., "my" with 17 rules) and the various grammatical cases (using 21 rules, examples given in Figure 2).

Finally the last column (labeled "Stemmer 2") lists the MAP obtained by a more aggressive stemmer, adapted to also remove some derivational suffixes (e.g., "féltékeny / féltékenység" = jealous / jealousy, or "talál / találat" = to hit / hit or "rend / rendel / rendes / rendelés / rendezett " = order / to order / orderly / reservation / ordered). We introduced 17 additional rules in order to achieve this goal, and some examples are given in Figure 3.

	Mean average precision		
\ Stemmer IR Model	None	Light	Stemmer 2
Okapi-npn	**0.1832**	**0.2842**	**0.3007**
GL2-nnn	0.1730*	0.2734*	0.2906*
LM-nnn	0.1661*	0.2638*	0.2830*
Lnu-ltc	0.1793	0.2656*	0.2808*
tf·idf	0.1552*	0.2067*	0.2238*
Improvement. %		+50.7%	+60.7%

Table 3. MAP of various stemmers using short queries (T) and of word-based indexing strategy

Using the best performance as a baseline (shown in bold in Table 3), we wanted to compare its retrieval effectiveness with other search models under the same condition (or same column). Statistically significant differences are indicated by an asterisk ("*") after the corresponding MAP value. Table 3 thus shows that the Okapi model always provided the best retrieval performance, usually significantly better than the other search approaches. The only exception was when comparing the Okapi (0.1832) and

Lnu (0.1793) models without stemming (labeled “None”). A query-by-query analysis revealed that the Okapi model produced better average precision for 43 queries (over a total of 98 queries), while for 40 others Lnu performed better; the same performance was achieved for the 98-43-40=15 queries. Compared to the classical *tf idf* IR model, the improvements resulting from the Okapi model varied from 37.5% (using the light stemmer) to 34.4% (with Stemmer 2).

```
If (length(word) >= 5)  then  {
  If final is “-{aoeö}k” then { remove final “-{aoeö}k”;  return }
\\       e.g., egyetemek → egyetem  (universities → university)
\\       e.g., diákok → diák  (students →  student)
  }
If (length(word) >= 4)  then  {
  if final is “-k” then { remove final “-k”;  return }
  }
```

Figure 1. The two rules used to remove the plural “-k” form

```
If (length(word) >= 6)  then  {
  If final is “-n[ae]k”  then  { remove final “-n[ae]k”;  return }
\\    e.g., háznak → ház  (of the house (dative/genitive)→ house))
  If final is “-b[ae]n”  then  { remove final “-b[ae]n”;  return }
\\    e.g., házban → ház  (in the house (inessive)→ house))
  …  }
If (length(word) >= 5)  then  {
  If final is “-b[ae]”  then  { remove final “-b[ae]”;  return }
\\    e.g., házba → ház  (into the house (illative)→ house))
  …  }
```

Figure 2. Examples of rules used to remove the suffixes associated with some grammatical cases

```
.If (length(word) >= 8)  then  {
  If final is “-oss[áé]g”  then { remove final “-oss[áé]g”; return }
\\       e.g., alázatosság → alázat  (humbleness)
  …  }
If (length(word) >= 5)  then  {
  If final is “-[áé]s”  then { replace final “-[áé]s”;  return }
\\       e.g., temetés → temet  (burial, funeral → to bury)
  If final is “-[ae]t”  then  { replace final “-[ae]t”;  return }
\\       e.g., találat → talál  (hit → to find, discover)
  …  }
```

Figure 3. Examples of rules used to remove certain derivational suffixes

A comparison of stemming strategies needs to be done column by column. As a first experiment, we used baseline IR performances obtained when ignoring the stemming procedure (column labeled “None”). After applying the light stemming (column “Light”) or our more aggressive stemmer (“Stemmer 2”), the performance obtained after applying stemming was always statistically better than that achieved when ignoring stemming. As depicted in the last line of Table 3, the mean improvement over the baseline was around 50.7% for the light stemmer, and 60.7% for the more aggressive stemmer.

