Identifying parallel corpora using Latent Semantic Indexing

Yuliya Katsnelson
Highland Technologies
4831 Walden Lane
Lanham, MD 20706
Tel: 301-306-2849 Fax: 301-306-8201
ykatsnelson@htech.com

Charles Nicholas
UMBC
1000 Hilltop Circle
Baltimore, MD 21250
Tel: 410-455-2594 Fax: 410-455-3969
nicholas@cs.umbc.edu

Abstract

Identifying parallel corpora can be an important step in a variety of tasks related to information retrieval. However, at present to identify parallel corpora requires human experts to examine the texts and evaluate their respective contents. We assume in this research that texts which are translations of each other have similarities in their semantic structure which are absent between independent documents.

Latent Semantic Indexing (LSI) (Landauer 1989, Deerwester 1990) is a statistical technique that brings out correlations between documents based on term co-occurrence patterns identified using the method of Singular Value Decomposition. LSI does not involve any knowledge of the actual content of the documents and therefore has no need for human intervention. If LSI could be used for parallel corpora identification, it would lower the costs incurred in this task. Our purpose is to determine if it is possible to identify a parallel corpus using the method of Latent Semantic Indexing.

We present evidence that LSI reveals similarities between parallel documents that do not exist between non-parallel documents and is therefore useful for identifying parallel corpora.

1 Introduction

Parallel corpora identification can be considered an area of Cross-Language Information Retrieval with applications both inside and outside of the area of IR. Recognizing the relevance of documents in multiple languages also includes recognizing the relevance of parallel documents. Parallel documents, in fact, make good test cases for evaluating the effectiveness of a Cross-Language IR system, since if a document is relevant to a query, then its translation should also be relevant. Also, being able to tell the parallel part of a corpus from the rest of the corpus allows one to reduce the amount of search, since only one portion of the parallel subsection needs to be explored.

Parallel corpora analysis, on the other hand, relates largely to the area of document translation, human- or computer-aided. Parallel documents are included in a corpus in order to provide the reader with a choice of language in which to read the document. This means that the different versions of the document need to convey the same information. This objective introduces a number of questions. Is the corpus parallel? What part of the corpus has a corresponding parallel subset in the same corpus? How close are the various multilingual translations? Which translation is the best among all the available versions in the same language? In this research we aim to answer the first question, believing that the insight that we get in the process will enable us to approach the others.

Latent Semantic Indexing (LSI) is a statistical technique that uses the method of Singular Value Decomposition (SVD) to represent the terms and documents in the corpus as points in a multi-dimensional space. The dimensions of this space represent various patterns of term co-occurrence. Thus, documents that have similar characteristics with respect to a particular term co-occurrence pattern have similar coordinates in the corresponding dimension of the LSI space. SVD allows a reduction of the dimensionality of the LSI space with minimal information loss.

We report on the results of applying Latent Semantic Indexing to identifying parallel corpora. In our research we used the unsupervised learning approach, relying on the hypothesis that the content "signature" of the collection is going to outweigh the difference in language-related "noise" and still

supply meaningful results. As measures for comparing the LSI representations for parallel and non-parallel corpora we use correlation coefficient analysis and visual inspection.

We used corpora in the English, French, Russian and Italian languages to assess the behavior of this method in the cases of more similar (English/French) and less similar (French/Russian) language pairs. The corpora included abstracts from "The Little Prince" by A. de Saint Exupery, short stories and the novel "Smoke Bellew" by J. London, "The Adventures of Sherlock Holmes" by Sir A. Conan Doyle, and "The Black Tulip" by A. Dumas. Our choices of documents were driven largely by their availability in different languages.

We present evidence that LSI reveals relationships between parallel documents that do not exist between independent documents and thus, provides a means of identifying parallel corpora.

The remainder of this paper is organized as follows. Section 2 lists the necessary definitions and gives a brief description of the Vector Space model and Latent Semantic Indexing. Section 3 explains the research hypotheses and experimental design. Section 4 discusses the results, and Section 5 gives the conclusions and outlines future work.

2 Definitions and Background

In this section we introduce the main concepts and methods used in the course of our research. We also briefly describe the Vector Space model (VSM) of Information Retrieval and explain in more detail the method of Latent Semantic Indexing (LSI) and its advantages as compared with the VSM. We will also discuss the reasons behind choosing LSI for the task of parallel corpora identification.

