Information Processing and Management 45 (2009) 714–720

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Indexing and stemming approaches for the Czech language

Ljiljana Dolamic, Jacques Savoy *

*Computer Science Department, University of Neuchatel, 2009 Neuchâtel, Switzerland*

ARTICLE INFO

*Article history:*
Received 12 December 2008
Received in revised form 4 June 2009
Accepted 14 June 2009
Available online 16 July 2009

*Keywords:*
Czech language
Stemming
Evaluation
Slavic languages

ABSTRACT

This paper describes and evaluates various stemming and indexing strategies for the Czech language. Based on Czech test-collection, we have designed and evaluated two stemming approaches, a light and a more aggressive one. We have compared them with a no stemming scheme as well as a language-independent approach (*n*-gram). To evaluate the suggested solutions we used various IR models, including Okapi, *Divergence from Randomness* (DFR), a statistical language model (LM) as well as the classical *tf idf* vector-space approach. We found that the *Divergence from Randomness* paradigm tend to propose better retrieval effectiveness than the Okapi, LM or *tf idf* models, the performance differences were however statistically significant only with the last two IR approaches. Ignoring the stemming reduces generally the MAP by more than 40%, and these differences are always significant. Finally, if our more aggressive stemmer tends to show the best performance, the differences in performance with a light stemmer are not statistically significant.

## 1. Introduction

Slavic languages dominate in Eastern and Central Europe (e.g., Serbo-Croatian, Russian, Polish, Bulgarian or Czech), and their distinct linguistics features (e.g., the use the various grammatical cases marked by suffixes) must be taken into account in an efficient IR system (Manning, Raghavan, & Schütze, 2008). However, the IR community has only a very small number of test-collections available for this family of languages. As an exception, we could mention the Bulgarian language for which the last two CLEF evaluation campaigns have produced a test-collection (Peters et al., 2008). Unlike the morphology of other Slavic languages however, the grammatical cases are usually not explicitly indicated by a given suffix in the Bulgarian morphology (with the exception of the infrequent vocative case). Thus, experiments drawn for this language cannot be applied directly to other Slavic languages.

The CLEF 2007 campaign (Dolamic & Savoy, 2008) produces also a shorter test-collection for the Czech language, and the main objective of this paper is to describe the main morphological difficulties when working with this language. We also proposed and evaluated a suitable stemmer for this Slavic language. In IR it is assumed that applying a stemmer will conflate several word variants into the same stem, and thus improve the pertinent matching between query and document surrogates. For example, when a query contains the word “horse,” it seems reasonable to also retrieve documents containing the related word “horses.” Moreover, stemming procedures will also reduce the size of inverted files.

When designing a stemmer, we may create a “light” suffix-stripping procedure by removing only the morphological inflections by conflating the singular and plural word forms (e.g., “door” and “doors”) or feminine and masculine variants (e.g., “actress” and “actor”) to the same stem. More sophisticated approaches will remove derivational suffixes (e.g., “enhance” and “enhancement”) use to generate a new part-of-speech word from a given stem. Even though a different stemming procedures have been suggested for various European languages (e.g., Snowball project, CLEF, TREC and NTCIR

* Corresponding author.
*E-mail addresses:* Ljiljana.Dolamic@unine.ch (L. Dolamic), Jacques.Savoy@unine.ch (J. Savoy).

0306-4573/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.06.001

campaigns (Harman, 2005; Peters et al., 2008), no stemming algorithm with its evaluation is available for the Czech language.

The rest of this paper is organized as follows. Section 2 describes different stemming approaches while Section 3 depicts the main characteristics of our test-collection. Section 4 briefly describes the IR applied during our experiments. Section 5 evaluates the performance of various IR models, in addition to two stemming approaches for the Czech language. The main findings of this paper are presented in the conclusion.

## 2. Related work

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; 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"). 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. *However, applying an algorithmic stemmer does not guarantee that we always obtain either the correct stem or an existing word in the corresponding language*.

