Océ at CLEF 2002

Roel Brand, Marvin Brünner

Océ Technologies
P.O. Box 101
NL-5900-MA Venlo
The Netherlands
{rkbr , mbru}@oce.nl

Abstract

This report describes the work done by the Information Retrieval Group at Océ Technologies B.V., for the 2002 edition of the Cross-Language Evaluation Forum (CLEF). We have participated in the mono, cross and multilingual tasks, using BM25 for ranking, Ergane, Logos and BabelFish for translations and the Knowledge Concepts semantic network for stemming and morphological expansion.

1 Introduction

To enlarge our knowledge and experience with information retrieval in multi-lingual document collections, we again participated in the Cross-Language Evaluation Forum (CLEF) this year. Last year we only participated in the Dutch monolingual task. Our goal for this year was to participate in all of the mono-lingual tasks, some of the cross lingual and in the multi-lingual task. Additionally we wanted to explore methods for combining results from different languages to obtain the best multi-lingual result.

This report describes the details of the retrieval system we built to create our contribution and the results we obtained in the contest.

2 Methods

2.1 Query construction

From the topics, queries were automatically constructed. Fields from the topics were split into terms on non-alphanumerical characters. Single character or stopword terms were removed. Each term was expanded with its root form by using morphological collapse (dictionary based stemming) from Knowledge Concepts' Content Enabler semantic network The root form was then expanded with the semantic network, with the morphological variants of the root form (such as plural form, etc.).

2.2 Compound Splitting

For both Dutch and German, compound words were split using a simple recursive algorithm that takes a word, tries to split a front part that is a complete word, and recurses with the remaining tail of the word. If a word is successfully split into N words, these words are added to the query. For instance the word 'waterbeddenfabrikant' is expanded into 'water bed den fabrikant' and 'water bedden fabrikant'. Sometimes, for pronunciation, an additional 's' is inserted between two parts of the compound. A simple rule in the compound splitter handles these characters. Optionally, related terms or synonyms could be added, but tests show worse results than without.

For the German task, the semantic network was not working properly; hence a stemmer was used to generate a mapping from a term into its stemming variants [3].

2.3 Topic translation

Because we participated in cross and multilingual retrieval, topics written in one language are used to query collections in another or multiple languages. In order to do these tasks, we translated the topics to the language(s) of the collection.
Experiments have been conducted using the semantic network for word-by-word translation. This however yields very bad results for translation, because a word can point to many concepts and we have no tools to select the proper concept. Therefore the semantic network based translation adds way too many terms, resulting in too broad queries. For instance, the English word 'baby' was translated into Dutch as three concepts with the following terms:

1. Liefje, lieveling, lieverd, schat, schatje, snoesje
2. Baby, dreumes, hummel, jong kind, kind, kleuter, pasgeborene, peuter, puk, uk, wurm, zuigeling
3. Bemoederen, in de watten leggen, koesteren, liefderijk verzorgen, moederen over, vertroetelen, verwekelijken, verwennen

Clearly, only a very few terms are correct given the fact that a term in a topic is often used in a specific way. Therefore, this method of translation can not be used in a non-interactive way and hence was abandoned for the CLEF.

For bilingual tasks we did English to Dutch, Dutch to English, English to Spanish and English to Spanish. The International Training Center of Océ translated English to Spanish using Logos machine translation software. The Internet babelfish translation was used for Spanish to English. For the translation of English to Dutch and vice versa, we used the Ergane dictionaries to translate on a word-by-word base. This translation did not have the problem of query broadening, probably because of its limited amount of words. After translation, the resulting topics were processed similar as the non-translated topics.

For the multilingual task we used English as the source language and translated it to German, Spanish, Italian and French, with Babelfish and Logos.

2.4 Indexing

In the indexes we built for each of the languages, documents were split on non-alphanumerical characters and single character words were ignored. We performed no stemming. For stop word elimination we used the stop list from Knowledge Concepts' Content Enabler semantic network.

2.5 Ranking

Instead of using our own model like last year, we based our system on BM25 [2] this year. Experiments with last year's topics showed that BM25 outperformed our own model. To work with expanded query terms in our implementation of BM25, we summed term frequencies over the expansion of each term and we defined document frequency as the number of documents in which a term or any of its expansions occurs (Equation 1).

2.6 Result merging

Due to lack of time we were not able to invest much effort in merging retrieval results from multiple languages into a single multilingual result. We therefore choose two basic approaches:

- round robin over all languages in the order es, fr, it, de, en ;
- merge sort based on document scores, normalised by dividing them by the score of the highest ranked document of the same language.

