Searching Strategies for the Bulgarian 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 indexing and searching with the Bulgarian language. For this language we propose a general light stemmer and demonstrate that it can be quite effective, producing significantly better MAP (around + 34%) than an approach not applying stemming. We implement the GL2 model derived from the *Divergence from Randomness* paradigm and find its retrieval effectiveness better than other probabilistic, vector-space and language models. The resulting MAP is found to be about 50% better than the classical *tf idf* approach. Moreover, increasing the query size enhances the MAP by around 10% (from T to TD). In order to compare the retrieval effectiveness of our suggested stopword list and the light stemmer developed for the Bulgarian language, we conduct a set of experiments on another stopword list and also a more complex and aggressive stemmer. Results tend to indicate that there is no statistically significant difference between these variants and our suggested approach. This paper evaluates other indexing strategies such as 4-gram indexing and indexing based on the automatic decompounding of compound words. Finally, we analyze certain queries to discover why we obtained poor results, when indexing Bulgarian documents using the suggested word-based approach.

Keywords: Cross-language information retrieval; Bulgarian IR; stemmer, evaluation, morphology.

1 Introduction

The Slavic languages (e.g., Russian, Polish, Czech, Slovenian, Serbo-Croatian or Bulgarian) predominate in Central and Eastern Europe, but only a very limited number of test collections are available for this family of languages. For example, a Russian test collection was created during the 2003 and 2004 (Peters *et al.*, 2005) CLEF campaigns, but due to its small size (16,716 documents or 68 MB) we were not able to draw any definitive conclusions. This was mostly due to the fact that numerous queries found only a fairy small number of relevant items. For example, for seven queries out of a total of 28 for 2003, or ten out of 34 for 2004, we found only one relevant document (and four other queries in 2003 and seven in 2004 found only two

pertinent items). These rather limited results have a clear impact on any comparative evaluations. For example, if a given IR system ranks the only pertinent document in the first position, the average precision (AP) obtained for this query is 1.0. On the other hand, if this item is ranked in second position, it obtains an AP of only 0.5. When repeating this swapping between first and second places for all requests having only one relevant item, the absolute difference in mean average precision (MAP) for the 34 queries processed is 0.147 (or [0.5·10] / 34), a relatively high value given that the average MAP for this test collection is around 0.35 (Peters *et al.*, 2005). As another example, we may mention experiments done on the Slovenian language (Popovic & Willett, 1992) based also on a very small collection (504 documents, 48 queries).

The main objective of our paper is to describe some of the morphological difficulties involved in working with the Bulgarian language, a Slavic language for which a larger test collection was made available during the 2005 and 2006 CLEF evaluation campaigns (Peters *et al.*, 2006). We will also propose and evaluate a suitable light stemmer for this Slavic language using different indexing and search strategies. The rest of 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, while Section 5 evaluates them under different indexing and stemming conditions and compares our suggested stemming and stopword list with other variants. A query-by-query analysis will conclude this evaluation. The main findings of this paper are summarized in Section 6.

2 Context and Related Work

2.1 Stopword List

In order to define pertinent matches between search keywords and documents, we removed very frequently occurring terms having no important significance (e.g., the, in, but, some). For the Bulgarian language, we first created a list of the top 200 most frequently occurring forms found in the corpus, from which we removed certain words (e.g., police, government, minister) as described in (Fox, 1990). The final list derived by adding certain articles (e.g., a = “един”, “една”, this = “този”, “тази”, “това”, these = ” тези”, …), pronouns (e.g., I = “аз”, he = “той”, she = “тя”, it = “то”, them = “те”, you = “тебе”, “вие”, “ти”, …), possessive pronouns (e.g., your = “твой”, “твоя”, “твое”, “твои”, …), prepositions (e.g., with = “със”, of = “от”, in = “в”, “във”, for = “за”, …), conjunctions (and = “и”, but = “но”, “пък” , …), very frequently occurring verb forms (e.g., am = “съм”, is = “е”, was = “беше”, to have = “имам”, …), and some words (e.g., yes = “да”). The final stopword list contains 258 Bulgarian terms (see Table A.1 in the Appendix).

2.2 Characteristics of Bulgarian Morphology

Bulgarian shares many characteristics with the other Slavic languages (e.g., Russian, Polish or Czech), some morphological features with other Balkan languages (Greek, Albanian or Romanian), and generally with certain Indo-European languages (Sproat, 1992). As with the Latin or the German languages, in the Slavic languages the various grammatical cases are usually marked by suffixes (e.g., the noun "city" in Russian could be written as "город" (nominative), "города" (genitive) or "городе" (locative)). With the exception of the vocative case however, these grammatical cases are usually not explicitly indicated by a given suffix in the Bulgarian language (Allières, 2000). As with the English language, traces of these declensions are still detectable upon inspecting certain pronouns (e.g., "I" (nominative) and "me" (accusative)). These variations are usually included in stopword lists and thus do not cause any specific IR problems.

Thus for the Bulgarian language we suggest that a light stemmer would be the easiest solution. Other morphological features must however be taken into account. Bulgarian has three genders (masculine, feminine and neutral), and plural forms comprising more variations than in English (where the usual suffix is the '-s', however there are certain exceptions as in "foot / feet"). In Bulgarian the plural is represented by various suffixes (e.g., "компютър" / "компютри" (computer /s), "име" / "имена" (name /s), or "град" / "градове" (city /-ies)). The same suffix may be used with different genders (e.g., the '-и' used usually to denote the plural). One of the difficult aspects of Bulgarian morphology is that the stem may vary (e.g., "място" / "места" (place /s) or in "ден" / "дни" (day /s)). To remove the suffix denoting the plural form, we created 10 rules for our stemmer.

Unlike the morphology of other Slavic languages, Bulgarian employs a suffix to indicate the definite article (the). For example, the neutral noun "море" (sea) becomes "морето" (the sea), which in the plural becomes "морета" (seas) and "моретата" (the seas). For feminine nouns the definite article is represented by various suffixes (e.g., '-та') and its plural form (e.g., '-те'). For masculine nouns, there are two possibilities (namely '-ът' or '-а' and '-ят' or '-я'), each with a long or short form. The selection of either the long or short form depends on the noun's function in the sentence. The long form is used when a masculine noun serves as verb subject and the short form for other grammatical cases (e.g., "син" (son) becomes "синът" (the son, long form) or "сина" (short form)). The second possibility is "кон" (horse), "конят" (the horse) or "коня", which in the plural becomes "коне" (horses) and "конете" (the horses). In our light stemmer, 8 rules are applied to control the removal of the definite article. Note also that in Bulgarian the indefinite articles (a/an) are not represented by a suffix, but they appear on their own (e.g. "едно море" (a sea), while other forms are "един" (for masculine noun), "една" (feminine) and "едни" (plural)).