Following Harman's study (Harman, 1991), we may assume that different stemmers do indeed produce different results, but performance differences are not statistically significant. To verify this assumption, we used the performance results of the light stemmer (column "Light") as a baseline. The statistical test indicated that when compared with this baseline, the performance is always statistically significant (we underlined the corresponding MAP values in Table 3). Using the more aggressive stemmer, we obtained significantly better performance than the light stemming approach. Unlike the English, the Hungarian morphology was more complex and thus a more aggressive word normalization procedure provided significantly better MAP.

The effect of applying a stemmer could be illustrated by inspecting some of the queries. Overall, the more aggressive stemming strategy was not able to find any relevant item for four requests over 98. When ignoring the stemming procedure however, the search system could not find any pertinent information for 11 requests. The greatest improvement after adopting "Stemmer 2" was obtained with Topic #279 ("Swiss referendums" or "Svájci népszavazások"), for which there were nine relevant articles. Ignoring the stemming, the Okapi model resulted in an average precision (AP) of 0.0257, by retrieving six relevant documents (in ranks 11, 37, 87, 203, 227, and 579). In this case, the underlying query was composed of two words "svajci" and "nepszavazasok" (the accents have been removed by the indexing system). Using Stemmer 2, the search terms were "svajc" and "nepszavaz", and this query obtained an AP of 0.8944, retrieving eight relevant documents in the first eight positions (the last one appears in position 182). The performance difference between these two queries was not related with the Swiss word ("svájci") appearing in all relevant documents, but with the word referendum ("népszavazások"). In relevant documents, we encountered the forms "népszavazással", "népszavazáson" or "népszavazást" that were not conflated under the same stem when we ignored the stemming stage.

On the other hand, when comparing the two stemmers, a query-by-query analysis revealed that the largest improvement was obtained with Topic #255 ("Internet Junkies" or "Internetfüggők") having six relevant documents. With the light stemmer, this query only retrieved three documents, with all of them being relevant (AP: 0.5). Using the aggressive stemmer, the Okapi model obtained an average precision of 0.9762, retrieving only seven documents. All these articles were pertinent, except for the item ranked sixth. With the light stemming, the query consisted of a single search term ("internetfugg") while the forms appearing in the relevant documents were usually different (e.g., "internetfüggőség", "internetfüggőséggel", "internetfüggőségben"). For this request, the stem was the verb form "függ" (to depend on), and the topic used a form indicating that a person was ill or dependent ("függők"). The form appearing in the relevant articles were dependence ("függőség") with various grammatical case endings ('-gel' or '-ben'). The light stemmer removed these endings, and obtained the form related to the illness ("függőség"), while the query used another form ("függők").

5.3 And the diacritics?

In the previous experiments, all diacritics were removed (both in the documents and in the queries). As shown in certain examples in Section 2, diacritics may be useful in clearly identifying a word's meaning, and without them increased polysemy could occur, such as evidenced by the word forms (e.g., "ver" (hurt) and "vér" (blood)). Removing diacritics when stemming may also be helpful however, as evidenced by the noun "tűz" (fire) that could appear as "tüzet" (the accusative case), formed not only by adding the suffix '-(e)t' but also by modifying the accent. As another example, the noun "levél" (letter) is written as "levelek" in the plural form, and the accent disappears.

In order to verify the impact on retrieval effectiveness caused by keeping or removing diacritics, we repeated the two experiments shown in Table 3 in their corresponding runs, after having preserving the diacritics.