Equation 1: calculating tf and df for expanded terms, and the total score for a document given a query

Let q_i be a query term in query Q

Let $q_{i,0}, q_{i,1}, .., q_{i,n}$ be the expansion of q_i in which $q_{i,0} = q_i$

Let $tf(q_{i,j}, d)$ be the term frequency of expansion term $q_{i,j}$

We now calculate the document and term frequency of q_i as follows:

$$tf(q_i, d) = \sum_j tf(q_{i,j}, d)$$

$$df(q_i) = \left| \bigcup_j \text{set of documents in which } q_{i,j} \text{ occurs} \right|$$

Then for a Document d, and Query Q, the score is calculated as:

$$\text{Rel}(d, Q) = \sum_{q_i \in Q} \frac{\log(N) - \log(df(q_i)) \cdot tf(q_i, d) \cdot (k_1 + 1)}{k_1 \cdot ((1-b) + (b \cdot ndl(d))) + tf(q_i, d)}$$

In which *ndl(d)* is the document length of d, divided by the average document length

3 Activities

3.1 BM25 Parameters

The performance of the BM25 ranking algorithm depends greatly on the choice for the values of the parameters *k1* and *b*. Using the CLEF 2001 relevance assessments, we have executed a brute-force search over the parameter space for Dutch, English and Spanish. In Appendix A, plots show the average precision measure as a function of *k1* and *b*. Based on this, we chose to use the parameter values as shown in Table 1 for our runs. Of course we are not sure that the parameter values we found are indeed language dependent. They might be strongly dependent on the document collections used in CLEF, on the way in which we split terms or on something else.

Table 1: BM25 parameter values

	k1	b
Dutch	1.2	0.6
German	1.2	0.6
English	2.0	0.75
Spanish	1.4	0.5
Italian	1.4	0.5
French	1.4	0.5

3.2 Different parts of the topic

We also experimented with using different parts of the topics generate query terms from. Listing 1 shows that each topic consists of three parts: a title, a description and a narrative. We experimented with building queries from the title only, title+description and title+description+narrative. We found that we got the best results using title+description. Probably because using title only yields too few query terms and using all parts yields too many irrelevant terms. We did not test the influence of choosing different settings for the BM25 parameters with using different parts of the topics.

3.3 Synonym expansion

As we described in section 2, we used the Knowledge Concepts semantic network for term expansion during query construction. We found that adding synonyms of query terms to their expansions only resulted in poorer average precision. Looking at the kinds of synonyms that are added we suspect that the problem lies in failing to select the right meaning for a term before adding synonyms and thereby adding many irrelevant terms. For instance, when adding synonyms for the term baby, we could choose to add the nouns *child* and *infant* or the verbs *to nurse* and *to care.*

3.4 Logos vs. Babelfish

In the cross lingual task, we experimented with Logos and Babelfish for translating English topics into Spanish and Spanish topics into English. Of these, Logos yielded the better translations and also the better average query results. For translating topics between English and Dutch, we used Ergane, which translates via Esperanto. With this, translations were not very good and neither were query results.

3.5 Official runs

This year we submitted the following official runs:

- Mono lingual Dutch, German, Spanish, Italian and French, queries built from title only and title+description;
- Cross lingual Spanish to English, English to Spanish, topic translation with Logos and Babelfish;
- Cross lingual Dutch to English and English to Dutch, topic translation with Ergane;
- Multi lingual from English topics, result merging using round robin and merge sort, topic translation with Logos and Babelfish.

4 Results for the Clef 2002

This section presents the improvement made, compared to the run we submitted in 2001. It also presents the results obtained on the monolingual tasks of 2001, compared with the groups that participated that year. The last part of this section presents the comparison of our runs to the median of all runs submitted this year. Only after the conference, a better comparison with the other participants will be possible.