As with many languages, the suffixes assigned to adjectives agree with the attached noun in gender and number (e.g., "луд" (mad) in masculine gives "луда" in feminine, "лудо" in neutral, and "луди" in plural). Such a general rule may hide certain particularities, such as in the sentence "бащата е добър" (father-the is good) or "добрият баща" (good-the father).

2.3 Stemming Strategies

The stemming process is used to conflate word variants into a common stem (or form when the string cannot be found in the language). When indexing documents or requests in IR, stemming is assumed to be a good practice. For example, when a query contains the word "horse," it seems reasonable to also retrieve documents containing the related word "horses." Effective stemming procedures may also be helpful for other purposes, such as text data mining, natural language processing or gathering statistics on a document corpus. The *n*-gram indexing strategy is however viewed as an exception to this rule (McNamee & Mayfield, 2004), given that this approach does not usually apply a stemming stage.

As a first approach to designing a stemmer, we begin by removing only inflectional suffixes so that singular and plural word forms (e.g., "dogs" and "dog") or feminine and masculine variants (e.g., "actress" and "actor") will conflate to the same root. Stemming schemes that remove only morphological inflections are termed as "light" suffix-stripping algorithms, while more sophisticated approaches have also been proposed to remove derivational suffixes (e.g., '-ment', '-ably', '-ship' in the English language). Those suggested by Lovins (1968) or by Porter (1980) are typical English language uses. When considering other Indo-European languages, we can find stemmers suggested for the German (Braschler & Ripplinger, 2004), Dutch (Kraaij & Pohlman, 1996), Swedish (Hedlund *et al*., 2001; Ahlgren & Kekäläinen, 2007), French (Savoy, 1999), Slovene (Popovic & Willett, 1992), modern Greek (Kalamboukis, 1995), Latin language (Schinke *et al*., 1998) or more generally during the various CLEF evaluation campaign (Peters *et al.*, 2006). Of course, stemmers for members of other language families can be found such as for the Finnish (Alkula, 2001), Hungarian (Savoy, 2007), or Turk language (Ekmekçioglu & Willett, 2000). Stemming procedures have been suggested for other non-European languages as for example the Arabic (Chen & Gey, 2003), (Savoy & Rasolofo, 2003), Malay (Ahmad *et al*., 1996) or Indonesia language (Asian *et al*., 2004), but such word normalization procedure has no or little impact in other cases such as for the Chinese, Japanese or Korean language (Savoy, 2005).

Stemming schemes are usually designed to work with general text in any given language. Certain stemming procedures may however be especially designed for a specific domain (e.g., medicine) or a given document collection. For example Xu & Croft (1998) suggest that statistical stemming procedures be developed using a corpus-based approach, more closely reflecting the language used (including characteristic 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. To measure the frequency of each possible suffix, Kettunen & Airo (2006) have studied the Finnish language. In theory Finnish nouns have around 2,000 different forms, yet most of these forms rarely occur in actual collections. As a matter of fact 84 to 88% of the occurrences of inflected nouns in Finnish are generated by only six out of a possible 14 cases.

Stemming procedures ignore word meanings and thus 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). In analyzing the IR stemming performance of three different stemming strategies, Harman (1991) demonstrated that no statistically significant improvements could be obtained. A query-by-query analysis revealed however that stemming did indeed affect performance, even though the number of queries showing improvements was nearly equal to the number of queries showing decreased performance. Other studies (limited to the English language only), show that applying a stemmer may lead to modest improvements (Hull, 1996) or small degradation (Abdou *et al*., 2006). When compared with approaches that ignored stemming however, differences were not always statistically significant (Abdou *et al*., 2006)

When evaluating two different stemming strategies, Di Nunzio *et al*. (2004) showed that relative retrieval performances vary for each of the five languages studied. This means that any given stemming approach may work well for one language but not for another. When compared to statistical stemmers, Porter's stemmers seem to work slightly better. Braschler & Ripplinger (2004) showed that for short queries in German, 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 (based on a dictionary of the language involved). Moreover for the Finnish and German languages, Tomlinson (2004) found that the lexical stemmer based on a dictionary and a more complex morphological analysis tended to produce statistically significant results, while for seven other languages the performance differences were small and insignificant.

2.4 Compound Words

Compound word construction (e.g., handgun, viewfinder) is another morphological characteristic that may have an impact on retrieval effectiveness. Most European languages allow some form of compound construction, indicated by a hyphen sign in some cases (e.g, in French "porte-clefs" (key ring)) or by a 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 a compound from two or more words, as in the English (viewpoint) or German language ("Bankangestelltenlohn"). Such word composition is not limited to the Germanic family, and in Finnish similar constructions are possible, such as "rakkauskirje" = "rakkaus" (love) and "kirje" (letter). In Bulgarian, we also encounter this word formation as, for example, "радиоапарат" = "радио" (radio) + "апарат" (receiver), or in "мироопазване" = "мир" (peace) + "опазване" (keeping).

The real underlying difficulty is not the presence of such compound forms but the fact that there may be variant forms found among requests and relevant documents. Recently, Braschler & Ripplinger (2004) showed that decompounding German words could significantly improve retrieval performance. In order to automatically break up compound words into their various components, Chen (2003) or Savoy (2004) sug-

gest using a word list and then obtaining their frequencies directly from the training corpus.

3 Test Collection

The corpus used in our experiments consists of articles extracted from the newspapers *Sega* and *Standart* published in 2002. This corpus was made available for the CLEF evaluation campaigns in 2005 (Peters *et al.*, 2006) and 2006, and contains 69,195 documents or around 213 MB of data, encoded in UTF-8. On average, each article contains about 133.7 indexing terms having a standard deviation of 145 (min: 1, max: 2,805). 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), as shown in Figure 1.

