A STEMMING PROCEDURE AND STOPWORD LIST FOR GENERAL FRENCH CORPORA

Jacques Savoy

Institut interfacultaire d'informatique
Université de Neuchâtel
Pierre-à-Mazel 7
CH - 2000 Neuchâtel (Switzerland)

to appear in
Journal of the American Society for Information Science, 50(10), 1999, 944-952

Abstract

Due to the increasing use of network-based systems, there is a growing interest in access to and search mechanisms for text databases in languages other than English. To adapt searching systems to those foreign languages with characteristics similar to the English language, all we need to do for the most part is to establish a general stopword list and a stemming procedure. This article presents the tools needed to establish these in the French language databases and some retrieval experiments that have been carried out using two medium-sized French language test collections. These experiments were conducted to evaluate the retrieval effectiveness of the propositions described.

Introduction

The browser technologies currently available for use on CD-ROMs and also local and wide-area networks (Internet and WWW), allow us to store, distribute and manage larger volumes of documents, many of which are not always written in English. To provide access and search mechanisms for these sources of information accessed through digital libraries (Lesk, 1997) or web browsers, we need to readapt portions of certain existing retrieval systems so that they can handle languages other than English.

Most European languages (e.g., French, Slovene, Italian) share many of the characteristics of Shakespeare's language (e.g., word boundaries marked in a conventional manner, variant word forms generated by adding suffixes at the

end of a root, etc.). Any adaptation therefore means the elaboration of a general stopword list and a fast stemming procedure. The stopword list contains non-significant words that are removed from a document or a request before beginning the indexing process. The stemming procedure tries to remove inflectional and derivational suffixes in order to conflate word variants into the same stem or root. In resolving this problem for the French language, it is important to remember that French and other European languages involve a more complex morphology than does English (Sproat, 1992). Previous examples of such adaptations are reported in (Popovic & Willett, 1992; Buckley *et al.*, 1995) where a stemming procedure is proposed for both the Slovene and Spanish languages respectively.

The aim of this article is therefore to propose a general stopword list and a simple stemming procedure required for a French corpora. Moreover, as a result of recent cooperation between various research groups, two medium-sized French test collections (see Appendix 2) have been created. These corpora, together with various current search strategies, were used to corroborate or invalidate prior assumptions or algorithms. This means that our findings are based on more solid arguments than on conclusions derived from a single retrieval model working on a small text collection (e.g., less than 500 records).

The rest of this paper is organized as follows. The first part describes the approach we used to establish a general stop list for French corpora. The second part details our "quick and dirty" inflectional stemming procedure based on a few general linguistic considerations. The third chapter summarizes and comments upon some of the experimental results that are used to justify both the suggested stopword list and the stemming procedure developed, and based on two French language test collections.

General Stopword List

For the purposes of this research, we consider a word to be each uninterrupted sequence composed of letters (a..z) , digits (0..9) or two special characters (@ and _). Thus, the phrase "la machine IBM-360" counts as four words but "la machine IBM360" as only three. In French, the apostrophe «'» is very often used as a word delimiter (e.g. "l'avenir" is composed of two words, namely the article "l" (the) and the noun "avenir" (future)). An exception would be the noun "aujourd'hui" (today), various English name transcriptions (e.g.,

McDonald's or K'NEX) and the comma used as a separator in numbers (e.g., 3,000,000 is written as 3'000'000 in French typography (Corthésy *et al.*, 1993)).

We defined a general stopword list for those words which serve no purpose for retrieval, but are used very frequently in composing the documents, and these stopword lists are developed for two main reasons: Firstly, we hope that each match between a query and a document will be based on good indexing terms. Thus, retrieving a document because it contains words like "be", "your" and "the" in the corresponding request does not constitute an intelligent search strategy. These non-significant words represent noise, and may actually damage the retrieval performance because they do not discriminate between relevant and nonrelevant documents. Secondly, we expect to reduce the size of the inverted file, hopefully in the range of 30% to 50%.

Although the objectives seem clear, we do not have a clear theoretical foundation upon which we can define a methodology for the development of a stop list, thus a certain arbitrariness is required (Fox, 1990). For example, the SMART system has 571 English words in its stopword list, Fox (1990) suggests 421 words while DIALOG Information services (Harter, 1986, p. 88) propose using only nine terms (namely "an", "and", "by", "for", "from", "of", "the", "to" and "with").