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, Laurikkala, Järvelin, & Juhola, 2004), 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 and Airo (2006) indicate for example that in theory Finnish nouns have around 2000 different forms, yet in actual collections the occurrence of most of these forms is rare. As a matter of fact in Finnish, 84–88% 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 and 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.

Few stemming procedures[1] have been suggested for European languages other than English. The proposed stemmers usually pertain to the most popular languages (Peters et al., 2008; Tomlinson, 2004) and some of them, like the Finnish language, seem to require a deeper morphological analysis (Korenius et al., 2004) to achieve good retrieval performance.

Algorithmic stemmer ignores word meanings and tends to make errors, usually due to over-stemming (e.g., "organization" is reduced to "organ") or to under-stemming (e.g., "European" and "Europe" do not conflate to the same root). Most of the studies so far have been involved in evaluating IR performance for the English language, while studies on the stemmer performance for less popular languages are less frequent. For example, Tomlinson (2004) evaluated the differences between Porter's stemmer strategy (Porter, 1980) and lexical stemmers (based on a dictionary of the corresponding language) for various European languages. For the Finnish and German languages, lexical stemmer tends to produce statistically better results, while for seven other languages performance differences were insignificant.

Based on these facts, the rest of this paper will address the following questions: (1) Does stemming affect IR performance for the Czech language (and to which extent)? (2) For this language, is a light stemming approach more effective than more complex suffix-stripping algorithms?

## 3. Czech morphology and stemming strategies

When creating stemming procedures for the Czech language we adopted the same strategy as for the other European languages for which we have created stemmers during the past years. We believe that effective stemming should focus mainly on nouns and adjectives (sustaining most of the meaning of a document), thus ignoring numerous verb forms (tending to generate more stemming errors when taken into account).

[1] Freely available at the Web site http://www.snowball.tartarus.org/ or http://www.unine.ch/info/clef/.

**Table 1**
Some statistics from the Czech test-collection.

| Size | # Docs | # Docs mean terms | # Queries | # Rel. docs/query |
|---|---|---|---|---|
| 178 MB | 81,735 | 212.6 | 50 | 15.24 |

The Czech language belongs to the Slavic languages and is written, as for example the Polish language with our Latin alphabet with the addition of eight diacritics used to specify a particular pronunciation (e.g., ‘č’, ‘ň’, ‘ř’, ‘ď’, ‘ť’). As with the Latin or the German languages, the Czech and usually other Slavic languages use various grammatical cases marked by suffixes (e.g., the noun “city” in Russian could be written as “город” (nominative), “города” (genitive) or “городе” (locative)). These linguistic elements indicate that Czech inflections are more complex than the English ones which are mainly limited to the final ‘-s’.[2]

All nouns in the Czech language belong to the three distinct genders (masculine, feminine, or neutral). Moreover, all nouns are declined both in number (singular, plural)[3]; and using seven grammatical cases (nominative, genitive, dative, accusative, vocative, locative, and instrumental), with very few exceptions (a handful of indeclinable borrowed words). Each combination gender-case has its own set of characteristic paradigms, including hard-stem types, soft-stem types, and special types. For example, masculine noun “muž” (husband) appears as such in the nominative case singular, but varies in other cases “muže” (genitive, accusative), “mužovi” (dative, locative), “muže” (accusative), “muži” (vocative) or “mužem” (instrumental) with plural forms of this noun being “mužové,” mužů,” “mužům,” “muže,” “mužích,” and “muži”. From this example, we can see that the suffix denoting a case could be ambiguous in the sense that the same suffix may appear in other cases (“muže” could be the accusative or genitive singular form). Moreover, the stem (e.g., “muž” in our case) does not change after adding the appropriate suffix (unlike other languages like Finnish (Korenius et al., 2004)). Although this phenomenon can also occur in the Czech language, it is less frequent that in other languages. Finally, it is important to know that suffixes denoting cases occur also with proper names (e.g., with Paris, “Paříž” (nominative), “Paříže” (genitive), “Paříži” (dative), or with Ann, “Anna” (nominative), “Anny” (genitive), and “Ann” (dative)). *It is also important to notice that the stemming unlike lemmatization doesn't not always produce result with a correct lexical meaning (e.g., neuter noun “moře” (sea) and its different forms “moři” (dative), “mořem” (instrumental) conflate into “moř”, the corresponding stem that does not appear as it in the dictionary*).