This test collection contains 99 topics (an example is given in Figure 2), 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., "Oil Price Fluctuation", or "Human Cloning and Ethics"), and include both regional ("Hungarian-Bulgarian Relationships") and international coverage. In order to work within more realistic conditions, we mainly evaluate our system using queries that contain only the title section (or, in short, T) or both the title and descriptive parts (TD).

<DOC>
<DOCNO> NST2002-04-17-043 </DOCNO>
<TITLE> Стриктен график за ползване на домашния компютър направила пристрастената двойка </TITLE>
<AUTHOR> Биляна Веселинова </AUTHOR>
<DATE> 11/02/2002 </DATE>
<RUBRIC> thecountry </RUBRIC>
<LEAD> Семейство се лекува от Интернет </LEAD>
<LEAD> Психолог ще помага на изпадналите в зависимост млади хора</LEAD>
<TEXT>
<P> Младо шуменско семейство, обхванато от Интернет-мания, близо 3 месеца говори само чрез """мрежата""". 33-годишният Иван К. и съпругата му Елица, 25 г., почти не излизали от чат-каналите и дори направили стриктен график за ползване на домашния компютър. Тъй като почасовият списък за достъп до Интернет не помогнал, Иван се принудил да остава до ранни зори във фирмения си офис, за да е нонстоп онлайн. Когато са вкъщи, двамата си пишат есемеси или си пускат съобщения по електронната поща. Родителите на семейната двойка били сериозно притеснени, тъй като от доста време двамата не отделяли никакво внимание за фамилните сбирки. Заради пристрастеността си към виртуалната комуникация семейството потърсило помощта на известен психолог. Най-малко два месеца щяла да продължи терапията на кибердвойката - казаха запознати. </P>
</DOC>

Figure 1. Example of an article about "addiction to Internet"

The relevance judgments were made by human assessors during the CLEF 2005 evaluation campaign for Topics #251 to #300, and in year 2006 for Topics #301 to 325 and Topics #351 to #375. Topic #292 was removed because no relevant information on it was found in the corpus. From an inspection of these relevance assessments, the average number of relevant articles per topic was 20.47 (median: 12; standard deviation: 22.51). Three topics (#258, #272, and #296) had only one pertinent document while Topic #316 ("Strikes") had the greatest number of relevant articles (158).

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

<NUM> 255 </NUM>
<TITLE> Пристрастяване към Интернет </TITLE>
<DESC> Дали честото ползване на Интернет води до пристрастяване? </DESC>
<NARR> Подходящите документи дикутират дали честото ползване на Интернет формира определени навици и може да доведе до психологическа или физическа зависимост </NARR>

Figure 2. Example of a topic description in English and Bulgarian languages

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, wherein 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 (for more information, see Chapter 2 in (Baeza-Yates & Ribeiro-Neto, 1999)).

Better weighting schemes were suggested during the TREC evaluation campaigns, especially in those schemes that assigned more importance to the first occurrence of a term, compared to any successive and repeated occurrences. Therefore, the *tf* component was computed as the $\ln(tf_{ij})+1$. 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) wherein the two information measures formulated below are combined:

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

where $Prob^1_{ij}$ 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. The GL2 model was based on the following formulae:

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

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

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 and based on occurrence frequencies in document d_i or the entire C corpus. 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 ($P[t_j | d_i]$) and corpus ($P[t_j | C]$).

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

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 is a constant related to the underlying corpus C.

In Equation 4, $P[d_i]$ is the prior probability that the document d_i is pertinent. This value was ignored in our experiments because it did not vary across documents and thus did not change the final ranking. For web searches, this probability may vary across different web pages, depending on the number of incoming links, page length or page position within the web site (Kraaij *et al*., 2002).

5 Evaluation

To evaluate our various IR schemes, we adopted the mean average precision (MAP) computed by the `trec_eval` software in order to measure retrieval performance (based on a maximum of 1,000 retrieved records). To statistically determine whether or not a given search strategy would be better than another, we applied the non-parametric bootstrap test (Savoy, 1997). In our statistical tests, the null hypothesis H_0 stated that the two retrieval schemes used in the comparison produce similar MAP performance. Thus, in the experiments presented in this paper, statistically significant differences were detected by a two-sided test (significance level 95%) based on the mean (more precisely the MAP), and the corresponding computations were done using R (Crawley, 2005). To complete such an overall evaluation we analyzed the retrieval performance of some queries, in order to obtain a better understanding of the effect of a given search strategy

5.1 IR Models & Stemming Evaluation

Table 1 depicts the MAP achieved by five different IR models with and without stemming. In this table, the best performance under a given condition is shown in bold. The first column indicates the tested IR model and the second (labeled "None") lists 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, the vocative case and the definite article. All the rules included in our light Bulgarian stemmer are depicted in Table A.2 in the Appendix

Using the best performance as a baseline, we wanted to compare the retrieval effectiveness with other search models under the same condition (or same column). Statistically significant differences are indicated by an asterisk ("*") next to the corresponding MAP value. Table 1 thus shows that the Okapi model provided the best retrieval performance when we ignored the stemming (under the label "None"), while the GL2 provided the best MAP after stemming. The performance differences between the three probabilistic models (Okapi, GL2, and LM) were not significant. The difference between the best IR model and the vector-space approaches were however usually statistically significant. When the GL2 was compared with the classical *tf idf* with stemming, the relative difference was around 50% (0.2590 vs. 0.1708).

	Mean average precision	
\ Stemmer IR Model	None	Light
GL2	0.1783	**0.2590**
Okapi	**0.1841**	0.2541
LM	0.1795	0.2537
Lnu-ltc	0.1821	0.2345*
tf·idf	0.1479*	0.1708*
Difference %		+33.8%

Table 1. MAP of stemming approach using short queries (T) and various IR models

Stemming strategies need to be compared column by column. As a baseline, we used the IR performances obtained when ignoring the stemming procedure. After applying the light stemming, the performance was always statistically better (values underlined in Table 1) than those achieved when stemming was ignored. Moreover, as depicted in the last row, the mean difference over the baseline was 33.8%.