In establishing a general stopword list for French, we followed the guidelines described in (Fox, 1990). Firstly, we sorted all the word forms appearing in our French corpora according to their frequency of occurrence and we extracted the 200 most frequently occuring words. Secondly, we inspected this list to remove all numbers (e.g., "1992", "1"), plus all nouns and adjectives more or less directly related to with the main subjects of the underlying collections (French articles were extracted from a newspaper as described in Appendix 2). For example, the words "France" ranked at the 66th position on the list as well as the noun "Président" (ranked at the 69th position) were removed from the list. Also removed were other nouns such as "janvier" (January), "Paris", "francs", "millions" or "Jean" (John) as well as adjectives (e.g., "premier" usually appearing in the expression "premier ministre" (prime minister) or "deux" (two)). From our point of view, such words can be useful as indexing terms in only some circumstances. Thirdly, we included some non-information-bearing words, even if they did not appear in the first 200 most frequent words. For example, we added various personal or possessive pronouns (such as "moi" (me),

"tien" (yours)), prepositions ("dessus" (upon)) and conjunctions ("cependant" (however)).

In the resulting stopword list there were thus a large number of pronouns, articles, prepositions and conjunctions. As in various English stopword lists, there were also some verbal forms ("être" (to be), "ont" (have), "sont" (are)). However there was only one noun ("aujourd'hui" (today) included as two words "aujourd" and "hui" because a quote is considered as a word boundary).

We did not included various frequently used words such as "world" ("monde" appearing in the 81st position of the 200 most frequent words in our corpora) or "political" ("politique" appearing in the 78th rank order), "years" ("ans", 71st position), "city" ("ville", 158th position), "ministre" (79th position), "day" and "days" ("jour" and "jours", 190th and 191st position), "life" ("vie", 152nd position). The presence of homographs represents another debatable issue, and to some extent, we had to make arbitrary decisions concerning their inclusion in a stopword list. For example, the French word "son" can be translated as "sound" or "his", and the French term "or" as "thus/therefore" or "gold".

The general stopword list suggested for French contains 215 words and is included in Appendix 1. When using such a stopword list, the size of the inverted file was reduced by about 21% for one test collection, and about 35% for the second corpus. Ordering the words according to their occurrence frequency also confirms Zipf's law, and based on our French corpora, the 10 most frequent words represent 23.2% of all occurrences in these text databases, while the 20 most frequent words cover 32.4% of all forms appearing in the documents.

Stemming Procedure

After removing high frequency words, an indexing procedure tries to conflate word variants into the same stem or root using a stemming algorithm. For example, the words "thinking", "thinkers" or "thinks" may be reduced to the stem "think". In information retrieval, grouping words having the same root under the same stem (or indexing term) may increase the success rate when matching documents to a query (van Rijsbergen, 1979, Chapter 2; Salton, 1989; Frakes, 1992). Such an automatic procedure may therefore be a valuable tool in enhancing retrieval effectiveness, assuming that words with the same stem refer to the same idea or concept and must be therefore indexed under the same form.

When defining a stemming algorithm, a first approach will only remove inflectional suffixes or, for English, such a procedure conflates singular and plural word forms as well as removing the past participle ending «-ed» and the gerund or present participle ending «-ing». More sophisticated schemes for English corpora have also been proposed for the removal of derivational suffixes (e.g., «-ize», «-ably», «-ship»). For example, Lovins' stemmer (Lovins, 1968) is based on a list of over 260 suffixes, while Porter's algorithm looks for about 60 suffixes (Porter, 1980). Most of these suffix-stripping algorithms are controlled by both quantitative constraints (e.g., a minimal stem length must be respected for a given suffix removal operation) and qualitative constraints (e.g., the ending must satisfy a certain condition). Finally, a set of recoding rules may be followed in order to alter stems and to improve the conflation (e.g., "hopping" minus «-ing» gives "hop" and not "hopp"). Various implementation strategies have also been suggested (Frakes, 1992).

In defining an inflectional stemmer for French, there are a greater number of irregularities to consider (Grevisse & Goose, 1988). Although English contains morphological irregularities (e.g., box / boxes, mouse / mice, keep / kept) there are even more in French and in other languages (e.g., Slovene, Italian). In fact, these include inflectional suffixes governed by gender variations (masculine vs. feminine) and number variations (singular vs. plural) both for nouns and adjectives. For verbs, we must add variations in tense and person. The resulting set of rules and exceptions is quite large, and, as an extreme example, the verb "être" (to be) possesses 40 different possible forms. As another stop list example, the one we suggest contains the variations in gender and number for various pronouns ("mien" in masculine singular, "miens" in masculine plural, "mienne" in feminine singular, and "miennes" in feminine plural) (Sproat, 1992).