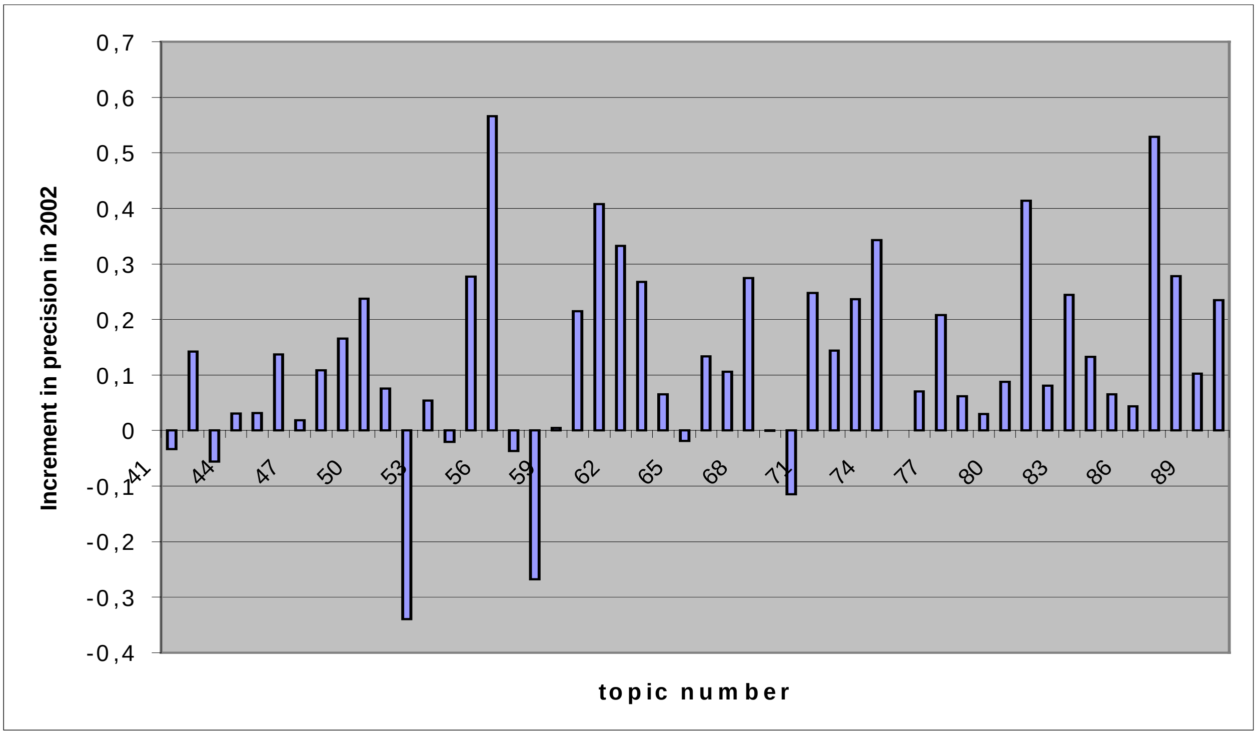


Figure 1: Dutch monolingual task 2001, improvement by using 2002 retrieval algorithm

Figure 1 presents the improvement made on the average precision per topic for the Dutch monolingual task. This considerable improvement has been reached primarily by abandoning the model described in [1] and using a classical BM25 algorithm instead. Additionally, using morphological tools from Knowledge Concepts and our own compound splitter yielded quite a lot too.

Dutch Monolingual	**Participant Name**	**Mean Average Precision**
	TNO	0.3917
	Hummingbird	0.3844
	Thomson L&R	0.3775
	• Océ	0.3767
	APL/JHU	0.3497

French Monolingual	**Participant Name**	**Mean Average Precision**
	U Neuchatel	0.5021
	TNO	0.4877
	Hummingbird	0.4825
	• Océ	0.4450
	Thomson L&R	0.3920

German Monolingual	**Participant Name**	**Mean Average Precision**
	Hummingbird	0.4474
	U Neuchatel	0.4309
	Thomson L&R	0.4205
	U Amsterdam	0.4172
	Eurospider	0.4132
	• Océ	0.3730

Italian Monolingual	**Participant Name**	**Mean Average Precision**
	U Neuchatel	0.4865
	IRST	0.4642
	Hummingbird	0.4555
	TNO	0.4534
	U Amsterdam	0.4485
	• Océ	0.4211

Spanish Monolingual	**Participant Name**	**Mean Average Precision**
	U Neuchatel	0.5800
	Hummingbird	0.5378
	TNO	0.5234
	UC Berkeley 2	0.5225
	Thomson L&R	0.5195
	• Océ	0.4887

The five tables above give an indication where we would have ended in the rank for performing the monolingual tasks last year, with our current algorithms. These figures will be available for 2002 after the conference in September 2002.

Statistics we can produce now are the comparison with the median of all submitted runs. It is important to note that a bilingual run, for instance English to Spanish will be compared to all bilingual runs with Spanish as the document collection language. These are presented in Figure 2.

