Evaluation of a Vector Space Similarity Measure in a Multilingual Framework

Romaric Besançon, Martin Rajman

Artificial Intelligence Laboratory
Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland
{Romaric.Besancon,Martin.Rajman}@epfl.ch

Abstract

In this contribution, we propose a method that uses a multilingual framework to validate the relevance of the notion of vector based semantic similarity between texts. The goal is to verify that vector based semantic similarities can be reliably transfered from one language to another. More precisely, the idea is to test whether the relative positions of documents in a vector space associated with a given source language are close to the ones of their translations in the vector space associated with the target language. The experiments, carried out with both the standard Vector Space model and the more advanced DSIR model, have given very promising results.

1. Introduction

The notion of textual similarity is very often used in Natural Language Processing applications designed for the exploration of large textual databases. For example, in Information Retrieval, the relevant documents retrieved by the system are in general the documents the most similar to the user's query, according to the similarity measure used by the search engine (Salton and McGill, 1983). Similarly, in the case of automated clustering, the documents are also clustered according to a given similarity measure (Salton et al., 1975a).

The similarity between documents strongly relies on the choice of the representation method for the texts. The most frequent method is the vector space representation (implemented for instance in IR systems such as SMART (Salton, 1971)). In this approach, a document is represented by a vector in a vector space in which each dimension is associated with a specific linguistic unit, called *indexing term* (e.g. a word, a stem or a lemma).

The vector space representation, simple and easy to implement, has already proven its efficiency in the specific framework of several applications (in particular, information retrieval and automated classification). However, in this paper, we propose a novel validation method, independent from any application.

The idea is to use an aligned multilingual corpus to validate the robustness of the vector space representation of the documents, by showing that the relative positions of the vectors representing the documents in a source language are close to the ones of their translations in a target language. In other words, we try to check if the similarity between any two vectors representing two documents in the source space is close to the similarity between the vectors representing their translations in the target space.

In section 2,we present the two vector space models used for our experiments, the standard vector space model, and the more advanced DSIR model, that integrates more semantic information using co-occurrences profiles; in section 3,we present the data with which the experiments have been carried out; in section 4,we present the first results obtained with a simple test on the similarity matrices. In section 5,we present another validation method and its results, and section 6 provides some preliminary conclusions.

2. Vector Space representations

2.1. The standard vector space model

Within the standard vector space model (VS), each document d is represented by a vector $d^{VS} = (d_1^{VS}, \ldots, d_{|T|}^{VS})$, called *lexical profile*. T is the set of indexing terms and the component d_j^{VS} represents the weight (or importance) in the document d of the indexing term t_j associated with the j^{th} dimension of the vector space. Generally, this weight is a function of the frequency of the term in the document, and also integrates a global weighting and a normalization factor (with respect to the document length). The function used in our experiments corresponds to the *ltn* weighting scheme of SMART (Salton and Buckley, 1988; Singhal, 1997):

$$d_j^{VS} = w_j = idf \times (1 + log(tf)) \quad (1)$$

where tf is the frequency of the indexing term in the document and idf is its inverse document frequency factor $idf = \log \frac{1}{df}$, with df denoting the document frequency of the indexing term. The idf factor allows to give more importance to terms that appear less frequently in the document collection and are therefore more discriminatory. No document length normalization factor have been used (notice however that an implicit normalization is performed through the choice of the cosine similarity, that is independent from the norm).

2.2. The DSIR model

The DSIR model is a vector model that integrates additional semantic information by taking into account the co-occurrences between terms (Rajman and Bonnet, 1992; Rajman et al., 2000; Besançon, 2001).

Within this model, the considered linguistic units u_i are first represented by a vector $c_i = (c_{i1}, \ldots, c_{i|T|})$, called the *co-occurrence profile*, each component c_{ij} of which is the co-occurrence frequency of the linguistic unit u_i with the indexing term t_j. A document d is then represented as the weighted sum of the co-occurrence profiles of the linguistic units it contains, *i.e.* d is represented by a vector $d^{VS} = (d_1^{VS}, \ldots, d_{|T|}^{VS})$ where each d_j^{VS} is defined as:

$$d_j^{VS} = \sum_{u_i \in d} w_i \; c_{ij}$$

where the weight w_i is the same as the one defined by equation (1), for the VS model.