In order to resolve this problem, Krovetz (1993) suggests using a stemming procedure based on both inflectional and derivational suffixes within which the suffix stripping process is under the control of an English dictionary. Hull (1996) presents a similar approach based on various linguistic tools. For French, Savoy (1993) proposes a suffixing algorithm also based on grammatical categories, although such an approach requires a French dictionary, an electronic resource that is not freely available. Moreover, the suggested procedure is time consuming compared to various approaches designed for the English language (e.g., Porter's stemmer) or for the Slovene language (Popovic & Willett, 1992).

Figure 1 below depicts a detailed description of our "quick and dirty" stemming procedure for the French language. The principal feature of this suggested stemming procedure is that it is based on only a few general morphological rules. In French the main inflectional rule is to add a final «-s» to denote the plural form for both nouns and adjectives. Another common morpheme for indicating the plural is adding a final «-x» (as in "hibou / hiboux" (owl / owls) or in a slightly more complex circumstance, for nouns ending with «-al» such as "cheval / chevaux" (horse / horses)). The suggested algorithm does not account for person and tense variations, or for the morphological variations used by verbs. Our procedure therefore corresponds to the English "S stemmer" which conflates singular and plural word forms (Harman, 1991).

```
For words of five or more letters
   if the final letter is «-x» then
         if final is «-aux» then replace final «-aux» by «-al»
                                                    (e.g., chevaux -> cheval)
         otherwise, remove final «-x»               (e.g., hiboux -> hibou)
   otherwise    (words not ending with «-x»)
         if final letter is «-s» then remove final «-s» (e.g., chantés -> chanté)
         if final letter is «-r» then remove final «-r» (e.g., chanter -> chante)
         if final letter is «-e» then remove final «-e»(e.g., chante -> chant)
         if final letter is «-é» then remove final «-é»(e.g., chanté-> chant)
            (a simple recoding rule, e.g., baronn-> baron)
         if final two letters are the same, remove final letter
otherwise does not alter words of four or less letters
```

Figure 1: Weak stemmer for French language

Using our stemming procedure, the French words "baronnes" (baronesses), "barons" and "baron" will be reduced to the same stem "baron". Of course, various counter-examples can also be found, such as "français" and "françaises" (the adjective "French" in its masculine and feminine plural forms) that cannot be reduced to the same root ("français" is reduced to "françai", a non-French word and "françaises" to "français"). Moreover, obtaining the exact semantic root of a given form is not always achieved by the automatic stemming procedures, so that we are faced with various conflation errors (see examples of various English stemming procedures in (Krovetz, 1993)). Working with "real" and large text collections reveals other problems such as conflating of misspelled terms or removing suffixes from the proper nouns appearing in a document or a request.

Experimental Results

To evaluate the retrieval effectiveness of our suggested stopword list and stemming procedure, we have used two French test collections. The first corpus, OFIL, contains selected articles from the French newspaper *Le Monde* (11,016 documents, 26 queries). INIST is our second test collection, composed of very short abstracts of scientific articles (163,308 documents and 30 queries). Various statistics regarding both test collections can be found in Appendix 2.

As a means of evaluation, we used the non-interpolated average precision at 11 recall values provided by the TREC2_EVAL software based on 1,000 retrieved items per request (Harman, 1995). To decide whether a search strategy is better than another, we need a decision rule. The following rule of thumb may be used to define such a rule: a difference of at least 5% in average precision is generally considered significant and a 10% difference is considered material (Sparck Jones & Bates, 1977, p. A25). For a more precise decision, we might also apply statistical inference methods such as Wilcoxon's signed rank test (Salton & McGill, 1983, Section 5.2; Hull, 1993) or hypothesis testing based on bootstrap methodology (Savoy, 1997).

Evaluation of stemming and nonstemming searches

In evaluating various search strategies, we considered the OKAPI probabilistic model (Robertson *et al.*, 1995) and various vector-processing schemes (retrieval status computed according to the inner product (Salton, 1989, p. 318)). Following Buckley *et al.*, (1995), we used three letters to denote the weighting method for documents, combined with three letters for the weighting method for queries. The exact formulation for each indexing scheme is described in more detail in Appendix 3. For example, one can find the simple coordinate match (doc = BNN, query = BNN) within which the retrieval status value of each document corresponds to the number of terms in common with the query. Another simple indexing strategy which uses only the occurrence frequency for each term in the document or the request is described using the label "doc = NNN, query = NNN". In addition to these two well-known indexing weighting schemes, we also suggest employing more complex indexing formulae (e.g., LTN, LTC, ATN) within which an indexing term weight depends on both its frequency of occurrence within a document and its importance within the entire collection (idf component). Finally, we also used the LNU and OKAPI weighting schemes, which take account of document length.