As with many languages, the suffixes assigned to adjectives agree with the attached noun in case, gender and number. These language characteristics result in large number of suffixes being added to adjectives compared to other languages like German (having a rather limited set of suffixes (e.g., ‘-en’, ‘-es’)). Our stemmer denoted as “light” contains 52 rules for removing these grammatical case endings from nouns and adjectives (inflectional suffixes only). *A complete description of this stemmer is given in the Appendix. In the case of part-of-speech other than nouns and adjectives sharing the same set of suffixes (this being rather rare in the Czech language), they will also be removed. In such cases, the suggested strategy will certainly produce an incorrect stem. However defining the POS of each surface word is the first step of a lexical stemmer. Their use depends also 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). On the other hand, light algorithmic stemmers have shown to be effective for different European languages* (Savoy, 2006).

Derivational Czech morphology is accomplished by means of prefixation and suffixation of a stem, a usual construction with the Indo-European languages. Usually, the part-of-speech of the stem changes after adding a suffix (e.g., ‘-ial’ in “commerce” and “commercial”). In our work we addressed only suffixes because adding a prefix usually changes more the original meaning of a stem (e.g., “prehistory” vs. “historic” from the stem “history”). In the Czech language, derivational suffixes are added before case endings. We designed and implemented a more aggressive stemmer denoted “aggressive” in this paper which, besides removing inflectional suffixes, removes certain frequent derivational suffixes as for example (e.g., “klavír” (piano) → “klavírista” (pianist)). Both suggested stemmers address other morphological characteristics of the Czech language as fleeting ‘e’ (e.g. “zámek” (lock, nominative sing.) → “zámku” (genitive, dative, vocative, and locative sing.)) or consonant alternations (e.g. “ruka” (hand, nominative sing.) → “ruce” (dative and locative sing.)). Such irregularities, also present in the English language, are usually integrated to smooth the pronunciation.

Finally, 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 Czech language, the suggested stopword list contains 467 forms (determinants, prepositions, conjunctions, pronouns, and some very frequent verb forms). *In the process generating this stopword list we have followed the guidelines suggested by* Fox (1990). Both stemmers and the suggested stopword list for the Czech language are freely available at http://www.unine.ch/info/clef/.

[2] As for other natural languages, the English knows exceptions such as “mouse” and “mice” or the “'s” in “Paul's book” to denote the genitive case in some circumstances.

[3] As for other natural languages, some words occur only in singular or plural form (e.g., “nůžky”, scissors).

## 4. Test-collections

The evaluations reported in this paper were based on the Czech collection built during the CLEF 2007 evaluation campaign. This corpus consists of newspaper articles extracted from the *Mladá fronta Dnes* (year 2002) and *Lidové Noviny* (year 2002) newspapers. A typical document begins with a short title (tag <title>), usually followed by the first paragraph under the <headings> tag, and finally the body (<text> tag). As shown in Table 1, the mean number of indexing terms per article is around 212.6 while the whole corpus contains 81,735 articles.

The topics available covered various subjects (e.g., "NATO Summit Security," "Human cloning," "VIP Divorces") including both regional ("Kostelic Olympic Medals") and more international coverage ("Causes of Air Pollution"). Topics #411 ("Best Picture Oscar") or #413 ("Reducing Diabetes Risk") owns the smallest number of pertinent articles (2) while Topic #415 ("Drug Abuse") has the greatest number of correct answers (47).