2.1 Definitions

Before going any further, we list the definitions of concepts that are going to be referred to throughout the paper.

1. *Parallel corpus* – a collection of documents in which at least a subset of documents has translations within the corpus.
2. *Fully parallel corpus* – a corpus in which each document in the corpus has a translation within the corpus.
3. *Partially parallel corpus* – a corpus in which only a subset of documents has a translation within the corpus.
4. *Literal translation* – translation in which the goal is to remain as close to the original choice of words and structures as possible.
5. *Literary translation* – translation in which the goal is to remain as close to the original meaning and imagery as possible.
6. *Synonymy* – different terms are used to describe the same concept.
7. *Polysemy* – the same term describes more than one concept.
8. *Mixed LSI space* – all documents in the corpus are used to create the LSI space for further analysis.
9. *Separate LSI space* – a separate LSI representation is created for each monolingual sub-collection of the corpus under consideration.
10. *Cognates* – words in different languages that are spelled similarly and have the same meaning.

2.2 Vector space model

The three main models of IR are the Boolean, vector space and probabilistic models. We are only going to discuss the Vector Space model in this paper as the most relevant to the issue in question. Information on this and other models may be found (Baeza-Yates 1999).

In the vector space model, both documents and queries are represented as vectors. The vector components are term frequencies, often normalized by the vector length to account for different lengths of documents. There are a number of methods of computing similarity between these vectors. The most popular is the cosine of the angle between the vectors.

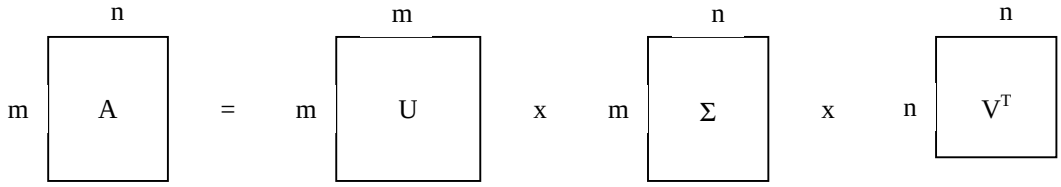
The vector space model presents more flexibility for evaluating the degree of similarity between documents than the Boolean model (Baeza-Yates 1999). However, it does not account for the fact that the occurrence of one query term in the document may influence the likelihood of the occurrence of another term. For instance, if the query is "computer networks security", then the fact that the term "computer" occurred in a document make it more likely that the term "networks" occurs in the same document than, for instance, the term "knitting".

2.3 Latent Semantic Indexing

Latent Semantic Indexing (LSI) can be considered an extension of the vector space model. LSI claims that the meaning of a document is not determined by its set of terms, but rather by a latent semantic structure that manifests itself by how each term is used with all other terms across the entire corpus. This means that replacing one (or more) term(s) in this structure with their synonyms does not change the meaning of the document, provided that the latent semantic structure remains intact. The goal of LSI then is to bring forth this latent semantic structure (Landauer 1989).

The entrance point for this method is constructing the term-document matrix A for the corpus. The size of the matrix A is [m x n] where m is the number of terms in the corpus and n is the number of documents. Element a_{ij} is the frequency of occurrence of term i in document j.

Once the term-document matrix is constructed, the LSI method uses Singular Value Decomposition (SVD) to create the LSI space. The result of the SVD is a 3-tuple of matrices U, Σ, and V^T such that:



where matrices U and V are orthonormal (i.e. $U*U^T = V*V^T = I$), and matrix S is diagonal (i.e. $a_{ij} = 0$, $\forall\ i \neq j$).

Matrix U has size [m x m]. An element of U u_{ij} contains the coordinate of term i in dimension j of the LSI space. Matrix V^T has size [n x n]. An element of V^T v_{ij} contains the coordinate of document j in dimension i of the LSI space.

The matrix S contains the singular values sorted in descending order. The size of this matrix is [m x n]. The number of non-zero diagonal elements in this matrix equals the rank of the original term-document matrix, which can be at most min(m,n). Therefore, only min(m,n) singular values at the most are significant for the analysis. Also, since the singular values decrease as they go down the main diagonal of S, the dimensions corresponding to those values are of decreasing significance. Any first k dimensions of the LSI space represent the best k-approximation of the original LSI space (Berry 1995). However, it may not be a good practice to establish a preset cutoff point for the number of dimensions k under analysis. In the course of our experiments we established $10 \leq k \leq 15$ as a number of dimensions that appeared to be the most useful for the purpose of parallel corpora identification.