To provide a more precise interpretation of these retrieval effectiveness results, in the following tables we have underlined statistically significant differences based on a one-sided Wilcoxon signed rank test with a significance level fixed at 5%. Our baseline performance shown in the second column of Table 1 is achieved by a retrieval scheme with does not use a stopword list and ignores our weak suffix-stripping procedure.

	Precision (% change)			
Collection Model	OFIL no stop. no stem.	OFIL with stop list no stemming	OFIL no stop list with stemming	OFIL with stop list & stemming
doc=OKAPI, query=NPN	32.21	34.92 (+8.41%)	33.65 (+4.47%)	35.59 (10.49%)
doc=LNC, query=LTC	32.75	33.77 (+3.21%)	34.76 (+6.14%)	36.90 (12.67%)
doc=LTC, query=LTC	32.71	32.86 (+0.46%)	36.65 (+12.05%)	36.34 (+11.10%)
doc=LNU, query=LTC	30.97	32.60 (+5.26%)	32.17 (+3.87%)	34.44 (+11.20%)
doc=ANC, query=LTC	29.88	31.25 (+4.58%)	32.21 (+7.80%)	33.56 (+12.32%)
doc=ATN, query=NTC	25.00	29.51 (+18.04%)	26.56 (+6.24%)	31.02 (+24.08%)
doc=LTN, query=NTC	23.24	26.22 (+12.82%)	23.11 (-0.56%)	26.47 (+13.90%)
doc=NNN, query=NNN	0.20	6.26 (+3030%)	0.21 (+5.0%)	4.70 (+2250%)
doc=BNN, query=BNN	6.37	10.63 (+66.88%)	5.01 (-21.35%)	8.62 (+35.32%)

Table 1a: Average precision of various indexing strategies (OFIL collection)

	Precision (% change)			
Collection Model	INIST no stop. no stem.	INIST with stop list no stemming	INIST no stop list with stemming	INIST with stop list & stemming
doc=OKAPI, query=NPN	10.70	15.47 (+44.58%)	14.83 (+38.60%)	19.17 (+79.16%)
doc=LNC, query=LTC	11.01	11.81 (+7.27%)	15.38 (+39.69%)	16.03 (+45.59%)
doc=LTC, query=LTC	11.49	11.42 (-0.61%)	15.68 (+36.47%)	15.45 (+34.46%)
doc=LNU, query=LTC	12.51	14.57 (+16.47%)	15.22 (+21.67%)	17.67 (+41.25%)
doc=ANC, query=LTC	11.38	11.72 (+2.99%)	15.17 (+33.30%)	15.56 (+36.73%)
doc=ATN, query=NTC	14.62	15.13 (+3.49%)	17.71 (+21.14%)	18.26 (+24.90%)
doc=LTN, query=NTC	12.50	13.85 (+10.80%)	15.24 (+21.92%)	16.83 (+34.64%)
doc=NNN, query=NNN	0.19	5.52 (+2805%)	0.20 (+5.26%)	6.46 (+3300%)
doc=BNN, query=BNN	2.21	7.33 (+231.67%)	2.51 (+13.57%)	7.21 (+226.24%)

Table 1b: Average precision of various indexing strategies (INIST collection)

From data depicted in Table 1, it can be seen that retrieval performance depends on the test collection. Average precision for the INIST collection is lower than that of the OFIL corpus. Based on various statistics shown in Appendix 2, we may point out that the average document length is much shorter for the INIST corpus than for the OFIL collection (52.0 words per document vs. 379.8). Short documents contain less evidence, resulting in poorer retrieval effectiveness.

Moreover, the number of documents included in the INIST test collection is 14 times greater than the size of the OFIL collection.

The last two rows of Table 1 displays the two poorest retrieval performances achieved by retrieval schemes ignoring collection-wide information ("doc = NNN, query = NNN"; "doc = BNN, query = BNN"). On the other hand, it could be inferred that the OKAPI probabilistic model results in very interesting retrieval performance for both test collections.

When presenting the results obtained by various vector-processing strategies, we rank them according to the retrieval performance achieved by the OFIL corpus when using both the suggested stopword list and stemming procedures (last column of Table 1). In the first line, we add the OKAPI model (representing a probabilistic retrieval model) which has good retrieval performance overall. Looking at the INIST corpus retrieval performance, it can be seen that we cannot obtain consistent ranking between the two test collections leading to the conclusion that the performance for a given search scheme depends on the underlying test collection characteristics.