	Mean average precision (% change)			
\ Stemmer IR Model	None & diacritics	None	Light & diacritics	Light
Okapi-npn	0.1801 (-1.7%)	**0.1832**	0.2572 (-9.5%)	**0.2842**
GL2-nnn	0.1707 (-1.3%)	**0.1730**	0.2551 (-6.7%)	**0.2734**
LM-nnn	0.1635 (-1.6%)	**0.1661**	0.2423 (-8.1%)	**0.2638**
Lnu-ltc	0.1791 (-0.1%)	**0.1793**	0.2526 (-4.9%)	**0.2656**
tf·idf	0.1549 (-0.2%)	**0.1552**	0.2060 (-0.3%)	**0.2067**
Improvement. %	-1.0%		-5.9%	

Table 4. MAP of various stemmers using short queries (T), with and without removing diacritics

Table 4 illustrates performance differences between runs with no stemming (under the label "None & diacritics" and "None") and those with our light stemmer (under "Light & diacritics" and "Light"). The "Light & diacritics" column displays the results when diacritics were preserved in the documents, the queries and stemming rules.

A comparison of both indexing strategies is shown in the "None" and "None & diacritics" columns, revealing that the performance differences are rather small and always is favor of diacritic removal. Taking the performance levels in the "None" column being used as a baseline, our statistical tests showed no statistically significant differences could be detected, e.g. both IR strategies resulted in the same performance levels. The last column in Table 4 demonstrates that removing the diacritic marks when applying a stemming procedure lead to more effective retrieval, and this improvement was more significant (5.9% in average) over a similar approach with diacritics. It is only with the Okapi model (0.2842 vs. 0.2572) however that performance differences were statistically significant. Finally, it is interesting to note that the performances listed in the "None & diacritics" column are those achieved when only the inverted file contained correctly spelled words.

5.4 Automatic Decompounding

As a second indexing strategy, we decided to automatically decompounding Hungarian compound words (e.g., “munkanap” = “munka” (work) + “nap” (day)). It is known that such linguistic constructions are used frequently in German, but they are also present in the Hungarian language. We had previously saw Topic #255 containing the compound term “Internetfüggők” (“Internet Junkies“). After applying our decompounding scheme, the query consisted of one compound construction (“internetfugg”) and two single terms (“intern” and “fugg”). From examining relevant items, we can see that some of them used the compound construction (“Internetfüggők” or “Internetfüggőség”) while in other articles the concept was expressed using two words separately (“internetezik” and “függőséggel”).

Our automatic decompounding approach (Savoy, 2004) increased the mean query size, from 2.21 to 3.22. As shown in Table 5, the IR performance increased but the previous findings were the same. First, the best MAP was obtained by the Okapi model and the performance differences with other approaches were usually statistically significant (as indicated by an “*”). Using the retrieval effectiveness obtained by IR models ignoring the stemming procedure as baseline (column “None”), the two stemmers performed significantly better and, as shown in the last line, the mean improvement that resulted was similar. Finally, using the light stemmer as a baseline, the more aggressive stemmer resulted in significantly better MAP for all IR models.

	Mean average precision		
\ Stemmer IR Model	None	Light	Stemmer 2
Okapi	**0.1964**	**0.3073**	**0.3308**
GL2	0.1871*	0.2967*	0.3268
LM	0.1804*	0.2880*	0.3140*
Lnu-ltc	0.1914	0.2878*	0.3124*
tf·idf	0.1588*	0.2215*	0.2427*
Improvement. %		+52.9%	+66.6%

Table 5. MAP of various stemmers using short queries (T) with automatic decompounding

When comparing word-only indexing scheme (Table 3) with an indexing scheme using compounds and their composite parts (Table 5), the largest difference was achieved by Topic #271 (“Gay Marriages “ or “Melegházasságok” with “meleg” = gay / hot / heat and “házasság” = marriage), having nine relevant items. The Okapi model using the word-only approach achieved an AP of 0.1111 (the result list was limited to a single document that was also pertinent). After applying our decompounding algorithm, the AP was 0.6184 (804 articles were retrieved, the first four were pertinent and other pertinent items were found in ranks 7, 8, 96, and 280). The reason of course for this performance difference is revealed upon inspecting both queries. In the first, the query is limited to one search term (the compound term “meleghazas”), while in the second case there are three stems (“meleg”, “hazas” and “meleghazas”) allowing a better matches to extract the relevant items.