Mean values, as with other summary statistics, may hide irregularities between queries and thus it is always advisable to take a closer look at certain performance differences. Using the GL2 model, the number of queries resulting in better average precision (AP) after stemming was 68, while for the 25 other queries, the search system without stemming performed better. For six queries, the same AP was achieved by both search strategies (namely Topic #272 "Czech President's Background" with an AP: 0.1429, Topic #281 "Radovan Karadzic" with an AP: 0.2778, Topic #306 "ETA Activities in France" with an AP: 0.5, #324 "Supermodels", AP 0.0, #360 "Water on Mars" with an AP 0.81, and Topic #367 "East Timor Independence" with an AP: 0.95). In some cases, the stemmer removed the final suffix, as for example the words "background" (Topic #272), "activities" (Topic #306) or "France"

(Topic #306), or the last letter of the words "supermodels" (Topic #324), or "water" (Topic #360). Such stemming modifications did not have any effect on retrieval effectiveness and thus both strategies performed with equal effectiveness. Finally, in some cases the stemming had no effect, as in Topic #281 ("Radovan Karadzic") which had the identical query, with or without stemming.

The largest performance difference between an approach with and without stemming was achieved by Topic #279 ("Swiss referendums"), having six relevant items. After stemming, the AP was 0.9167 (relevant items ranked in positions 1, 2, 3, 4, 5, and 12) and only 0.2753 without stemming (relevant items in positions 8, 14, 18, 19, 30, 153). The plural form of the term "referendum" occurs only in 32 documents and thus cannot be very helpful in promoting relevant articles that contain the singular form. For this query, removing the plural suffix was clearly more effective. Of course we encountered the same difficulty with the second term "Швейцария" (Switzerland) which was not able to retrieve articles containing the adjective form ("Швейцарски"), when we ignored the stemming procedure.

5.2 Using Different Topic Formulations

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

For all these topic formulations, the GL2 probabilistic model performed the best, but the performance differences with the Okapi or the LM model were never statistically significant. When comparing the GL2 model with the vector-space approaches, performance differences were always significant (indicated by an "*").

Using the performance achieved by the shortest query formulation (T) as a baseline, the data depicted in Table 2 indicates that with the GL2 and Okapi models, including the descriptive part (TD), did not significantly improve IR performance. However, when including both the descriptive and narrative (TDN) parts, the MAP was always statistically significant as compared to the T formulation (values underlined).

	Mean average precision		
\ Stemmer	T	TD	TDN
\ mean query size	2.52	7.48	15.8
GL2	**0.2590**	**0.2826**	**<u>0.2994</u>**
Okapi	0.2541	0.2805	<u>0.2922</u>
LM	0.2537	<u>0.2822</u>	<u>0.2950</u>
Lnu-ltc	0.2345*	<u>0.2615</u>*	<u>0.2769</u>*
tf·idf	0.1708*	<u>0.1937</u>*	<u>0.2044</u>*
Difference %		+11.1%	+16.9%

Table 2. MAP of various topic formulations

As shown in Figure 2, the inclusion of the descriptive part (D) in the query generation may add related and pertinent terms such, as "frequent use" or "addiction" with Topic #255 ("Internet Junkies"). The second row in Table 2 shows the average num-

ber of distinct search terms per query, a value that increased from 2.52 for the shortest query formulation (T) to 7.48 for the title and descriptive parts (TD), and to 15.8 for the longest query formulation (TDN).

Although it is important to apply a statistical test, it is also important to inspect the actual data. Upon inspecting the differences between the Okapi and GL2 model using TDN query formulation, for example, the MAP values were 0.2922 and 0.2994 respectively, and thus the differences were quite small (0.0072 in absolute value, or 2.5%). Using the bootstrap test, the difference detected was not significant, due to the small performance differences of many queries. For 63 queries the GL2 obtained better AP, while for the 33 others the Okapi model performed better (for three queries, we obtained the same AP). Using the Sign-test (where only the direction difference was taken into account), the *p*-value would be 0.002879, indicating that the 63 "+" and 33 "-" were not simply the result of a random effect. Even though in this particular case both statistical tests based on different information did not agree, usually their conclusions tended to corroborate and lead to the same conclusion (Abdou & Savoy, 2006).

The largest differences between the T and TD query formulations were achieved with Topic #256 ("Creutzfeldt-Jakob Disease"), having two relevant items. With the shortest query formulation (T), the AP was 0.2551 and the relevant documents were ranked in positions 2 and 198. The TD query improved the AP (0.625) by ranking the relevant articles in positions 1 and 8. In this case, the T query was composed of two terms, namely "Болест" (disease, with a document frequency (*df*) of 1,118), and "Кройцфелд-Якоб" (*df*=5). This short request was not able to rank the second relevant document higher because it contained the form "Кройцфелд-Якобс" (with a final '-c'). For this request, the TD formulation was able to rank the relevant items higher in the output list, given increased number of terms in common with the query. For example, they included the terms "луда" (mad), and "крава" (cow). However other words included in the D part and that were not present in the pertinent articles did not hurt the ranking process (e.g., "Spongiform" occurred in a single document).

5.3 Another Stopword List and Stemmer

It should be noted that when developing our stopword list, we had to make certain arbitrary decisions as to whether or not we would include a particular form (Fox, 1990), (Savoy, 1999). Thus another stopword list could very well have achieved the same objective, namely to allow pertinent matches between search keywords and documents. For the Bulgarian language, such an alternative stopword list was suggested during the CLEF-2005 evaluation campaign. Listed under the heading "BTB", this list contains 804 forms and is available at www.bultreebank.org/resources/BTB-StopWordList.zip. Clearly it is longer than our list of 258 entries, but there are 176 terms (or 68%) common to the two lists. By contrast, commercial information systems tend to adopt a more conservative approach, using only a few stopwords. The DIALOG system for example uses only 9 items when indexing English documents (namely "an," "and," "by," "for," "from," "of," "the," "to," and "with") (Harter, 1986).

Table 3 lists the retrieval effectiveness of both stopword lists using either the short query formulation (T) or using both the title and descriptive sections (TD) of topic descriptions. As shown in Table 3, both stopword lists performed equally well. For example, using the GL2 model and with T query formulation, the difference between the two stopword lists is rather small (0.2590 vs. 0.2555 with an absolute difference of 0.0035, or 1.3%). A query-by-query analysis reveals that only three queries out of 99 resulted in an AP difference greater than 0.05. For 37 queries, our stopword list resulted in better AP, while for 38 others, the BTB stopword list performed better (for 24 queries, we obtained the same AP). Using the MAP achieved by our stopword list as baseline, the statistical test did not detect any significance differences between the performances achieved by both stopword lists.