Based on the TREC model, each topic was structured into three logical sections comprising a brief title (examples given upper), a one-sentence description, and a narrative part specifying the relevance assessment criteria. In our experiments, we used only the title part of the topic formulation in order to reflect more closely queries sent to commercial search engines. Using only the title section, queries had a mean size of 2.98 search terms.

Finally, since the title part of the request "Cosmetic procedures" was corrupted in the original topic formulation (replaced by the narrative part of the previous topic) we changed this topic title part into "kosmetický procedury" (the Czech translation of the corresponding English version).

## 5. IR models

To evaluate our proposed two stemming approaches with respect to various IR models, first we used the classical *tf idf* model wherein the weight attached to each indexing term was the product of its term occurrence frequency ($tf_{ij}$ for indexing term $t_j$ in document $d_i$) and the logarithm of its inverse document frequency ($idf_j$). To measure similarities between documents and the request, we computed the inner product after normalizing (cosine) the indexing weights (Manning et al., 2008).

To complement this vector-space model, we have implemented probabilistic models, such as the Okapi (or BM25) approach (Robertson, Walker, & Beaulieu, 2000), and one model derived from *Divergence from Randomness* (DFR) paradigm (Amati & van Rijsbergen, 2002) wherein two information measures formulated below are combined:

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

in which $\mathrm{Prob}^1_{ij}$ is the pure chance probability of finding $tf_{ij}$ occurrences of the term $t_j$ in a document. On the other hand, $\mathrm{Prob}^2_{ij}$ 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.

To model these two probabilities, we used the $I(n_e)$C2 model based on the following estimates:

$$\begin{aligned}
&\mathrm{Prob}^1_{ij} = \left(\frac{n_e + 0.5}{n+1}\right)^{tfn_{ij}} \\
&\text{and } \mathrm{Prob}^2_{ij} = 1 - \left(\frac{tc_j + 1}{df_j \cdot (tfn_{ij} + 1)}\right) \\
&\text{with } tfn_{ij} = tf_{ij} \cdot \ln\left(1 + \frac{c \cdot mean\ dl}{l_i}\right) \text{and } n_e = n \cdot \left(1 - \left(\frac{n-1}{n}\right)^{tc_j}\right)
\end{aligned} \tag{2}$$

where $tc_j$ is the number of occurrences of term $t_j$ in the collection, $df_j$ indicates the number of documents in with the term $t_j$ occurs, $n$ the number of documents in the corpus, $l_i$ the length of document $d_i$, *mean dl* (=212), the average document length, and $c$ a constant (fixed empirically at 1.5).

Finally, we also used an approach based on a language model (LM) (Hiemstra, 2000), known as a non-parametric probabilistic model. Various implementations and smoothing methods might also be considered within this language model paradigm. In this paper we adopted a model proposed by Hiemstra (2002, 2002) as described in Eq. (3) using the Jelinek-Mercer smoothing (Zhai & Lafferty, 2004), a combination of an estimate based on document ($P[t_j|d_i]$) and one based on the whole corpus ($P[t_j|C]$).

$$\begin{aligned}
&\mathrm{Prob}[q_i|q] = \mathrm{Prob}[d_i] \cdot \Pi_{t_j \in Q}[\lambda_j \cdot \mathrm{Prob}[t_j|d_i] + (1 - \lambda_j) \cdot \mathrm{Prob}[t_j|C]] \\
&\text{with } \mathrm{Prob}[t_j|d_i] = \left(\frac{tf_{ij}}{1_i}\right) \\
&\text{and } \mathrm{Prob}[t_j|C] = \left(\frac{df_j}{lc}\right) \quad \text{with } lc = \sum_{k=1}^{t} df_k
\end{aligned} \tag{3}$$

where $\lambda_j$ 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 size of the underlying corpus $C$.