The second column of Table 1 depicts the average performance obtained without using stopword list and stemming procedures. The overall retrieval effectiveness is poor compared to the other columns leading to the general conclusion that for retrieval purposes both stemming and removing highly frequent words are overall beneficial.

As a study of the relative merit of the stopwording and stemming procedures, the third column of Table 1 depicts the average performance obtained with stopwording but without using our weak suffix-stripping procedure. The data shows that stopwording is strongly advantageous for both collections when using the OKAPI search strategy. In our set up, we removed any search keyword having a negative indexing weight which correspond to very frequent words. Such a context is also strongly advantageous for the two poorest retrieval schemes. With the third search strategy ("doc = LTC, query = LTC"), there appears to be no advantage in using a stopword list. For the remaining strategies, stopwording seems to be beneficial, but the extent of the effect is varied and rather inconsistent across test-collections.

The fourth column shows the average performance achieved by various retrieval schemes with our suffix removing procedure but without the removal

of the highly frequent words included in the stopword list. The stemming procedure seems particularly beneficial for the INIST collection.

Our stemming procedure can also be evaluated when looking at average precision results depicted in the last column of Table 1. The comparative performance between the conflated and nonconflated document representation indicates that a stemming procedure significantly favors the system performance and thus is confirmed by other studies based on English language corpora (Krovetz, 1993; Hull, 1996) and partially by Harman's study (1991) in which the differences in average precision are close to 5%, the limit value of our significance level.

Leaving the two poorest strategies aside, stemming is highly beneficial for the INIST collection, but only modestly beneficial for OFIL. This is presumably related to the different document lengths and collection sizes. In OFIL documents (an average document length of about 379.8), key concepts are likely to be mentioned several times, so both singular and plural forms will be represented: a search term is therefore likely to match in the document whether its form is singular or plural. In the INIST collection (a mean document length of 52.0), this will apply much less frequently, so the benefits of stemming will be greater. For the two least effective strategies, stemming is significantly damaging for OFIL documents based on the simple coordinate search strategy ("doc = BNN, query = BNN"), but is neutral or advantageous for INIST documents.

As usual, average performance may hide performance irregularities among requests. We performed a more detailed analysis of the performance achieved by the OKAPI model for stemming vs. nonstemming searches for both test collections and without stopwording. In a per-query analysis, the stemming procedure performs better for 19 of the searches, and worse for the remaining 7 for the OFIL corpus (an average precision of 32.21 vs. 33.65 (+4.47%)). For the INIST collection, the stemming search performs better for 25 requests, and worse for the remaining 5 (an average precision of 10.70 vs. 14.83 (+38.60%)). Based on the Wilcoxon signed ranking test (significance level fixed at 5%), the null hypothesis stating that both retrieval schemes produce similar retrieval performance must be accepted for the OFIL collection. This null hypothesis is rejected for the INIST corpus (average precision is therefore significant between the two retrieval schemes). A similar conclusion can be drawn when using bootstrap methodology.

In another experiment, we studied the retrieval effectiveness of various French stemming procedures. We developed another stemming algorithm which also removes most frequent French derivational suffixes defined by conducting a quantitative study of the frequency of various endings. When we compared such a strategy with our weak suffix-stripping approach, the difference in average precision was not significant (about 1.1%) and was in favor of our simple weak stemmer. These results tended to confirm other studies carried out on English stemming (Frakes, 1992, Section 8.3; Harman, 1991; Krovetz, 1993) in which the differences between various stemming procedures were not significant. However, since French morphology is far more complex than English morphology, a direct comparison cannot be made. According to Popovic & Willett (1992), and when trying to remove a large number of suffixes for a morphologically complex natural language, a simple stemming procedure seems to be more useful and effective than a more complex one which results in more conflation errors.

And the accents?

In most European languages, one of the first problems encountered is the requirement for storing each character using 8 bits (e.g., using the ISO LATIN standard) instead of the standard ASCII code. In French, as in most European languages, accents are used to indicate the precise pronunciation and to identify some homographs (e.g., "où" means "where " and "ou" "or", "mais" means "but" and "maïs" "corn").

Thus, according to the strict rules of composition (Corthésy *et al.*, 1993), words containing letters with accents must be written with the accents, even when these words appear as capitals. The word "Québec" must always therefore be written with its accent (even in a title as in "QUÉBEC"). However, if only the first letter in a word is a capital, any accent on it must be removed (e.g., "état" must be composed as "Etat").