	Mean average precision			
\ Stopword list	T	T (BTB)	TD	TD (BTB)
GL2	**0.2590**	**0.2555**	**0.2826**	0.2782
Okapi	0.2541	0.2539	0.2805	**0.2796**
LM	0.2537	0.2527	0.2822	0.2750
Lnu-ltc	0.2345*	0.2360	0.2615*	0.2616
tf·idf	0.1708*	0.1708*	0.1937*	0.1930*
Mean difference %		-0.0%		-0.0%

Table 3. MAP using two different stopword lists and topic formulations

Recently Nakov (2003) suggested a stemmer for the Bulgarian language, based on a large morphological dictionary (889,665 forms) and a learning algorithm. In this case, the machine learning process develops suffix removal rules in accordance with the part of speech class, a short remainder context (the ending for the proposed stem), and their frequency. In accordance with the recommended setting, we loaded 22,199 rules out of a total of 30,755 rules. In this case, the removal of suffixes is based on the longest possible rule and the stemmer may also remove certain derivational endings (e.g., as '-ment', '-ably', '-ship' in the English language). Moreover, while Nakov's approach takes numerous verb forms into account, the scope of the suggested light stemmer is limited to nouns and adjectives. Trying to remove most of the inflectional suffixes for a given language implies that numerous verb forms must be taken into account during the suffix removal process. Trying numerous suffixes may consequently impair overall effectiveness, as shown for other languages such as German, Portuguese and Hungarian (Savoy, 2006). An overall evaluation for the light and Nakov stemmers is listed in Table 4, under two different topic formulations.

	Mean average precision			
\ Stemmer	T (light)	T (Nakov)	TD (light)	TD (Nakov)
GL2	**0.2590**	**0.2655**	**0.2826**	**0.2800**
Okapi	0.2541	0.2584	0.2805	0.2642*
LM	0.2537	0.2629	0.2822	0.2677
Lnu-ltc	0.2345*	0.2421*	0.2615*	0.2651
tf·idf	0.1708*	0.1802*	0.1937*	0.2013*
Mean difference %		3.3%		-0.1%

Table 4. MAP using two different stemmers and topic formulations

The data in Table 4 indicates that MAP differences between the two stemmers are usually small. As mentioned previously, none of the performance differences can be viewed as statistically significant. For example, for the GL2 model and T query formulation, the light stemmer results in a MAP of 0.2590 vs. 0.2655 for Nakov's stemmer (absolute difference of 0.0065, or 2.4%). In this case Nakov's stemmer results in better AP for 52 queries, while the light approach performs better for 37 other queries (the same AP was obtained for the 10 remaining queries). An analysis of the largest AP differences between the two stemmers would provide us with a better understanding of their respective strengths and weakness.

The largest performance difference in favor of the light stemmer was obtained with Topic #320 ("Energy Crises" owning seven relevant documents). With the light stemmer, the AP is 0.6167 (relevant items ranked in positions 1, 2, 3, 8, 15, 24 and 30) while with Nakov's approach this query achieved an AP of 0.008 (relevant items ranked in positions 195, 201, 230, 273, 714, 914 and 1230). From the topic's title "енергийни кризи", the light stemmer produced the query "енергийн криз" (it removed the last letter '-и' for both terms) while the corresponding query based on the Nakov's stemmer was "енерги кризи". The noun "криза" is the singular form of "кризи" (crises). The singular form appears in all relevant documents and the stem "криз" as produced by the light stemmer is useful for extracting it. On the other hand, the second term "енергийни" was the adjective plural form ('-и'), from the term "енергия" (energy) and it is used in the sense "of the energy". With the Nakov's approach, the resulting stem "енерги" is correct but it appears in 2,029 documents, while the longest form produced by the light stemmer occurs in only 1,166 documents, including all relevant articles.

On the other hand, Nakov's stemmer resulted in the largest performance difference with Topic #296 ("Public Performances of Liszt" appearing in one relevant article). The GL2 model using the light stemmer achieved a moderate AP of 0.05 (the relevant item appears in the 20th rank) and with Nakov's stemmer, the AP was 0.5 (the single relevant item appeared in the second position). This difference can be explained in the following way. In Bulgarian, the topic title is "Публични изпълнения на творби на Лист". With Nakov's algorithm, the same plural form "творби" appears both in the query and in the relevant document. Moreover, the personal name ("Лист" – Liszt) appears only in 260 documents when using Nakov's stemmer as compared to 1,090 with the light stemmer. In the latter case, the search keyword "лист" (also meaning "leaf" in Bulgarian) was conflated with the form "листа" ("list", "menu") using the light stemmer.

5.4 Automatic Decompounding

As a third indexing strategy, we decided to automatically decompound Bulgarian compound words (e.g., "радиоапарат" = "радио" (radio) + "апарат" (receiver)) according to our decompounding algorithm (all details are given in (Savoy, 2004)). In German compound constructions are frequent, and we found decompounding them may have a positive impact (Braschler & Ripplinger, 2004) on retrieval effectiveness. As shown in Table 5, our automatic decompounding approach slightly increased the

mean query size from 2.52 to 2.87 for the T query formulation. The IR performances stayed relatively the same. With word-based queries as a baseline, we found no statistically significant difference.

	Mean average precision			
\ Indexing \ mean size	T (word) 2.52	T (decomp) 2.87	TD (word) 7.48	TD (decomp) 8.36
GL2	**0.2590**	**0.2633**	**0.2826**	**0.2809**
Okapi	0.2541	0.2505*	0.2805	0.2735
LM	0.2537	0.2482	0.2822	0.2707
Lnu-ltc	0.2345*	0.2434*	0.2615*	0.2690
tf·idf	0.1708*	0.1820*	0.1937*	0.1995*
Difference %		0.7%		-1.0%

Table 5. MAP for various topic formulations and indexing strategies

A query-by-query analysis reveals that by using the GL2 model (T queries), 74 queries out of 99 resulted in absolute AP differences of less than 0.05 (92 out of 99 had an absolute difference of less than 0.1). An analysis of the largest AP differences between the two indexing schemes would thus provide us a better understanding of their respective strengths and weakness.