## 6. Evaluation

In order to measure retrieval performance, we have adopted the mean average precision (MAP) computed by trec_eval (Buckley & Voorhees, 2005) based on maximum of 1000 retrieved items. To statistically determine whether or not a given search strategy is statistically better than another, we have applied the bootstrap methodology (Savoy, 1997), with the null hypothesis $H_0$ stating that both retrieval schemes produce similar performance. In the experiments presented in this paper statistically significant differences were detected by a two-sided test (significance level $\alpha = 5\%$). Such a null hypothesis would be accepted if two retrieval schemes returned statistically similar means, otherwise it would be rejected.

### *6.1. IR models evaluation*

Given the methodology previously described, Table 2 depicts the MAP using three stemming approaches with four IR models. In the last column we have also included a language-independent indexing approach based on 4-gram (McNamee and Mayfield (2004). Under this indexing scheme, words are decomposed by overlapping sequences of four letters (this value of 4 was selected because it produced the best IR performance). For example, the sequence "prime minister" generates the following 4-grams {"prim," "rime," "mini," "inis," ..., "ster"}. In the Table 2, statistically significant differences compared to the best performing model (depicted in italic) are marked with "*".

Finally, we have compared the retrieval effectiveness of the IR model with and without the stopword list. The performance differences were small (in mean, around 1%) and did not give any evidence of significant impact of stopword list removal on MAP, for this language at least. Of course, the inverted file was reduced as well as the query processing time.

### *6.2. Stemming evaluation*

Facing a language with more complex inflectional morphology than English, we may infer that applying stemming will improve the MAP. However to which extent (if it really exists) is not, a priori, known. This section will address these questions using different IR models.

If we use retrieval performance without stemming, marked "none" in Table 2 as a baseline, we can see that both stemming strategies, "light" and "aggressive", performed better than the baseline. Applying our statistical testing, we found that all performance differences were always statistically significant when compared to an approach ignoring the stemming stage. If we average the performance over four models given, we find an increase of 42% with the "light" stemmer and 46% with the more "aggressive" one. These relative improvements are clearly large and more important than with other languages (Tomlinson, 2004) (+4% with the English language, +4.1% Dutch, +7% Spanish, +9% French, +15% Italian, +19% German, +29% Swedish, +40% Finnish).

When comparing different stemming strategies we can see that the "aggressive" stemmer performs slightly better, 2.7% in average over four models. The retrieval performance differences were in this case never statistically significant.

Denoted as "4-gram" in Table 2 are shown retrieval performances of the given IR models when language independent 4-gram indexing strategy (without applying a stemming procedure). The performance difference between 4-gram indexing strategy and word-based indexing is rather small (e.g., in average −1% over "light" and −3.5% over "aggressive") and is never statistically significant.

When analyzing query-by-query the effect of applying a stemmer, and limiting our investigations of the best performing model (DFR-I($n_e$)C2), we found that after applying our light or more aggressive stemmer, the performance was increased for 41 queries while, for the remaining nine queries, the average precision (AP) decreases. In this case, Topic #418 ("Bülent Ecevit's Statements") has the greatest improvement, starting with an AP of 0.25 without stemming to 0.6797 (+172%) with our light stemmer and 0.7526 (+201%) with the more aggressive approach. Explanation for this improvement could be found in the fact that personal names in Czech, as in other Slavic languages are changed through cases. Genitive form of the name found in this query ("Prohlašení Bülenta Ecavita") as well as other forms found in relevant documents, after stemming conflate to its nominative form enabling a pertinent matching. Also, Topic #441 ("Space tourists") cannot retrieve any relevant articles without stemming (AP 0.0), retrieves the first relevant document in second place with both stemmers (e.g., AP 0.3568 with light stemmer). None of the terms forming the query ("Vesmírní turisté"), exists in relevant documents in the same