As in every rule of usage, this principle is not always respected, and usually the words in a title written in capitals appear without any accent. To account for this usage, the stopword list contains, for example, both the correct form of the verb "to be" ("être") and the form without an accent ("etre" which is no longer a French word).

To evaluate the relative importance of the accents for retrieval purposes, we modified the queries that included accented words. For those terms, we automatically included a copy of the corresponding word without its accents in the request. For example, an original request written as "chômage et économie" (unemployment and economics) will be treated as "chômage chomage et économie economie". Our prior assumption was that such a modification could be valuable because a search keyword included without its accent would now match any identically word appearing in a title (and written in capitals without its accents). Of course, we have also assumed that a match with a word included in a title can be considered as an important match. On the other hand, we also knew that the exact meaning of a phrase is often affected when the accents are removed as, for example, the noun phrase "un dossier critiqué" (a criticized case) and "un dossier critique" (a critical case).

	Precision (% change)		
Collection Model	OFIL baseline	OFIL modified queries	OFIL all accents removed
doc=OKAPI, query=NPN	35.59	33.61 (-5.56%)	36.81 (+3.43%)
doc=LNC, query=LTC	36.90	35.33 (-4.25%)	37.36 (+1.25%)
doc=LTC, query=LTC	36.34	33.95 (-6.58%)	36.52 (+0.50%)
doc=LNU, query=LTC	34.44	33.36 (-3.14%)	35.06 (+1.80%)
doc=ANC, query=LTC	33.56	31.93 (-4.86%)	34.10 (+1.61%)
doc=ATN, query=NTC	31.02	27.64 (-10.90%)	31.44 (+1.35%)
doc=LTN, query=NTC	26.47	25.80 (-2.53%)	27.64 (+4.42%)
doc=NNN, query=NNN	4.70	4.96 (+5.53%)	4.84 (+2.98%)
doc=BNN, query=BNN	8.62	8.71 (+1.04%)	8.67 (+0.58%)

Table 2a: Evaluation of various indexing strategies (OFIL collection)

	Precision (% change)		
Collection Model	INIST baseline	INIST modified queries	INIST all accents removed
doc=OKAPI, query=NPN	19.17	17.86 (-6.83%)	19.78 (+3.18%)
doc=LNC, query=LTC	16.03	15.18 (-5.30%)	15.90 (-0.81%)
doc=LTC, query=LTC	15.45	14.28 (-7.57%)	15.57 (+0.78%)
doc=LNU, query=LTC	17.67	17.17 (-2.83%)	18.05 (+2.15%)
doc=ANC, query=LTC	15.56	14.90 (-4.24%)	15.66 (+0.64%)
doc=ATN, query=NTC	18.26	16.59 (-9.15%)	18.82 (+3.07%)
doc=LTN, query=NTC	16.83	15.49 (-7.96%)	17.26 (+2.55%)
doc=NNN, query=NNN	6.46	6.46 (0.0%)	6.43 (-0.46%)
doc=BNN, query=BNN	7.21	7.11 (-1.39%)	7.38 (+2.36%)

Table 2b: Evaluation of various indexing strategies (INIST collection)

The retrieval results depicted in the third column of Table 2 show that this modification significantly decreases average precision based on the Wilcoxon signed rank test.

In another experiment, we removed all accents from French documents, requests and stopword list, and in doing so we hoped to achieve significantly improved retrieval performance. We might thus assume that in such circumstances various weighting schemes based on document frequency information (e.g., LTC, LTN, ATN, NTC, NPN) could now more properly weight the accentuated words (e.g., the term "économie" or "economie" would now have the same indexing weight).

The performance achieved by such a practice is depicted in the fourth column of Table 2. The resulting improvement can be considered as marginal across the various retrieval schemes and does not vary to any large extent between the two test collections. Thus, ignoring the accents does not significantly enhance retrieval effectiveness.

Removing the accents may improve overall recall but this advantage is counterbalanced by a loss of precision, due to false conflation. Such findings resemble those often encountered when expanding queries by adding synonyms extracted from a thesaurus.

Conclusions

In this article, we discussed the requirements for a general stopword list to be used for French corpora and a fast procedure for removing plural suffixes for the French language. Generating a general stopword list is subject to various arbitrary decisions; however we believe that the resulting stop list does not seem to include many controversial items. Moreover, we believe that the suggested list can be adapted for a specific domain by excluding some terms or adding new ones.