With T queries and the GL2 model, the decompounding indexing strategy resulted in better AP (0.5485) for Topic #373 ("Hungarian-Bulgarian Relationships", with 44 relevant items). With the word-based indexing approach, we obtained an AP of 0.0232 with the query for the words {"българо-унгарск" and "връзк"}. The decompounded query contained three stems, namely {"българ", "унгарск" and "връзк"}. In the first case, the order was imposed, and country names had to be joined by a dash. In the relevant articles, these names did not always appear in this order and when they occurred together in the same sentence, they were not always adjacent.

For the GL2 model and Topic #322 ("Atomic Energy" or "Атомната енергия", returning four relevant items), the word-based indexing approach produced better AP (0.6167) than did the decompounding approach (AP = 0.0991). The underlying query was however identical {the stems were "атомн" and "енерг"}, but due to the decompounding scheme the stem "energy" appeared in more documents. Thus the *idf* value for this search term was lower, and the resultant ranking was less effective than that of word-based indexing.

5.5 N-gram Indexing Strategy

As a language-independent indexing strategy, we might apply an *n*-gram character tokenization approach in which each surface form is subdivided into sequences composed of *n* consecutive letters (McNamee & Mayfield, 2004). For example, the form "computers" will generate the following 4-grams: "comp", "ompt", "mput", "pute", "uter" and "ters". This indexing approach is usually relatively effective across different languages and for languages such as Korean or Chinese it could be the best indexing strategy (Abdou & Savoy, 2006). Moreover, such an approach does not require the application of a stemming process before segmenting the surface forms. On the

other hand, the *n*-gram approach requires a larger inverted file and tends to slow the search process.

During the CLEF-2005 evaluation campaign (Peters *et al*., 2006), McNamee (2006) suggests that this indexing scheme be used for the Bulgarian language. In this case, the best performance was achieved using a 4-gram indexing strategy (MAP 0.3203 vs. 0.2768 for the 5-gram scheme (McNamee, 2006)). We used the same *n* value in our experiments where the most frequent *n*-grams were also removed, based on our suggested stopword list (see Table A.1. in the Appendix). The mean average precision of this indexing strategy is depicted in Table 6, together with the word-based approach.

	Mean average precision			
\ Indexing	T (word)	T (4-gram)	TD (word)	TD (4-gram)
GL2	**0.2590**	0.2421*	**0.2826**	0.2630*
Okapi	0.2541	**0.2560**	0.2805	**0.2771**
LM	0.2537	0.2325*	0.2822	0.2405*
Lnu-ltc	0.2345*	0.2122*	0.2615*	0.2573*
tf·idf	0.1708*	0.1672*	0.1937*	0.1856*
Mean difference %		-5.2%		-5.7%

Table 6. MAP for various topic formulations and indexing strategies

As shown in Table 6 in which the word-based (light stemming) is used as baseline, the performance differences were usually not statistically significant. The only exception to this finding was for the LM model with TD queries, where the difference 0.2822 vs. 0.2405 could be viewed as statistically significant. Moreover, retrieval performance usually tended to be slightly better when using a word-based indexing approach. For example, with the GL2 model and T queries, the MAP was 0.2590 for the word-based and 0.2421 for the 4-gram indexing scheme (a relative difference of 7%).

A query-by-query analysis revealed that the word-based indexing approach (GL2 model, T queries) produced a better AP for 53 queries, while for 45 other queries the 4-gram indexing strategy performed better (the same AP was obtained with Topic #301 "Nestlé Brands"). These values tended to explain why the differences between the two indexing strategies were usually not statistically significant.

Upon examining larger differences, we found that Topic #320 ("Energy Crises", seven relevant items) resulted in better retrieval performance than did the word-based approach. In this case, we achieved an AP of 0.6167 (relevant items ranked in positions 1, 2, 3, 8, 15, 24 and 30) vs. 0.0581 for the 4-gram scheme (relevant items ranked in positions 5, 33, 59, 91, 195, 358 and 767). As explained previously, it is important to maintain the longest form (the adjective, "енергийни") in the topic. Generating a 4-gram from this search term (e.g., "енер," "нерг" or "ерги") will also match the noun ("енергия"), and thus will extract many non-relevant documents from the corpus.

On the other hand, Topic #255 ("Internet Junkies" having three relevant items) with the 4-gram approach obtained an AP of 0.5 while the word-based model produced only an AP of 0.1139. With the *n*-gram scheme, the two relevant documents were found in the first and fourth positions, while for the word-based approach these

two articles were ranked in positions 3 and 239 respectively. The title of this topic was written as "пристрастяване (addiction) към (to) интернет (internet)". Both main search keywords were found in the first relevant document, thus explaining their high position under both indexing schemes. For the second relevant article, only the search term "internet" appeared as it (as well as in 1,217 other documents). The search form for "addiction" (written as "пристрастяване" in the topic) differed in the second relevant item where the form "пристрастеността" appeared (the corresponding document appears in Figure 1). Because the *n*-gram indexing strategy is more robust in the event of slight orthographic or morphological variations, the 4-gram indexing strategy was nevertheless able to find six matches (underlined in the previous example) between the query term and the form used in the document. This fact means it is possible for the search engine to rank this particular document higher on the result list, more precisely in the fourth position in the current case.

5.6 Hard Topics

Until now, the mean was the only single measure used for any given search model, under a specific condition. Although this measure has the advantage of summarizing sample values into one number, it hides individual performances. For the shortest query formulation (T) and the GL2 model, Figure 3 indicates the distribution of individual query performances. In this figure, the MAP (0.2590) is indicated by a dashed line (standard deviation 0.2424). For this right-skewed distribution, the minimum AP was 0.0 (Topic #324 "Supermodels") while the maximum AP was 0.95, obtained by Topic#367 "East Timor Independence".