**Table 2**
Mean average precision (MAP) of various IR models and different stemmers.

| | None | Light | Aggressive | 4-Gram |
|---|---|---|---|---|
| *tf idf* | 0.1357* | 0.2040* | 0.2095* | 0.1918* |
| Okapi | 0.2040* | 0.2990 | 0.3065 | 0.2957* |
| DFR-$I(n_e)$C2 | *0.2208* | *0.3042* | *0.3135* | *0.3125* |
| LM | 0.2054* | 0.2813* | 0.2882* | 0.2785* |

word form (they occur as “vesmírný”, “vesmírnou”, “turista”). Of course, applying a stemmer may sometimes hurt the AP as shown by Topic #407 (“Australian Prime Minister”) having an AP of 0.9325 without stemming to 0.5616 (−39.8%) with our light stemmer and 0.5925 (−36.5%) with the more aggressive approach. In this case nouns “premiér” (prime minister) and “premiéra” (first night, premiere) conflate to the same stem resulting in retrieving large number of non-relevant articles.

Finally it is interesting to know that some topics could be classify as hard because for all indexing strategies and IR models they achieve a MAP smaller than 0.1. In our experiments, we have found seven such topics (#403, #411, #422, #425, #428,

```
CzechStemmer(word) {
  RemoveCase(word);
  RemovePossessives(word);
  Normalize(word);
  return;
}
RemoveCase(word) {
  if (word ends with “-atech”) then remove “-atech” return;
  if (word ends with “-ětem”) then remove “-ětem” return;
  if (word ends with “-etem”) then remove “-etem” return;
  if (word ends with “-atům”) then remove “-atům” return;
  if (word ends with “-ech”) then remove “-ech” return;
  if (word ends with “-ich”) then remove “-ich” return;
  if (word ends with “-ích”) then remove “-ích” return;
  if (word ends with “-ého”) then remove “-ého” return;
  if (word ends with “-ěmi”) then remove “-ěmi” return;
  if (word ends with “-emi”) then remove “-emi” return;
  if (word ends with “-ému”) then remove “-ému” return;
  if (word ends with “-ěte”) then remove “-ěte” return;
  if (word ends with “-ete”) then remove “-ete” return;
  if (word ends with “-ěti”) then remove “-ěti” return;
  if (word ends with “-eti”) then remove “-eti” return;
  if (word ends with “-ího”) then remove “-ího” return;
  if (word ends with “-iho”) then remove “-iho” return ;
  if (word ends with “-ími”) then remove “-ími” return;
  if (word ends with “-ímu”) then remove “-ímu” return;
  if (word ends with “-imu”) then remove “-imu” return;
  if (word ends with “-ách”) then remove “-ách” return;
  if (word ends with “-ata”) then remove “-ata” return;
  if (word ends with “-aty”) then remove “-aty” return;
  if (word ends with “-ých”) then remove “-ých” return;
  if (word ends with “-ama”) then remove “-ama” return;
  if (word ends with “-ami”) then remove “-ami” return;
  if (word ends with “-ové”) then remove “-ové” return;
  if (word ends with “-ovi”) then remove “-ovi” return;
  if (word ends with “-ými”) then remove “-ými” return;
  if (word ends with “-em”) then remove “-em” return;
  if (word ends with “-es”) then remove “-es” return;
  if (word ends with “-ém”) then remove “-ém” return;
  if (word ends with “-ím”) then remove “-ím” return;
  if (word ends with “-ům”) then remove “-ům” return;
  if (word ends with “-at”) then remove “-at” return;
  if (word ends with “-ám”) then remove “-ám” return;
  if (word ends with “-os”) then remove “-os” return;
  if (word ends with “-us”) then remove “-us” return;
  if (word ends with “-ým”) then remove “-ým” return;
  if (word ends with “-mi”) then remove “-mi” return;
  if (word ends with “-ou”) then remove “-ou” return;
  if (word ends with “-[aeiouyáéíýě]”) then remove “-[aeiouyáéíýě]” return;
  return;
}
RemovePossessives(word) {
  if (word ends with “-ov”) then remove “-ov” return;
  if (word ends with “-in”) then remove “-in” return;
  if (word ends with “-ův”) then remove “-ův” return;
  return;
}
Normalize(word) {
  if (word ends with “čt”) then replace by  “ck” return;
  if (word ends with “št”) then replace by  “sk” return;
  if (word ends with “c” or “č”) then replace by  “k” return;
  if (word ends with “z” or “ž”) then replace by  “h” return;
  if (word ends with “e*”) then replace by  “*” return;
  if (word ends with “*ů*”) then replace by “*o*” return;
  return;
}
```