In the second part we presented a weak stemming procedure which essentially attempts to remove the plural inflections of words in French language document collections. Such a practice seems to be an adequate measure for improving retrieval effectiveness; and it will always be possible to derive more specific stemming procedures for given domains (e.g., French medical terminology).

The various experiments carried out with French document collections show that:

- Strategies which took collection-wide term distribution into account resulted in much better retrieval performance that strategies which did not (the "good strategies");
- Relative ranking of "good retrieval schemes" is different between the two collections, so comparative work with other collections is desirable;
- For both corpora, stopwording is highly advantageous with the OKAPI retrieval model, and not at all with the "doc = LTC, query = LTC" strategy; for the other strategies it is generally advantageous, but the picture is not consistent;
- The impact of a stemming procedure is clearly more beneficial when dealing with a collection of short documents; for longer document corpora, thre is some improvement but to a lesser extent;
- The relative complexity of French morphology seems to favor a simpler versus a more complex stemmer, one that would also try to remove derivational suffixes and produce more conflation errors;
- Using both stopwording and stemming significantly improve retrieval effectiveness;
- Ignoring accents, at least for French text collections, does not significantly enhance average precision; the overall result is only marginally improved retrieval effectiveness.

Acknowledgments

The author would like to thank C. Buckley for writing and supporting the SMART text retrieval system, without which this study could not have been conducted. This research was supported by the SNSF (Swiss National Science Foundation) under grants 20-43'217.95 and 20-50'578.97.

References

Buckley, C., Salton, G., Allen, J., & Singhal, A. (1995, April). *Automatic query expansion using SMART*. Proceedings of the TREC'3 Conference, (pp. 69-80). Gaithersburg, MA: NIST publication # 500-225.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996, October). *New retrieval approaches using SMART*. Proceedings of the TREC'4 Conference, (pp. 25-48). Gaithersburg, MA: NIST publication # 500-236.

Corthésy, G., Porchet, B., Umiglia C., & Chatelain R. (1993). *Guide du typographe romand*. Lausanne: Association Suisse de Typographie.

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

Frakes, W. B. (1992). Stemming algorithms. In W. B. Frakes, R. Baeza-Yates (Eds.), *Information retrieval, data structures & algorithms* (pp. 131-160). Englewood Cliffs: Prentice-Hall.

Grevisse, M. & Goose, A. (1988). *Le bon usage*. Paris: Duculot.

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

Harman, D. (1995). *Overview of the third text retrieval conference (TREC-3)*. Proceedings TREC'3 (pp. 1-19), Gaithersburg, MD: NIST .

Harter, S. P. (1986). *Online information retrieval*. San Diego: Academic Press.

Hull, D. (1993). *Using statistical testing in the evaluation of retrieval experiments*. Proceedings of the 16th International Conference of the ACM-SIGIR'93, (pp. 329-338). Pittsburgh, PA: ACM.

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

Krovetz, R. (1993, June). *Viewing morphology as an inference process*. Proceedings of the 16th International Conference of the ACM-SIGIR'93, (pp. 191-202). Pittsburgh, PA: ACM.

Lesk, M. (1997). *Practical digital libraries: books, bytes, and bucks*. San Francisco: Morgan Kaufmann.

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

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, 384-390.

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

van Rijsbergen, C. J. (1979). *Information retrieval*. 2nd ed. London: Butterworths.

Robertson, S. E., Walker, S., & Hancock-Beaulieu, M. M. (1995). Large test collection experiments on an operational, interactive system: OKAPI at TREC. *Information Processing & Management*, 31, 345-360.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New-York: McGraw-Hill.

Salton, G. (1989). *Automatic text processing, the transformation, analysis, and retrieval of information by computer*. Reading: Addison-Wesley.

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

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33, 495-512.

Sparck Jones, K., & Bates, R. G. (1977). *Research on automatic indexing 1974-1976*. Technical Report, Computer Laboratory, University of Cambridge, UK.

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

Appendix 1: Stopword List for French Corpora (215 words)