5.5 Using Different Topic Formulations

Previously we only considered the shortest topic formulation (see examples given in Table 2). During the CLEF campaigns, the official evaluation was based on the query composed of the title and descriptive parts (TD) of the topic. Finally, we also consider the longest query formulation using all topic fields (TDN).

Table 6 shows the evaluation obtained with these three topic formulations, using the best stemmer (namely "Stemmer 2") and after decompounding the Hungarian terms. In the second row of this table, we indicated the mean query size of these three topic formulations. When considering longer topic formulations (TD or TDN), the GL2 probabilistic model performed better than the Okapi, and the performance difference with the Okapi model was even statistically significant when using the longest topic formulation (TDN) (as indicated by an "*").

	Mean average precision		
\ Query	T	TD	TDN
\ mean query size	3.22	10.31	20.61
Okapi	**0.3308**	0.3412	0.3411*
GL2	0.3268	**0.3451**	**0.3525**
LM	0.3140*	0.3401	0.3281*
Lnu-ltc	0.3124*	0.3288	0.3339*
tf·idf	0.2427*	0.2624*	0.2664*
Improvement. %		+6.1%	+6.4%

Table 6. MAP of various topic formulations with decompounding indexing terms (using Stemmer 2)

While using the title-only query formulation as a baseline and comparing the two longer topic formulations, performance differences were statistically significant for the classical *tf idf* model (values underlined in Table 6). Compared to the title-only queries, the mean improvement was rather small, +6.1% when using TD queries or +6.4% for the longest topic formulation.

5.6 Using different indexing units

In order to represent documents and queries, we used a word-based indexing approach and the words resulting from decompounding. As a language-independent approach, we might consider 4-gram indexing strategy (McNamee & Mayfield, 2004). The evaluation of these three indexing strategies was done using title-only topic formulation as shown in Table 7. The Okapi probabilistic model produces the best IR performance, and is usually significantly better (denoted by an "*") than other IR models (differences from the GL2 model are however usually not significant). When using the decompounding strategy as a baseline, the performance differences were only statistically significant with word-only indexing strategies (values underlined).

Comparing the 4-gram strategy with an indexing scheme using compounds and their composite parts, the largest difference was achieved by Topic #306 ("ETA Activities in France" or "ETA-tevékenységek Franciaországban"), consisting of six relevant items. The Okapi model combined with a 4-gram indexing scheme achieved an AP of 0.0101 (relevant items ranked in positions 68, 81, 306, 646, and 932) while the decompounding indexing approach resulted in an AP of 0.5807 (relevant items ranked in positions 1, 2, 7, 8, and 9). The underlying query was composed of five search terms, namely “eta”, “tevekenyseg” (activity), “franciaorszag” (France), “franci” (French) and “orszag” (country). In this case, the 4-gram generated multiple matches for the compound construction “Franciaországban” and the word “tevékenységek”. These matches retrieved many non-relevant documents that did not have the right actor (ETA in this case) but others such as “France Télécom” or “Jacques Chirac”.

	Mean average precision		
\ Indexing	Word only	Word & Decompound	4-gram
\ mean query size	2.21	3.22	11.91
Okapi	**0.3007**	**0.3308**	**0.3236**
GL2	0.2906*	0.3268	0.3212
LM	0.2830*	0.3140*	0.3114*
Lnu-ltc	0.2808*	0.3124*	0.2819*
tf·idf	0.2238*	0.2427*	0.2496*
Improvement. %		+10.6%	+8.0%

Table 7. MAP of various stemmers using short queries (T) with different indexing terms (using Stemmer 2)

By contrast, Topic #315 (“Doping in Sports” or “Doppingolás a sportban”) consisting of 73 relevant items) obtained an AP of 0.6713, using the Okapi model combined with 4-gram indexing scheme, and only 0.289 with the same search model combined with the decompounding scheme (the query was “doppingol”, “spor”). The *n*-gram indexing scheme had the advantage of allowing multiple matches (for the doping concept in this case) which clearly boosted the number of relevant articles for this request.