Notice that, in the pure DSIR model, the terms contained in the document are only taken into account through their co-occurrence profile, which often leads to unsatisfactory representations. To correct this property and to also directly take into account the presence of the term in the document, a hybrid DSIR model has been proposed (Rungsawang, 1997; Rajman et al., 2000), in which:

$$d_j^{VS} = \alpha\ w_j + (1-\alpha)\ \sum_{u_i \in d} w_i\ c_{ij} \qquad (2)$$

where α is the hybridization coefficient controlling the relative importance of the occurrence and co-occurrence information.

3. Aligned bilingual data

The goal of this paper is to use an aligned bilingual corpus to validate a vector-based similarity between documents. The aligned bilingual data used for the tests were extracted from the *JOC* corpus, containing written questions and answers of the Official Journal of the European Community, published in several languages. We considered a corpus containing 6729 documents in each languages and for our experiments, we first restricted to French and English.

Formally, we can consider that our aligned bilingual corpus is composed of the corpora $\mathcal{C}$ and $\mathcal{C}'$ respectively in the languages L and L'.

A preprocessing phase has been performed to extract from each of the corpora $\mathcal{C}$ and $\mathcal{C}'$ a corresponding set of indexing terms, denoted respectively T and T':

- the corpora have been analyzed by a lemmatizer based on the syntactic analyzer SYLEX (Constant, 1995): each word has been associated with its part-of-speech and its lemma, the combination of which constitutes the considered linguistic units;
- the produced linguistic units have been filtered according to their *document frequency* (*i.e.* the number of documents containing the considered unit); the indexing set T was defined as the set of linguistic units with document frequency in $\left[\frac{|\mathcal{C}|}{100}, \frac{|\mathcal{C}|}{10}\right]$, where $|\mathcal{C}|$ denotes the number of documents in the corpus $\mathcal{C}$. This interval is often considered as providing terms with good discriminating power (Salton et al., 1975b);
- similarly, the indexing set T' was derived from $\mathcal{C}'$. Notice that the two lexica of linguistic units have been extracted independently.

The same selection process was used for French and English, but in an independent way and each selection is only based on the corpus of the corresponding language. In particular, the two indexing sets are not aligned: the term associated with the i^{th} dimension in one of the vector spaces does not necessarily correspond to the translation of the term associated with the i^{th} dimension in the other vector space.

The sizes of the corpora, the lexica (lemmas of nouns, verbs, and adjectives) and the indexing sets (selected with document frequency in $[70, 700]$) are given in table 1.

	French corpus	English corpus
number of words in the corpus	1 160 877	1 053 945
size of lexicon (number of terms)	25 322	24 469
size of the indexing set	1 062	1 102

Table 1: Sizes of corpora, lexica and indexing sets for the English and French parts of the JOC corpus

The interesting property of a bilingual corpus is the underlying hypothesis that a document and its identified translation are assumed to share a similar semantic content. In such case, if the vector-based similarity is indeed representative of a semantic proximity, then the similarity between two documents should be close to the similarity between their translations.

4. Mantel test

We first propose to test whether the shape of the global representation of the document collection is stable when changing from one language to another. One way of representing this global shape is the matrix of pairwise similarities between the documents.

The comparison of two similarity matrices can be performed with the Mantel test (Mantel, 1967; Legendre, 2000). This test is a widely used method for assessing the relationships between two distance matrices or, more generally, two resemblance or proximity matrices. It involves a measure of the association between the elements in two matrices by a statistic r, and then assesses the significance of this measure by comparing it with the distribution of the values found by randomly reallocating the order of the elements in one of the matrices. The statistic r used in our experiments is the sum of the products of the standardized similarities; for two similarity matrices A and B of size N, it is defined by:

$$r\ =\ \frac{1}{N^2-N-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{(a_{ij}-\overline{a})}{\sigma_a} \frac{(b_{ij}-\overline{b})}{\sigma_b}$$

where $\overline{a}$ (resp. $\overline{b}$) is the average similarity in matrix A (resp. B), and σ_a (resp. σ_b) is the standard deviation of the similarities in matrix A (resp. B). The statistic r ranges from -1 to 1, respectively indicating perfect negative and positive correlation. The zero value indicates no correlation.

In our experiments, we used the implementation of the Mantel test by Eric Bonnet (Bonnet, 2001). For the English and French corpora, we obtained a correlation value $r = 0.5463$ with significance 0.0001. This first result provides us with the confirmation of a positive correlation between the matrices of the similarities between the documents in each language. Additional experiments were how-

ever necessary to analyze the precise nature of this correlation.