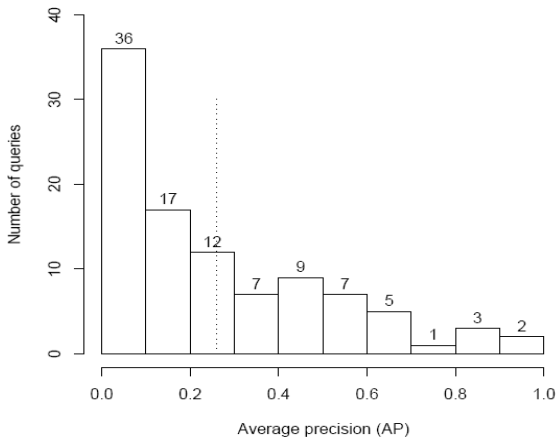

Figure 3. Distribution of 99 average precision measures using GL2 model and title-only queries

Under this condition, Topic #324 "Supermodels" proves to be the most difficult topic (15 relevant documents). Using only the topic title, the query response was limited to one term occurring in five documents, all of which were judged as non-relevant. Even when including the descriptive part (containing the related term "top

models"), the request was still difficult (AP = 0.0015). All IR models failed to retrieve one relevant item in the top ten results. The relevant articles usually cited the name of a model ("C. Crawford", "N. Campbell") or used synonyms in Bulgarian language (e.g., "mannequin" meaning also "top model" in Bulgarian).

Another interesting case is Topic #297 "Expulsion of Diplomats", which had five relevant documents. With T query formulation, this request obtained an AP of 0.0525 (GL2 model, relevant items ranked in positions 17, 29, 30, 248 and 272). However, using the same topic formulation with the classical *tf idf* model, we obtained an AP of 0.1563. The same query produced clearly two different rankings, but in this case the classical *tf idf* performed better (relevant items in positions 4, 17, 21, 118 and 121). The relevant documents had only one term in common with the query, namely the term "diplomats", occurring in 1,027 articles. The second query term "expulsion" appeared in 495 documents, thus having a higher *idf* value. Although three documents containing both search terms ("expulsion" and "diplomat") were ranked higher in the result list they were judged not relevant. In these three articles, the search terms did not appear in the same sentence and were not related (e.g., one document dealt with the expulsion of Saddam and the arrival of American diplomats).

6 Conclusion

In this paper we describe the most significant linguistic features of the Bulgarian language, from an IR perspective. Belonging to the Slavic family, this language has a rich morphology and includes the use of suffixes to denote the definite article (the). Using a test collection extracted from the CLEF 2005 & 2006 test suites containing 99 requests, we evaluate three probabilistic and two vector-space models. When using the title-only queries, the GL2 model derived from the *Divergence from Randomness* (Amati & van Rijsbergen, 2002) paradigm tends to result in the most effective retrieval, under a variety of conditions. However, performance differences between this IR model and the Okapi or the language model usually tend to be statistically non-significant. When comparing the GL2 model with other vector-space models, the MAP differences are usually significant.

When topic size increases, so does retrieval effectiveness. As shown in Table 2, the GL2 model having short topic formulations (2.52 search keywords per query on average) produces a MAP of 0.2590 while for the model having longer topic formulations (in average 7.48 terms per query) the MAP increases to 0.2826 (enhancement of around 9%).

This paper examines a stopword list composed of 258 entries (forms depicted in Table A.1) and compares it with another stopword list composed of 804 forms. The data depicted in Table 3 reveals that performance differences between these two lists are small and insignificant. Also described in this paper is a light stemming strategy used to remove only inflectional suffixes (feminine and plural forms, and definite articles). When compared to IR models that ignore the stemming procedure, the mean difference is around +30% (see Table 1). We then evaluate a more complex Bulgarian stemmer based on a large dictionary that removes inflectional and certain derivational suffixes. Upon comparing the performances achieved by both stemmers

(see Table 4), we do not find any statistically significant differences. Furthermore, various query-by-query analyses reveal situations in which a one stemming strategy is better than another.

The word-based indexing strategy results in slightly better retrieval effectiveness than does the indexing method, comprising a decompounding stage (see Table 5) or is clearly better than an indexing strategy based on a *n*-gram approach (see Table 6). Finally, the distribution of the AP for the 99 queries (GL2, T queries) is found to vary, and our analysis of some of the most difficult topics explains why the search system based on our stopword list and light stemmer was not able to rank a single relevant retrieved item in the top ten results.

Acknowledgments
This research was supported in part by the Swiss NSF under Grants #200020-103420 and #200021-113273.

References

Abdou, S., Ruck, P., & Savoy, J. (2006). Evaluation of stemming, query expansion and manula indexing approaches for the Genomic task. In *Proceedings of TREC-2005*. NIST Publication #500-266, Gaithersburg (MA).

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

Ahlgren, P., & Kekäläinen, J. (2007). Indexing strategies for Swedish full text retrieval under different user scenarios. *Information Processing & Management*, 43(1), 81-102.

Ahmad, F., Yusoff, M., & Sembok, T.M.T. (1996). Experiments with a stemming algorithms for Malay words. *Journal of the American Society for Information Science*, 47(12), 909-918.

Alkula, R. (2001). From plain character strings to meaningful words: Producing better full text databases for Finnish with morphological analysis software. *IR Journal*, 4 (3-4), 195-208.

Allières J. (2000). Les langues de l'Europe. Presses Universitaires de France, Paris.

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.

Asian, J., Williams, H.E., & Tahaghoghi, S.M.M. (2004). A testbed for Indonesian text retrieval. In *Proceedings of the ADCS*. Melbourne, 55-58.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. The ACM Press, New York.

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*. NIST Publication #500-236, Gaithersburg (MA), 25-48.

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

Chen, C., & Gey, F. (2003). Building an Arabic stemmer for Information retrieval. In *Proceedings of TREC-2002*. NIST Publication #500-251, Gaithersburg (MA), 631-640.

Crawley, M.J. (2005). *Statistics. An Introduction using R*. John Wiley & Sons, Chichester.

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*, LNCS #3237, Springer-Verlag, Berlin, 220-235.

Ekmekçioglu, F.C., & Willett, P. (2000). Effectiveness of stemming for Turkish text retrieval. *Program*, 34(2), 195-200.

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

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

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

Hedlund, T., Pirkola, A., & Järvelin, K. (2001). Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. *Information Processing & Management*, 37(1), 147-161.

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.

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

Kraaij, W., & Pohlman, R. (1996). Viewing stemming as recall enhancement**.** In *Proceedings of ACM-SIGIR*. Tempere, 40-48.

Kraaij, W., Westerveld, T., & Hiemstra, D. (2002). The importance of prior probabilities for entry page search**.** In *Proceedings of ACM-SIGIR*. Tempere, 27-34.

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

Nakov, P. (2003). BulStem: Design and evaluation of inflectional stemmer for Bulgarian**.** In *Proceedings of Workshop on Balkan Language Resources and Tools*. Thessaloniki.