6 Conclusion

In this paper we described the most important linguistic features of the Hungarian language, from an IR perspective. Not only does this language use a relatively large set of unambiguous suffixes, but its morphology is also complex, due to the use of possessive pronouns being sometimes added to the suffix construction. Using a test collection extracted from the CLEF-2005 & 2006 suite containing 98 requests, we evaluated three probabilistic and two vector-space models. When using the title-only queries, the Okapi model resulted in the most effective retrieval, under a variety of conditions.

This paper also presents a light stemming strategy used to remove only inflectional suffixes, as well as a more aggressive algorithmic stemmer used to remove some derivational suffixes. Compared to IR models ignoring the stemming procedure, the mean improvement is around +53% for the light stemmer, and +67% for the more aggressive stemmer. When considering the English language (Harman, 1991) in which both stemmers tend to produce statistically similar performance, a comparison of these two stemmers shows that a more aggressive approach produces significantly better results. These performance differences become evident upon analyzing some of the queries.

Also evaluated in this paper is the application of an automatic decompounding algorithm in order to separate compound construction (e.g., viewpoint) into their composite parts. Such an approach produces significantly better MAP (around +10%) than an approach based on a word-only indexing scheme. Finally, including more search terms into the topic formulation (T vs. TD) improves retrieval effectiveness by 6%, an enhancement that is not always statistically significant.

For the Hungarian language, additional work and experiments are needed to obtain a more complete view of the stemming problem. One solution may be to apply more complex morphological analysis based on a lexical stemmer or on a lemmatizer (Haláscy, 2006). A second may be to consider the language usage more closely through adding, modifying or removing rules applied in an algorithmic stemming approach, so that they take the frequency of various grammatical rules into closer consideration.

Acknowledgments
This research was supported in part by the Swiss NSF under Grant #200021-113273.

References

Abdou, S. & Savoy, J. (2006). Statistical and comparative evaluation of various indexing and search models. In Proceedings AIRS 2006 (pp. 362-373). LNCS #4182. Berlin: Springer.

Amati, G., & van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM-Transactions on Information Systems, 20(4), 357-389.

Braschler, M., & Peters, C. (2004). Cross-language evaluation forum: Objective, results, achievements? IR Journal, 7(1-2), 7-31.

Braschler, M., & Ripplinger, B. (2004). How effective is stemming and decompounding for German text retrieval? IR Journal, 7(3-4), 291-316.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996). New retrieval approaches using SMART. In Proceedings of TREC-4 (pp. 25-48). Gaithersburg: The NIST Publication.

Buckley, C., & Voorhees, E. (2005). Retrieval system evaluation. In TREC, Experiment and Evaluation in Information Retrieval (pp. 53-75). Cambridge: The MIT Press.

Chen, A. (2003). Cross-language retrieval experiments at CLEF 2002. In Advances in Cross-Language Information Retrieval (pp. 28-48). LNCS #2785, Berlin: Springer.

Di Nunzio, G.M., Ferro, N., Melucci, M., & Orio, N. (2004). Experiments to evaluate probabilistic models for automatic stemmer generation and query word translation. In Comparative evaluation of multilingual information access systems (pp. 220-235). LNCS #3237, Berlin: Springer.

Fox, C. (1990). A stop list for general text. SIGIR Forum, 24(1-2), 19-35.

Gordon, M., & Pathak, P. (1999). Finding information on the world wide web: the retrieval effectiveness of search engines. Information Processing & Management, 35(2), 141-180.

Haláscy, P. (2006). Benefits of deep NLP-based lemmatization for information retrieval. http://clef.iei.pi.cnr.it/2006/working_notes/workingnotes2006/halacsyCLEF2006.pdf

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

Harman, D.K. (2005). The TREC ad hoc experiments In TREC, Experiment and Evaluation in Information Retrieval (pp. 79-97). Cambridge: The MIT Press.