3 Method and experimental design

The goal of the present research is to see whether it is possible to identify parallel corpora with computational means rather than through inspection by a trained multi-lingual human professional. We assume that given an original document A in a particular language, any adequate translation of this document into another language is going to possess some similarity to the original document. The hypothesis that is based on this assumption is that Latent Semantic Indexing (LSI) is a suitable method for showing the similarities between parallel documents.

Thus, the approach taken in the course of this research is to represent both parallel and non-parallel document collections in LSI space and measure the similarities between corresponding documents. In this section we discuss the rationale behind applying LSI to identifying parallel corpora and describe the particular ways LSI was applied for this purpose. We are also going to describe the experimental strategy taken and the data that we used for the experiments.

3.1 Method

The intuitive way of parallel corpora identification is to create some kind of a dictionary, which provides a mapping between a term in one language and its equivalents in other languages. One problem with such a method is that it requires knowledge of what particular languages appear in the corpus. It is also subject to *synonymy* and *polysemy* problems. A good method would allow analyzing a corpus without needing to know the languages of the documents that comprise it.

Latent Semantic Indexing has one property that is very attractive for processing multilingual corpora: No knowledge of the term meanings is required for the analysis, thus eliminating the need for dictionaries. Once the term-document matrix has been constructed, the only "reality" that exists in the analysis space is the frequencies of term occurrences with no inherent meaning associated with those frequencies. Of course, this property can also be construed as a shortcoming of this method, since some built-in knowledge could be useful in some situations. For instance, it is possible that there exist two documents with completely unrelated contents, but very similar or identical lexical structure. In that case, LSI will regard the two documents as similar, despite their semantic differences. However, in our experience such occurrences are unlikely. The LSI method provides us with the means to compare the co-occurrence patterns within the corpora in ways that will hopefully bring out the similarities between parts of the corpus that belong to different languages.

3.2 Structure of LSI space

The critical question that needs to be answered when applying the Latent Semantic Indexing method is how to construct the LSI space.

The issue of finding translations (mates) within a corpus also relates back to the work of M. Littman, S. Dumais and T. Landauer on Cross-Language Information Retrieval using LSI (Littman 1996). Their approach was to create a training corpus in which each document consisted of corresponding multi-lingual documents pasted together. This technique allowed them to establish co-occurrence mappings between languages. Thus, when a newly introduced query produced a pattern similar to the one in the training corpus, the documents that had similar co-occurrence characteristics in all languages participating in the training corpus were returned as relevant.

Our assumption that translations of the same document are similar in ways that independent documents are not allows us to use an alternative strategy, i.e. to eliminate the training corpus, and consider only the correlation between coordinates in LSI space of multi-lingual documents.

We used two different ways of mapping documents from a multilingual corpus into the LSI space. One way was to process the entire multilingual corpus simultaneously and construct a single term-document matrix. This means that every vector in matrix U of the LSI decomposition contains the values for all the terms, and that the LSI space will be based on the term co-occurrence patterns not only across documents, but across language boundaries as well. Such a space is called *mixed LSI space* throughout this paper. The second way was to create a *separate LSI space* for every monolingual part of the corpus and compare the representations of corresponding documents or document vectors in the same way as in the case of mixed LSI space. We used both mixed LSI space and separate LSI space approaches in this work, but an in-depth analysis of their respective merits is beyond the scope of this paper.

This research does not rely on the existence of *cognates* (i.e. words in different languages that are spelled similarly and have the same meaning). In the case of the mixed LSI space, every column of the term-document matrix contains frequencies, raw or weighted, for every term in the corpus. If cognates were present in significant numbers, then the actual frequency with which a given term occurred in patterns in a particular language would be obscured. The assumption for this analysis is that no or very few cognates exist in the corpora and they have no significant impact on its outcome.

3.3 "Mate" dimensions

Since we are interested in measuring parallelism of corpora, rather than concentrating on individual documents, the objects of our analysis are the vectors of document coordinates within a particular dimension, rather than the coordinates of a particular document in the LSI space.