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

McNamee, P. (2006). Exploring new languages with HAIRCUT at CLEF-2005. In *Accessing Multilingual Information Repositories*. LNCS #4022, Spinger-Verlag, Berlin, 155-164.

Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M. & Magnini, B. (Eds.) (2005). *Multilingual Information Access for Text, Speech and Images*. LNCS #3491. Springer-Verlag, Berlin, 2005.

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, Spinger-Verlag, Berlin.

Popovic, M. & 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(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. (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., & Rasolofo, Y. (2003). Report on the TREC-11 experiment: Arabic, named page and topic distillation searches. In *Proceedings of TREC-2002*, NIST publication #500-251, Gaithersburg (MD), 765-774.

Savoy, J. (2004). Report on CLEF 2003 monolingual tracks. In *Comparative Evaluation of Multilingual Information Access Systems*, LNCS #2785, Springer, Berlin, 322-336.

Savoy, J. (2005). Comparative study of monolingual and multilingual search models for use with Asian languages. *ACM Transactions on Asian Languages Information Processing*, 4(2), 163-189.

Savoy, J. (2006). Light stemming approaches for the French, Portuguese, German and Hungarian languages. In *Proceedings ACM-SAC*, Dijon, 1031-1035.

Savoy, J. (2007). Searching strategies for the Hungarian language. *Information Processing & Management*, to appear.

Schinke, R., Greengrass, M., Robertson, A.M., & Willett, P. (1998). Retrieval of morphological variants in searches of Latin text databases. *Computers and the Humanities*, 31(1), 409-432.

Sproat, R. (1992). *Morphology and Computation*. The MIT Press, Cambridge.

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*. LNCS #3237, Springer-Verlag, Berlin, 286-300.

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:

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

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

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

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

а	добро	ме	първата
автентичен	добър	между	първи
аз	докато	мек	първо
ако	докога	мен	пъти
ала	дори	месец	равен
бе	досега	ми	равна
без	доста	много	с
беше	друг	мнозина	са
би	друга	мога	сам
бивш	други	могат	само
бивша	е	може	се
бившо	евтин	мокър	сега
бил	едва	моля	си
била	един	момента	син
били	една	му	скоро
било	еднаква	н	след
благодаря	еднакви	на	следващ
близо	еднакъв	над	сме
бъдат	едно	назад	смях
бъде	екип	най	според
бяха	ето	направи	сред
в	живот	напред	срещу
вас	за	например	сте
ваш	забавям	нас	съм
ваша	зад	не	със
вероятно	заедно	него	също
вече	заради	нещо	т
взема	засега	нея	тази
ви	заспал	ни	така
вие	затова	ние	такива
винаги	защо	никой	такъв
внимава	защото	нито	там
време	и	нищо	твой
все	из	но	те
всеки	или	нов	тези
всички	им	нова	ти
всичко	има	нови	
всяка	имат	новина	то
във	иска	някои	това
въпреки	й	някой	тогава
върху	каза	няколко	този
г	как	няма	той
ги	каква	обаче	толкова
главен	какво	около	точно
главна	както	освен	три
главно	какъв	особено	трябва
глас	като	от	тук
го	кога	отгоре	тъй
година	когато	отново	тя
години	което	още	тях
годишен	които	пак	у
д	кой	по	утре
да	който	повече	харесва
дали	колко	повечето	хиляди

два	която	под	ч
двама	къде	поне	часа
двамата	където	поради	че
две	към	после	често
двете	лесен	почти	чрез
ден	лесно	прави	ще
днес	ли	пред	щом
дни	лош	преди	юмрук
до	м	през	я
добра	май	при	як
добре	малко	пък	

Table A.1. Our Bulgarian stopword list

```
BulgarianLightStemmer (input/output: word)
  i := length(word);
  if (i > 5) {
     if (word ends with « -ища ») { remove « -ища »; return }
     };
  if (i < 4) { return };                                    # word too short
  RemoveArticle(word);
  RemovePlural(word);
  i := length(word);
  if (i > 3) {
     if (word ends with « -я ») { remove « -я »; i-- };        # normalize adjective
     if (word ends with « -[aoe] ») { remove « -[aoe] »; i-- };  # final “a”, “o” or “e”
     if (word ends with « -ен  ») { replace by « -н »; i-- }; # rewriting rule
     };
  if (i > 4) {
     if (word ends with « -ен  ») { replace by « -н »; i-- }; # rewriting rule
     }
  if (i > 5) {
    if (word ends with « -..ъ.  ») { remove « -ъ »; i-- };     # remove “ъ” near the end
  }
  return;

RemoveArticle(input/output: word)       # Mainly remove the definite article
  i := length(word);
  if (i > 6) {
     if (word ends with « -ият ») { remove « -ият »; return }; # for adjectives
     }
  if (i > 5) {
     if (word ends with « -ът ») { remove « -ът  »; return };    # masculine
     if (word ends with « -то ») { remove  « -то »; return };    # neutral
     if (word ends with « -те ») { remove « -те »; return };     # plural
     if (word ends with « -та ») { remove « -та »; return };     # feminine
     if (word ends with « -ия ») { remove « -ия »; return };     # for adjectives
     }
  if (i > 4) {
     if (word ends with « -ят ») { remove « -ят »; return };     # masculine
     }
  return;
```

```
RemovePlural(input/output: word)              # Mainly remove the plural suffix
  i := length(word);
  if (i > 6) {
     if (word ends with « -овци ») { replace by « -о »; return };   # for adjectives
     if (word ends with « -ове ») { remove « -ове  »; return };     # masculine
     if (word ends with « -еве ») { replace by « -й »; return };    # masculine
     }
  if (i > 5) {
     if (word ends with « -ища ») { remove « -ища  »; return };   # for adjectives
     if (word ends with « -та ») { remove « -та  »; return };         # feminine
     if (word ends with « -ци ») { replace by « -к »; return };      # rewriting
     if (word ends with « -зи ») { replace by « -г »; return };
     if (word ends with « -..е.и ») { replace by « -..я. »; return }; # rewriting
     }
  if (i > 4) {
     if (word ends with « -си ») { replace by « -х »;return };
     if (word ends with « -и ») { remove « -и »; return };            # other plural
     }
  return;
```

Table A.2. Our light Bulgarian stemmer