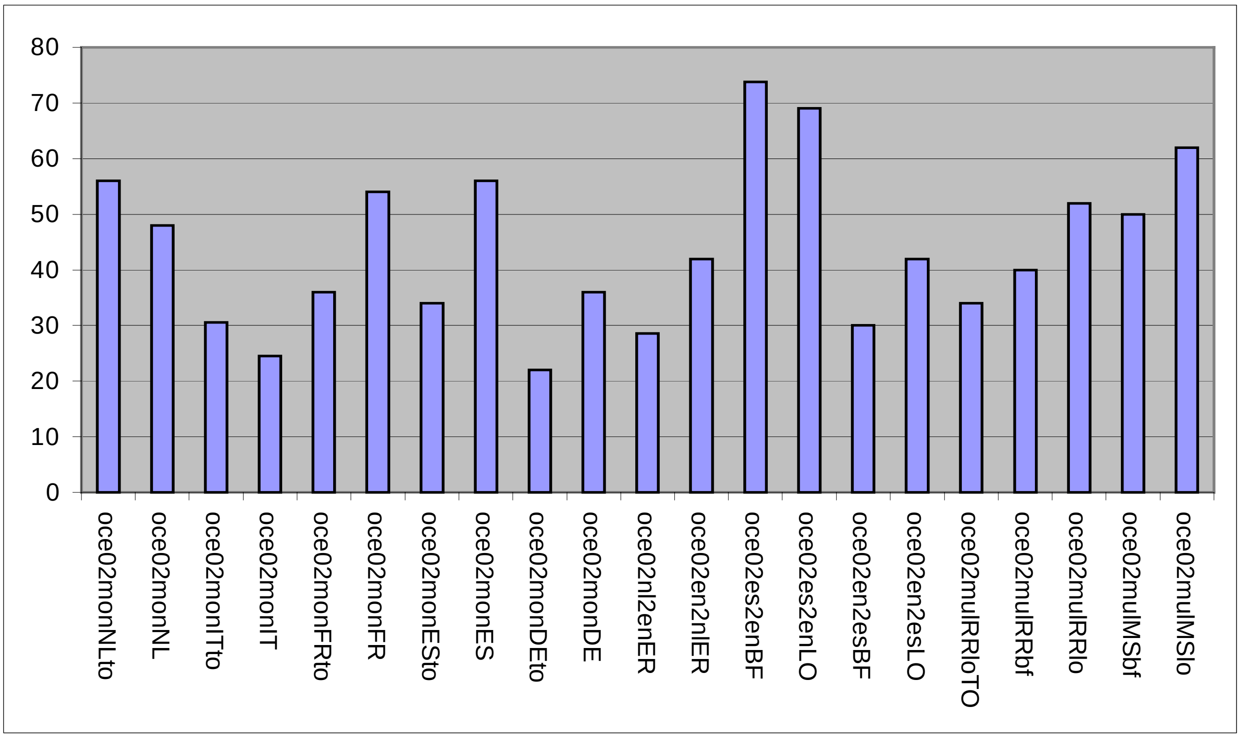


Figure 2: Percentage of topics on or above the median, for each submitted run

Run Name	Description
oce02monNLto	Monolingual Dutch Topic Title Only
oce02monNL	Monolingual Dutch Topic Title + Description
oce02monITto	Monolingual Italian Topic Title Only
oce02monIT	Monolingual Italian Topic Title + Description
oce02monFRto	Monolingual French Topic Title Only
oce02monFR	Monolingual French Topic Title + Description
oce02monESto	Monolingual Spanish Topic Title Only
oce02monES	Monolingual Spanish Topic Title + Description
oce02monDEto	Monolingual German Topic Title Only
oce02monDE	Monolingual German Topic Title + Description
oce02nl2enER	Bilingual Dutch to English Ergane Dictionary Topic Title + Description
oce02en2nlER	Bilingual English to Dutch Ergane Dictionary Topic Title + Description
oce02es2enBF	Bilingual Spanish to English Babelfish Translation Topic Title + Description
oce02es2enLO	Bilingual Spanish to English Logos Translation Topic Title + Description
oce02en2esBF	Bilingual English to Spanish Babelfish Translation Topic Title + Description
oce02en2esLO	Bilingual English to Spanish Logos Translation Topic Title + Description
oce02mulRRloTO	Multilingual English Round Robin Logos Translation Topic Title
oce02mulRRbf	Multilingual English Round Robin Babelfish Translation Topic Title + Description
oce02mulRRlo	Multilingual English Round Robin Logos Translation Topic Title + Description
oce02mulMSbf	Multilingual English Merge Sort Babelfish Translation Topic Title + Description
oce02mulMSlo	Multilingual English Merge Sort Logos Translation Topic Title + Description

Based on the comparison with the median, there is not really much to estimate on how well we currently do, compared to others. It is also true that we did not participate with a revolutionary new information retrieval model, but with a proven approach. A common technique we did not use in our system is Blind Relevance Feedback. This method extracts from the top-n documents retrieved from the initial query terms, and adds them to the original query. This query is then executed, and these retrieved documents are the final ranking. Literature shows that it helps to improve the mean average precision, but we did not manage to implement it.