5. Nearest translation test

To further test the stability of the representations by translation, we checked whether the translation of a given document d can be reliably recovered as the translated document the most similar to d. We call this test the *nearest translation test*.

In order to compute a similarity between a document and its translation, we first represent any document by the vector of its similarities with the elements of an aligned reference document set and then take as a measure of the similarity between a document d in language L and a document d' in language L' the similarity between the similarity vector associated with d and the similarity vector associated with d'.

The stability test procedure is then the following:

1. we build aligned bilingual test and reference sets;
2. we represent the source (resp. target) documents of the test set by their similarities with respect to the source (resp. target) documents of the reference set;
3. we compare the obtained representations to check how often the translation of a source document is indeed the nearest target document.

These three steps are presented in more details in the following sections.

5.1. Building the test and reference sets

From the available corpus, we randomly extract a number n of documents ($n = 500$ in our experiments), which constitute the test set. The $N = |\mathcal{C}| - n$ remaining documents constitute the reference set.

We denote:

- d_i the n documents of the test set TEST-L in the source language L;
- d'_i the n documents of the test set TEST-L' in the target language L';
- D_i the N documents of the reference set REF-L in the source language L;
- D'_i the N documents of the reference set REF-L' in the target language L';

5.2. Document representation

All documents in the test and reference sets are first represented in the vector space associated with their own language. Notice that in this case, two documents from different language are not directly comparable, since they are not represented in the same space.

However, to each document d_i (resp. d'_i) from the test set TEST-L (resp. TEST-L'), we can associate a similarity vector in which the j^{th} component is the similarity between this document and the j^{th} document from the reference set REF-L (resp. REF-L').

Let us denote $V_s(d_i)$ (resp. $V_s(d'_i)$) this similarity vector; we then have:

$$V_s(d_i) = (\delta(d_i, D_1), \ldots, \delta(d_i, D_N))$$
$$V_s(d'_i) = (\delta(d'_i, D'_1), \ldots, \delta(d'_i, D'_N))$$

where δ is the chosen similarity. In our experiments, we took the cosine similarity, defined, for two document representations d and d', by:

$$\delta_{cos}(d, d') = \frac{d \cdot d'}{||d||\ ||d'||}$$

The similarity vector characterizes the relative position of the considered document with respect to the documents in the reference set. Since the reference sets are aligned, the similarity vectors computed in different languages are still comparable. They can therefore be used to test the proximity between a document and its translation and then to test the invariance by translation of the position of a document with respect to the reference set.

This representation of documents from different languages with respect to a reference set is close to methods used in multilingual information retrieval, such as the Generalized Vector Space Model (Carbonell et al., 1997; Yang et al., 1998) or to the application of the *Latent Semantic Indexing* model (LSI) to multilingual IR (Dumais et al., 1996; Littman and Jiang, 1998).

5.3. Invariance test

For each document $d_i \in$ TEST-L, we consider the similarity between the associated similarity vector $V_s(d_i)$ and each of the similarity vectors associated with the n documents in TEST-L'. The similarities between similarity vectors are also evaluated using the cosine measure. The idea, illustrated by figure 1, is that if the invariance by translation is indeed verified, then the similarity between $V_s(d_i)$ and $V_s(d'_i)$ should be significantly larger than any of the similarities between $V_s(d_i)$ and $V_s(d'_j)$ with $j \neq i$.

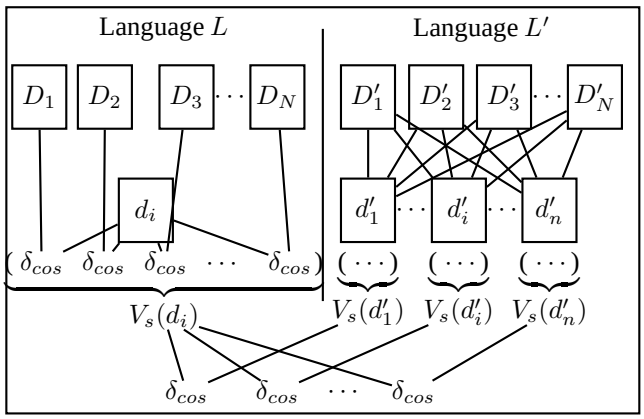

Figure 1: Validation method

In order to test this hypothesis, we count the number of documents $(d'_j)_{j \neq i}$ in TEST-L' for which the similarity $\delta_{cos}(V_s(d_i), V_s(d'_j))$ is smaller than the similarity $\delta_{cos}(V_s(d_i), V_s(d'_i))$, *i.e.* the number of document $(d'_j)_{j \neq i}$

that have a relative position (with respect to the reference set) that is further away from the relative position of d_i than the one of d'_i. We denote f_i the proportion of such documents in TEST-L', we have:

$$f_i = \frac{1}{n-1} \Big| \{d'_k \in \text{TEST-}L' \mid k \neq i \text{ and } \delta_{cos}(V_s(d_i), V_s(d'_k)) < \delta_{cos}(V_s(d_i), V_s(d'_i))\} \Big|$$