**Fig. A1.** Our Czech light stemmer.

#436, #439). Those topics mostly contain either too general terms (e.g., Topic #436 "VIP divorces") or certain spelling errors (e.g., in Topic #411 "Best Picture Oscar", Academy Award's name was spelled with a K ("Oskar") in the topic and with a C ("Oscar") in relevant documents).

## 7. Conclusions

In this paper, we present the main aspects of the Czech morphology and we suggested two stemmers for this Slavic language, one removing only inflectional suffixes (denoted "light") and a second algorithm that removes also some frequent derivational suffixes (denoted "aggressive"). Both approaches contain some rules to correct orthographic irregularities. A stopword list containing 467 forms was also suggested. These linguistic tools are freely available on the Internet.

Using the most effective current IR models, we have evaluated our stemming approaches and found that the best performing IR model is derived from *Divergence from Randomness* (DFR) paradigm. This approach performs statistically better than a language model or the classical *tf idf* while the difference with the Okapi model was not statistically significant.

Our various experiments clearly show that a stemming procedure improves retrieval effectiveness when applied to the Czech language (mean improvement of around +45%, larger than those found for other European languages). From a statistical point of view, the differences are always significant when comparing to an approach ignoring stemming.

From comparing different stemming strategies, it seems that the more aggressive stemming approach produces better MAP than does a light stemmer, but the difference between these two stemming schemes is never statistically significant.

## Acknowledgment

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

## Appendix A. Description of our Czech light stemmer

See Fig. A1.

## References

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.

Buckley, C., & Voorhees, E. M. (2005). Retrieval system evaluation. In E. M. Voorhees & D. K. Harman (Eds.), *TREC. Experiment and evaluation in information retrieval* (pp. 53–75). Cambridge, MA: The MIT Press.

Dolamic, L., & Savoy, J. (2008). Stemming approaches for east European languages. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, & A. Peñas, et al. (Eds.), *Advances in multilingual and multimodal information retrieval. LNCS #5152* (pp. 37–44). Berlin: Springer-Verlag.

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

Harman, D. K. (2005). Beyond English. In E. M. Voorhees & D. K. Harman (Eds.), *TREC experiment and evaluation in information retrieval* (pp. 153–182). Cambridge, MA: The MIT Press.

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

Hiemstra, D. (2002). Term-specific smoothing for the language modeling approach to information retrieval. The importance of a query term. In *Proceedings of the ACM-SIGIR*, Tempere, pp. 35–41.

Kettunen, K., & Airo, E. (2006). Is a morphologically complex language really that complex in full-text retrieval? In *Advances in natural language processing. LNCS #4139* (pp. 411–422). 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). Washington, DC: The ACM Press.

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

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press.

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

Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D. W., & Peñas, A., et al. (Eds.). (2008). *Advances in multilingual and multimodal information retrieval. LNCS #5152*. Berlin: Springer-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. (2006). Light stemming approaches for the French, Portuguese, German and Hungarian languages. In *Proceedings ACM-SAC* (pp. 1031–1035). Dijon: The ACM Press.

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

Tomlinson, S. (2004). Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003. In *Comparative evaluation of multilingual information access systems. LNCS #3237* (pp. 286–300). Berlin: Springer-Verlag.

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

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems, 22*(2), 179–214.