It is clear that we have participated in many more runs than last year, that we have improved our retrieval algorithms, made a more serious implementation of the indexer and ranker.

5 Conclusions

Our goals for taking part in the CLEF this year were to enter official runs for all tasks and to explore techniques for merging results from different languages. Due to time constraints, we succeeded in doing the former but not the latter.

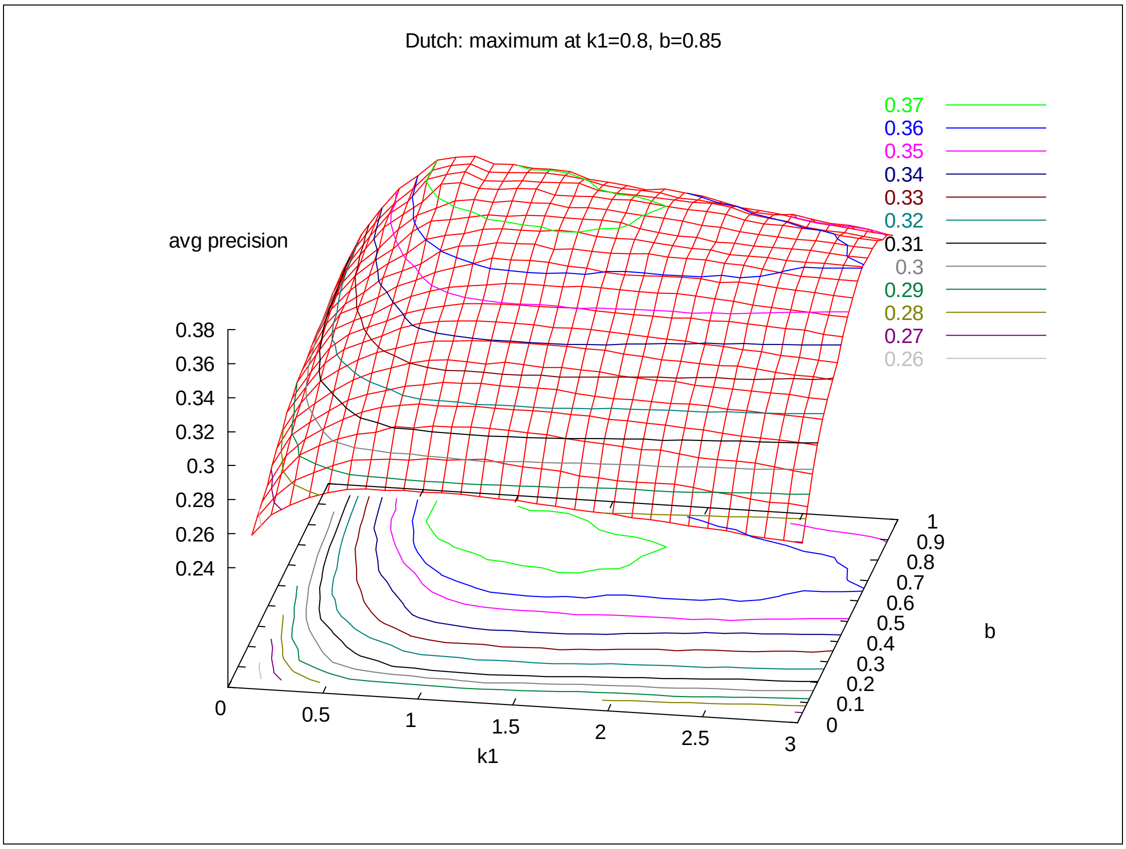
6 References

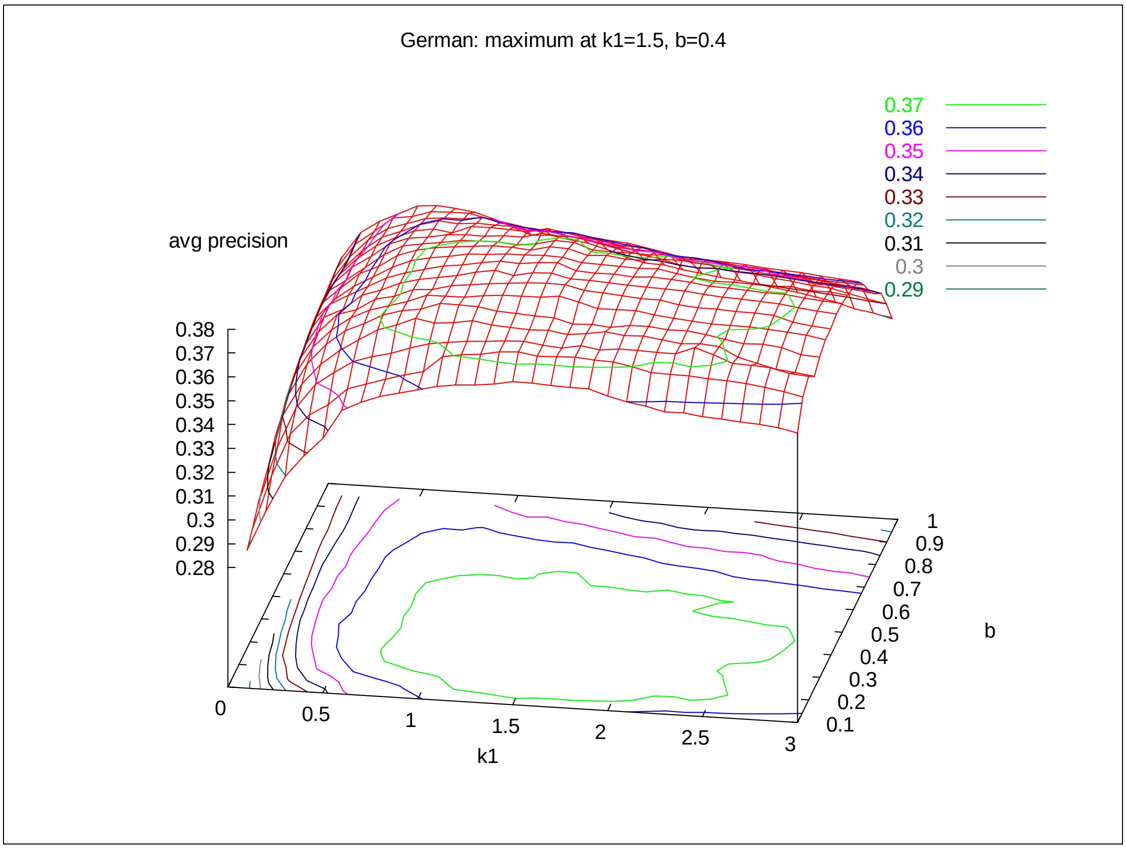
[1] Jakob Klok, Marvin Brünner, Samuel Driessen *Océ retrieval engine at the Cross-Language Retrieval Forum*, Lecture Notes on Computer Science, to appear.

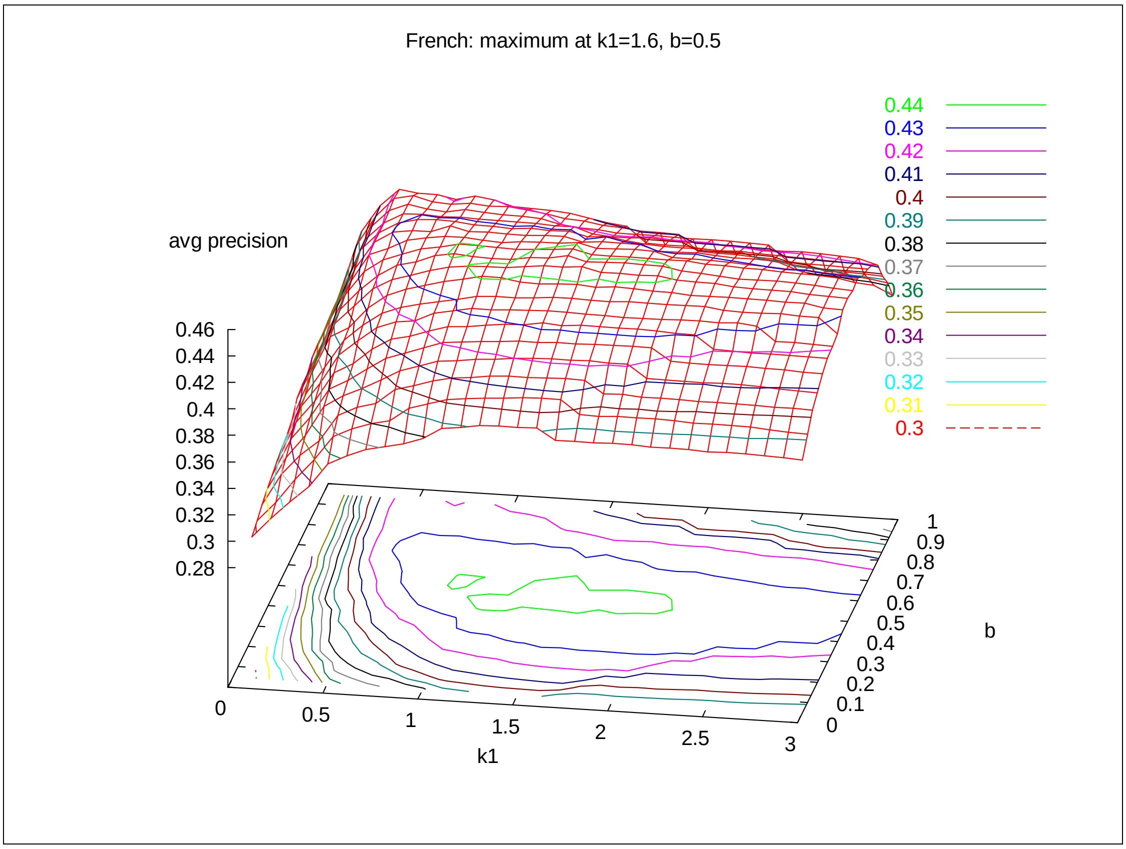
[2] S. Robertson and K. Jones, *Simple proven approaches to text retrieval*, Tech. Rep. TR356, Cambridge University Computer Laboratory, 1997.