To test whether only few documents in language L' are closer to a document d_i than its translation (according to the chosen measure), we perform a statistical test on this proportion. More precisely, we want to test that the proportion f_i is significantly larger than an a priori given threshold p_0, which corresponds to the following alternative:

$$\begin{cases} H_0 \ : f_i = p_0 \\ H_1 \ : f_i > p_0 \end{cases}$$

For a large enough sample (in our case, $n = 500$), we reject the null hypothesis H_0 at an error level α if (Grais, 1986, p. 261):

$$\frac{f_i - p_0}{\sqrt{p_0(1-p_0)}} \times \sqrt{n} > t_\alpha$$

where t_α is the value of the standard normal random variable for an error level α. This leads to reject the hypothesis H_0 if:

$$f_i > p_0 + t_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} \tag{3}$$

5.4. Results

For the statistical test, we chose $p_0 = 0.9$. The considered hypothesis is then that the proportion of documents having a relative position closer than the one of the translation is less than 10%. The error level was chosen at $\alpha = 2.5\%$, which corresponds to $t_\alpha = 1.96$. With these parameters, the equation (3) is then:

$$f_i > 0.9 + 1.96\sqrt{\frac{0.09}{500}} = 0.9263 \tag{4}$$

and one can verify that this leads to accept the null hypothesis H_0 if there is more than 37 documents (out of 500) which are "better" than the translation itself.

In our experiments, we produced the following results:

- the number of documents N_R (out of the 500) for which the hypothesis H_0 is accepted, *i.e.* the documents for which the invariance hypothesis is not verified;
- the number of documents N_0 (out of the 500) for which the translation is *the* document the most similar to the source document ($f_i = 0$), *i.e.* the documents for which the invariance hypothesis is perfectly verified;

For the chosen languages (French and English), the test has been performed in both directions:
- *fr-en*: $\{L$ =French, L' =English$\}$
- *en-fr*: $\{L$ =English, L' =French$\}$

The table 2 presents the results obtained with a standard vector space (using the *ltn* weighting scheme) and with the hybrid DSIR model, for several hybridization values. The value $\alpha = 1$ corresponds to the standard VS model. All results correspond to average values on 30 independently extracted random test sets.

From table 2, one can see that the invariance hypothesis is accepted for more than 99% of the documents within the VS model (99.4% for *fr-en*, 99.6% for *en-fr*). In addition, the results show that for 95% of the documents, the translation is indeed *the* document most similar to the source document (according to the chosen measure) among the 500 documents of the test set.

These results constitute a strong support evidence for the fact that the relative position of the documents does not vary significantly when translated from one language to another. This is in itself an interesting confirmation of the robustness of a simple vector-based similarity measure between documents.

Notice that this result was not obvious, in particular because the indexing sets are selected independently. In other words, even if the representation of the documents in each language only depends on the corpus in this language, the model leads to a representation of the corpus that is stable from one language to another.

Although the results obtained for the standard vector space model are already very good, it is interesting to notice that the hybrid DSIR model further improves them. This model leads to a proportion of documents for which the invariance hypothesis is verified to almost 99.8% (see figure 2). The best hybridization value seems to be between 0.5 and 0.6 but the representation based only on co-occurrences ($\alpha = 0$) give very bad results. A possible explanation for this is that representing the documents only with co-occurrence profiles tends to smooth the representations: vectors in the VS model are very sparse (the only non-zero values are the one associated to terms contained in the document), whereas vectors in the DSIR model are average of co-occurrence profiles, and therefore correspond to less discriminatory profiles. Nevertheless, the experiments show that the co-occurrence information is useful and can be exploited through the hybrid model.