Harter, S.P. (1986). Online Information Retrieval: Concepts, Principles and Techniques. San Diego: Academic Press.

Hiemstra, D. (2000). Using language models for information retrieval. CTIT Ph.D. Thesis.

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

Kettunen, K. & Airo, E. (2006). Is a morphologically complex language really that complex in full-text retrieval? In Advances in Natural Language Processing (pp. 411-422). LNCS #4139, Berlin: Springer.

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of Finnish text documents. In Proceedings of the ACM-CIKM (pp. 625-633). New York: The ACM Press.

Krovetz, R. (1993). Viewing morphology as an inference process. In Proceedings of the ACM-SIGIR, (pp. 191-202). New York: The ACM Press.

Lovins, J.B. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11(1), 22-31.

McNamee, P., & Mayfield, J. (2004). Character *n*-gram tokenization for European language text retrieval. IR Journal, 7(1-2), 73-97.

Moulinier, I. (2004). Thomson Legal and Regulatory at NTCIR-4: Monolingual and pivot-language retrieval experiments. In Proceedings NTCIR-4, (pp. 158-171). Tokyo: National Institute of Informatics.

Noriko, K., (Ed) (2005). NTCIR Workshop 5 Meeting. Tokyo: National Institute of Informatics.

Peters, C., Gonzalo, J., Braschler, M. & Kluck, M., (Eds) (2004). Comparative evaluation of multilingual information access systems. LNCS #3237. Berlin: Spinger-Verlag.

Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B. & de Rijke, M. (Eds) (2006). Accessing multilingual information repositories. LNCS #4022. Berlin: Spinger-Verlag.

Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137.

Robertson, S.E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. Information Processing &Management, 36(1), 95-108.

Savoy, J. (1993). Stemming of French words based on grammatical category. Journal of the American Society for Information Science, 44(1), 1-9.

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. Information Processing & 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. (2003). Report on CLEF 2002 experiments. In Advances in Cross-Language Information Retrieval (pp. 66-90). LNCS #2785, Berlin: Springer.

Savoy, J. (2004). Report on CLEF 2003 monolingual tracks. In Comparative evaluation of multilingual information access systems (pp. 322-336), LNCS #2785, Berlin: Springer.

Sproat, R. (1992). Morphology and computation. Cambridge: The MIT Press.

Tomlinson, S. (2004). Lexical and algorithmic stemming compared for 9 European languages with Humminbird SearchServer™ at CLEF 2003. In Comparative Evaluation of Multilingual Information Access Systems (pp. 286-300). LNCS #3237, Berlin: Springer.

Xu, J., & Croft, B. (1998). Corpus-based stemming using cooccurrence of word variants. ACM-Transactions on Information Systems, 16(1), 61-81.

Appendix: Term weighting formulae

When assigning an indexing weight w_{ij} to reflect the importance of the term t_j in a document d_i, the Lnu model is based on the following weighting formula:

$$w_{ij} = [(\ln(tf_{ij})+1)/(\ln(\textit{mean dl})+1)]/[(1-\textit{slope}) \cdot \textit{pivot} + \textit{slope} \cdot nt_i] \qquad (A.1)$$

where nt_i indicates the number of indexing terms included in d_i, *slope,* and *pivot* are a constant (fixed at *slope*=0.1 and *pivot*=75 in our experiments), and *mean dl* indicates the average document length. The Okapi model is based on the following weighting formula:

$$w_{ij} = [(k_1+1) \cdot tf_{ij}]/(K + tf_{ij}) \quad \text{with } K = k_1 \cdot [(1-b) + ((b \cdot nt_i)/\textit{mean dl})] \qquad (A.2)$$

where b, k_1, are constants fixed at $b = 0.75$, $k_1 = 1.2$ in our experiments.