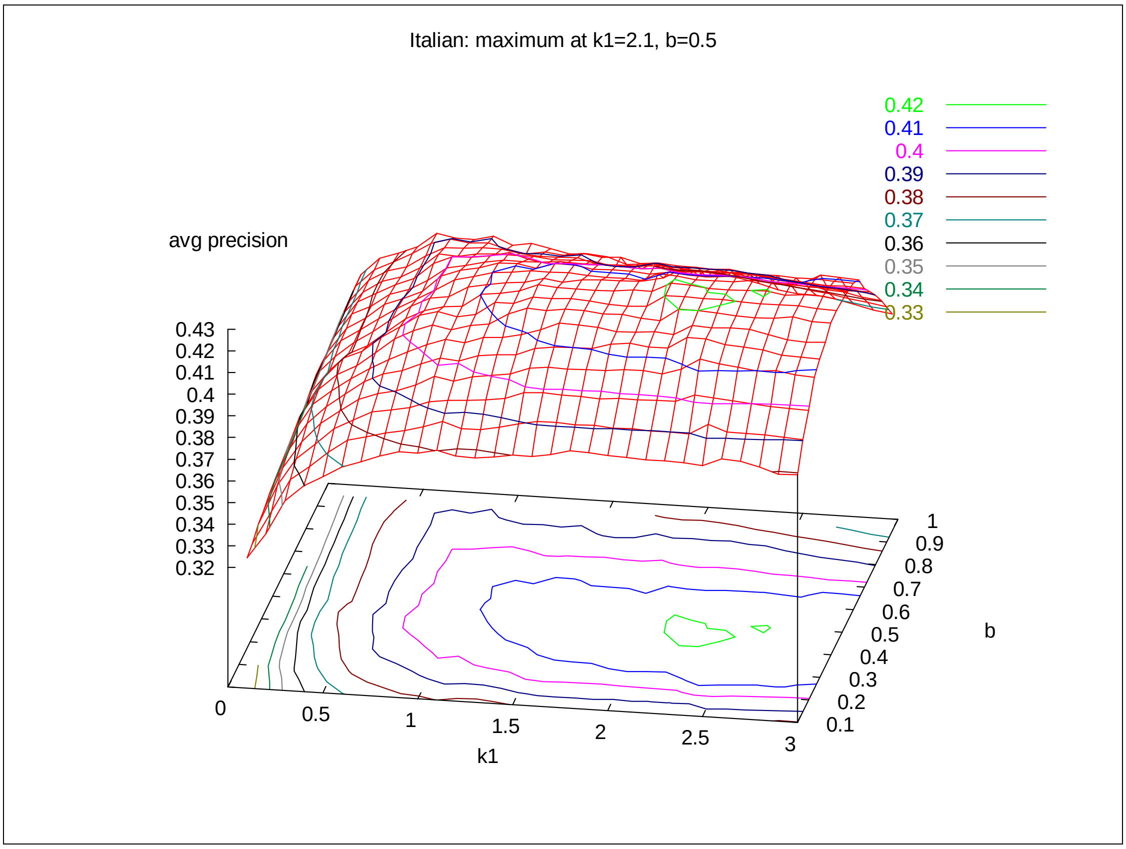
[3] German Stemming Algorithm, http://snowball.sourceforge.net/german/stemmer.html