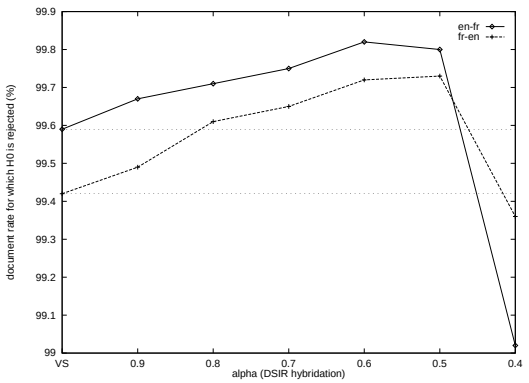

Figure 2: Impact of the hybrid DSIR model on the validation of the relevance of a vector-based similarity in a multilingual framework

	en-fr		*fr-en*	
α	N_R	N_0	N_R	N_0
1 (VS)	2.03 (0.41%)	476.47 (95.3%)	2.9 (0.58%)	473.7 (94.74%)
0.9	1.67 (0.33%)	476.3 (95.26%)	2.57 (0.51%)	474.4 (94.88%)
0.8	1.47 (0.29%)	477.5 (95.5%)	1.97 (0.39%)	475.77 (95.15%)
0.7	1.23 (0.25%)	**478.67 (95.73%)**	1.73 (0.35%)	476.63 (95.33%)
0.6	**0.9 (0.18%)**	478.33 (95.67%)	1.4 (0.28%)	**476.7 (95.34%)**
0.5	1 (0.2%)	474.6 (94.92%)	**1.33 (0.27%)**	473.37 (94.67%)
0.4	4.9 (0.98%)	449.83 (89.97%)	3.2 (0.64%)	452.1 (90.42%)
0	352.43 (70.49%)	40.8 (8.16%)	344.87 (68.97%)	52.1 (10.42%)

Table 2: Validation results for the similarity measure within the VS model and the DSIR model

5.5. Tests with other languages

To further test the generality of our invariance hypothesis, we performed the tests with all the eight languages for which aligned data was available in the JOC corpus: French, English, Italian, Spanish, Portuguese, German, Dutch and Danish. For all these languages (but English and French), we did not have preprocessing advanced tools. We therefore performed a simple stemming of the corpus (using the Snowball stemmers (Porter, 2001)) and used stoplists to filter out the common words.

The documents were then represented in eight independent vector spaces, using the standard vector space model. We performed the nearest translation test for each pair of languages (in both directions), and the average values of N_R and N_0 for each pair are presented in table 3.

The average value on all pairs of languages show that the hypothesis is verified for almost 98.7%, and the number of documents for which the translation is the closest document is more than 83.8%. Hence, the relevance of the vector-based similarity seems to hold for the other languages as well.

A more detailed analysis shows that some languages seems to be closer than others (*i.e.* for these languages, the invariance hypothesis is better verified): in particular, French, English, Italian, Portuguese and Spanish seem to form a set of "close" languages whereas German, Dutch and Danish lead to less significant results. This tendency is also confirmed by the correlations found with the Mantel test[1], presented in table 4.

However, these observations have to be moderated since the results obtained depend on several parameters, including the quality of the translations and the quality of the preprocessing phase (*i.e.* the stemmers). Both have not been evaluated in these experiments.

6. Conclusion

The notion of textual similarity, fundamental for numerous NLP systems, is still to be validated. In this paper, we propose a novel validation method relying on the invariance by translation of the relative positions of the documents represented in a vector space. This validation is interesting as it confirms the hypotheses made on the robustness of a vector-based similarity measure.

The validation method can also be used to compare different representation models: for instance, we showed that the DSIR model integrating co-occurrences improves the relevance of the textual similarity.

However, the results presented are global quantitative results. A more detailed analysis should be carried out, especially in order to identify the properties of the documents for which the null hypothesis cannot be rejected, and to understand why the DSIR model might improve the representation of these documents.

The method presented also have applications for the automated comparison of machine translation systems or for multilingual information retrieval, and further experiments for these applications should be undertaken.

[1]Due to a lack of time, the values presented here have been calculated only on a random subset of 672 documents (10% of the collection), and have a significance 0.001

7. References

Romaric Besançon. 2001. *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes*. Ph.D. thesis, Swiss Federal Institute of Technology, Lausanne.

Éric Bonnet. 2001. simple mantel test – version 1.2. http://www.geocities.com/eb_ce/mantel.html.

Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. 1997. Translingual information retrieval: A comparative evaluation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 708–715.

Patrick Constant, 1995. *Manuel de développement SYLEX-BASE*. INGÉNIA-LN, Paris, France.

Susan Dumais, Thomas Landauer, and Michael Littman. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In *SIGIR'96 - Workshop on Cross-Linguistic Information Retrieval*, pages 16–23, August.