a	dans	je	ou	soit
afin	de	jusqu	outre	son
ai	debout	jusque	où	sont
ainsi	dedans	l	par	sous
après	dehors	la	parmi	suivant
attendu	delà	laquelle	partant	sur
au	depuis	le	pas	ta
aujourd	derrière	lequel	passé	te
auquel	des	les	pendant	tes
aussi	desquelles	lesquelles	plein	tien
autre	desquels	lesquels	plus	tienne
autres	dessous	leur	plusieurs	tiennes
aux	dessus	leurs	pour	tiens
auxquelles	devant	lorsque	pourquoi	toi
auxquels	devers	lui	proche	ton
avait	devra	là	près	tous
avant	divers	ma	puisque	tout
avec	diverse	mais	qu	toute
avoir	diverses	malgré	quand	toutes
c	doit	me	que	tu
car	donc	merci	quel	un
ce	dont	mes	quelle	une
ceci	du	mien	quelles	va
cela	duquel	mienne	quels	vers
celle	durant	miennes	qui	voici
celles	dès	miens	quoi	voilà
celui	elle	moi	quoique	vos
cependant	elles	moins	revoici	votre
certain	en	mon	revoilà	vous
certaine	entre	moyennant	s	vu
certaines	environ	même	sa	vôtre
certains	est	mêmes	sans	vôtres
ces	et	n	sauf	y
cet	etc	ne	se	à
cette	etre	ni	selon	ça
ceux	eux	non	seront	ès
chez	excepté	nos	ses	été
ci	hormis	notre	si	être
combien	hors	nous	sien	ô
comme	hélas	néanmoins	sienne	
comment	hui	nôtre	siennes	
concernant	il	nôtres	siens	
contre	ils	on	sinon	
d	j	out	soi	

Appendix 2: Test Collections Statistics

In an effort similar to that of the ARPA-TISPTER project, the Agence francophone pour l'enseignement supérieur et la recherche (AUPELF-UREF) and DISTNB have launched the AMARYLLIS project with the aim of exploring the underlying problems concerning the French language in relation to computer science technologies.

In a first cycle, two French test collections were created. The first corpus named OFIL contains selected articles from the French newspaper *Le Monde* (11,016 documents, 26 queries). On the average, each document is relatively small, having only 379.8 words (or 207 distinct words). When considering only the indexing words upon which each article is indexed, there is an average of 182.2 single terms.

The second test collection INIST is composed of very short abstracts of scientific articles extracted from the humanities, the arts and the sciences (163,308 documents, 30 queries). For this corpus, the mean number of words is 52 per document (or 37.9 distinct words) leading to an average of 24.5 indexing terms per article.

Collection	OFIL	INIST
Size	32.3 MB	65 MB
# of documents	11,016	163,308
# of queries	26	30
# of index terms / query	26.65	29.3
# of relevant documents	587	1,407
# of words / per document		
mean	379.8	52.0
standard deviation	399.2	31.6
maximum	5,883	319
minimum	4	1
# of distinct words / doc.		
mean	207.0	37.9
standard deviation	167.4	18.9
maximum	1,484	185
minimum	2	1
# of indexing terms / doc.		
mean	182.2	24.5
standard deviation	154.0	12.9
maximum	1,359	135
minimum	2	1

Table A.1: Various statistics associated with each test collection

Appendix 3: Weighting Schemes

To assign an indexing weight w_{ij} that reflects the importance of each single-term T_j in a document D_i, we may take three different factors into account. They are represented by the three code letters respectively:

- the within-document term frequency, noted tf_{ij} (first letter);
- the collection-wide term frequency, noted df_j (second letter);
- the normalization scheme (third letter).

N	new_tf = tf_{ij} (occurrence frequency of T_j in the document D_i)
B	new_tf = binary weight (0 or 1)
A	new_tf = $0.5 + 0.5 \cdot (tf_{ij}$ / max tf in $D_i)$
L	new_tf = $\ln(tf_{ij}) + 1.0$
L	new_tf = $[\ln(tf_{ij}) + 1.0]$ / $[1.0 + \ln(\text{mean (tf in } D_i))]$
N	new_wt = new_tf (no conversion is to be done)
T	new_wt = new_tf $\cdot \ln[N / df_j]$
P	new_wt = new_tf $\cdot \ln[(N - df_j) / df_j]$
N	w_{ij} = new_wt (no conversion is to be done)
C	divide each new_wt by sqrt (sum of (new_wts squared)) to get w_{ij}
U	w_{ij} = new_wt / $[(1\text{-}c) \cdot \text{mean}(nt) + c \cdot nt_i]$

Table A.2 : Weighting schemes

In Table A.2, the document length (the number of indexing terms) of D_i is noted by nt_i, the mean(nt) stands for the collection mean and the constant c is fixed at 0.2. Finally, the OKAPI weighting scheme correspond to:

$$w_{ij} = \frac{2 \cdot tf_{ij}}{C + tf_{ij}}$$

within which C is computed as $0.5 + 1.5 \cdot [\text{sum}(tf_i) / \text{mean}(tf)]$ (the ratio between the length of D_i noted by $\text{sum}(tf_i)$ and the collection mean noted by mean(tf)).