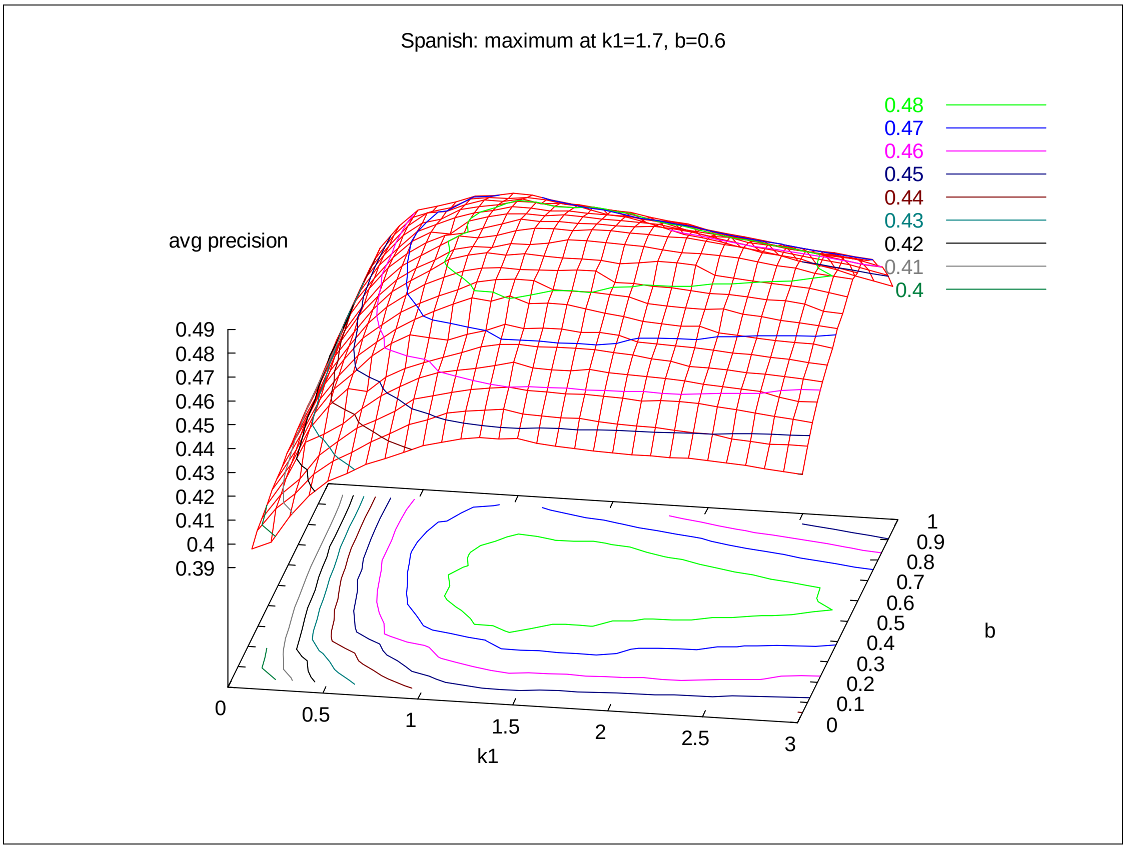
7 Appendix A Parameter optimisation






French: maximum at k1=1.6, b=0.5
avg precision
0.46
0.44
0.42
0.4
0.38
0.36
0.34
0.32
0.3
0.28
0
0.5
1
1.5
2
2.5
3
k1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
b
0.44
0.43
0.42
0.41
0.4
0.39
0.38
0.37
0.36
0.35
0.34
0.33
0.32
0.31
0.3


Italian: maximum at k1=2.1, b=0.5
avg precision
0.43
0.42
0.41
0.4
0.39
0.38
0.37
0.36
0.35
0.34
0.33
0.32
0
0.5
1
1.5
2
2.5
3
k1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
b
0.42
0.41
0.4
0.39
0.38
0.37
0.36
0.35
0.34
0.33


Spanish: maximum at k1=1.7, b=0.6
avg precision
0.49
0.48
0.47
0.46
0.45
0.44
0.43
0.42
0.41
0.4
0.39
0.48
0.47
0.46
0.45
0.44
0.43
0.42
0.41
0.4
0
0.5
1
1.5
2
2.5
3
k1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
b