Bernard Grais. 1986. *Méthodes Statistiques*. Dunod.

P. Legendre. 2000. Comparison of permutation methods for the partial correlation and partial mantel tests. *Journal of Statistical Computation and Simulation*, 67:37–73.

Michael Littman and Fan Jiang. 1998. A comparison of two corpus-bases methods for translingual information retrieval. Technical Report CS-1998-11, Department of Computer Science, Duke University, Durham, North Carolina 27708-0129, June.

L' / L	Italian	Spanish	Portuguese	French	English	German	Dutch	Danish
Italian	–	**0.223%**	**0.298%**	**0.342%**	**0.253%**	1.82%	1.7%	2.2%
	–	**94.3%**	**93.2%**	**94%**	**92.6%**	76.3%	80%	78%
Spanish	**0.699%**	–	**0.372%**	**0.595%**	**0.476%**	1.67%	1.67%	1.96%
	95.6%	–	**97.4%**	**96.2%**	**94.8%**	85.9%	88.9%	86.6%
Portuguese	**0.58%**	**0.193%**	–	**0.298%**	**0.342%**	1.49%	1.31%	1.53%
	89.8%	**94.2%**	–	**93.1%**	**91.7%**	75.6%	79.4%	78.1%
French	**0.432%**	**0.312%**	**0.283%**	–	**0.298%**	2.01%	1.67%	2.16%
	96.4%	**96.8%**	**97.2%**	–	**95.6%**	85%	88.5%	85.7%
English	1.47%	**0.565%**	**0.699%**	**0.938%**	–	3.27%	2.56%	3.07%
	89.7%	**92.3%**	**92.3%**	**92%**	–	73.3%	80.1%	77.7%
German	3.41%	1.55%	1.96%	1.86%	1.79%	–	1.12%	1.56%
	69.7%	74.3%	74.6%	75%	75.3%	–	81.5%	79.2%
Dutch	2.74%	1.32%	1.64%	1.61%	1.43%	**0.744%**	–	1.1%
	64.3%	70.2%	71.1%	70.2%	73%	73.6%	–	73.8%
Danish	2.59%	1.28%	1.43%	1.82%	1.47%	1.01%	**0.982%**	–
	75.8%	81.5%	82.2%	81%	82.5%	83.7%	85.1%	–
							average	1.32414%
								83.8547%

Table 3: Results (N_R and N_0 values) of the validation for 8 different languages, using the VS model: the results in bold font are the ones verifying $N_R \leq 1\%$ or $N_0 \geq 90\%$.

L1 / L2	Spanish	Portuguese	French	English	German	Dutch	Danish
Italian	**0.6**	**0.599**	**0.628**	**0.57**	0.464	0.478	0.466
Spanish		**0.641**	**0.62**	**0.57**	0.479	**0.502**	0.491
Portuguese			**0.638**	**0.573**	0.484	**0.51**	**0.511**
French				**0.593**	0.491	**0.506**	0.487
English					0.474	0.495	0.483
German						**0.52**	**0.509**
Dutch							**0.519**

Table 4: Mantel test correlations r for 8 different languages: the results in bold font are the ones verifying $r \geq 0.5$

N. Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27:209–220.

M.F. Porter. 2001. Snowball: A language for stemming algorithms. http://snowball.sourceforge.net/

Martin Rajman and Alain Bonnet. 1992. Corpora-base linguistics: new tools for natural language processing. In *1st Annual Conference of the Association for Global Strategic Information*, Bad Kreuznach, Germany.

Martin Rajman, Romaric Besançon, and Jean-Cédric Chappelier. 2000. Le modèle DSIR : Une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement Automatique des Langues*, 41(2):549–578.

Arnon Rungsawang. 1997. *Recherche Documentaire à base de sémantique distributionnelle*. Ph.D. thesis, ENST, Paris.

Gerard Salton and Chris Buckley. 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523.

Gerard Salton and M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw Hill.

Gerard Salton, A. Wong, and C. S. Yang. 1975a. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Gerard Salton, C. S. Yang, and Clement T. Yu. 1975b. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44.

Gerard Salton, editor. 1971. *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall.

Amit Singhal. 1997. *Term Weighting Revisited*. Ph.D. thesis, Department of Computer Science, Cornell University.

Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, and Robert E. Frederking. 1998. Translingual information retrieval: Learning from bilingual corpora. *Artificial Intelligence*, 103(1–2):323–345.