By comparing the monolingual parts of these vectors to each other, we claim that if the corpus is, indeed, parallel, then there is at least one pair of dimensions in the LSI space in which the corresponding monolingual sub-patterns are similar. This claim implies that the sub-vectors of different dimensions may be mated, i.e. the patterns in one language reflected in one dimension may be coupled with a similar pattern in another language. Therefore, for some, if not all dimensions there exists a "mate" which has the patterns of the multilingual sub-vectors that are similar in a quantifiable way. This hypothesis is a "twist" on the idea of finding mates developed by Dumais et. al. (Dumais 1997).

In the case of separate LSI spaces, one can argue that, considering the inherent similarities in the sub-corpora of a parallel corpus, there would be also similarities in their LSI representations that would

manifest themselves despite the grammatical differences between the languages in which these sub-corpora are expressed. This might mean that the idea of mate dimensions applies also to separate LSI, although in this case there is no common “context” that was created in the case of mixed LSI by processing the entire corpus into a single LSI space.

A special case of this "mate dimension" hypothesis exists for languages with similar syntactic structure (e.g. English and French). That is, the closer the grammatical structures of the languages, the more the "mate" dimensions converge to the point where in some dimensions the sub-vectors, corresponding to each language form a similar pattern.

3.4 Data

We constructed a parallel corpus from various digital libraries. We included literary works by Jack London, Sir Arthur Conan Doyle, A. de Saint-Exupery, A. Dumas pére, and the United Nations corpus. The experiments were performed on the following language combinations: English/French, English/Russian, French/Russian, and English/Russian/Italian.

The texts were divided into abstracts of 550-600 words (about one page long). The terms used for this research were 5-grams (Damashek 1995). An element a_{ij} of the term-document matrix A was the frequency of occurrence of the 5-gram i in the document j. Thus, the number of terms in each document was around 3000. The use of n-grams allowed us to use the same parser for each language, bypassing stemming and stop-listing. N-grams have also proven useful with other LSI-based techniques (Nicholas 1998).

The documents in the corpus were grouped together by language. In a parallel corpus consisting of n monolingual documents and their translations, document d_i has document $d_{i+n/2}$ as its corresponding translation. The number of dimensions in the LSI spaces constructed in the course of this research varied from 20 to 80.

3.5 Plotting and measuring

We used the standard plotting capability provided by MATLAB to plot the values in the Σ^*V^T matrix. The plots were all two-dimensional. The X-axis shows the sequence number of the document (and of its corresponding translation) in the collection, and the Y-axis shows the coordinate of that document in a particular dimension. Lines connect the document points in order to make the patterns easier to trace.

We also analyzed the correlation coefficients between all permutations of monolingual sub-vectors of each row of the Σ^*V^T matrix. This correlation analysis produced as a result a matrix of correlation coefficients of size [n, n]. Of this matrix only the highest values were chosen to illustrate the results of the experiment. These results are shown as in Table 1, below. The first and the second column hold the LSI dimension numbers of the sub-vectors participating in the correlation analysis. The third column holds the actual correlation values. The header row indicated what were the languages of sub-collections that produced the corresponding sub-vectors. The highest absolute values were selected for display. We also present graphs in which the two sub-vectors for each dimension are plotted for comparison, as shown in Figure 1, for example.

Table 1. Correlation coefficient table sample.

English sub-vector dimension	Russian sub-vector dimension	Correlation coefficient
1	2	0.8334
3	3	0.8902
3	4	0.8610
3	5	0.7996
7	7	0.7831
7	8	0.8342

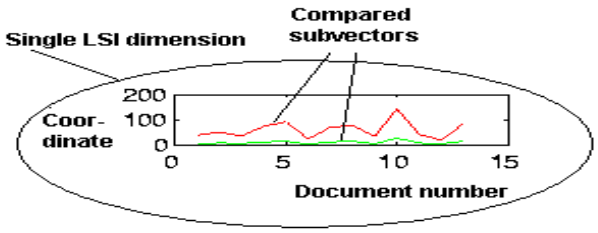


Fig. 1. Plotting schema.

The coordinate points for the documents are connected in the plots in order to make the pattern formed by them more evident. Each language participating in the experiment has its own color legend: red for English, green for French, blue for Russian, and black for Italian.

4 Results

This section is structured in the following way: we will first show the difference between the parallel and non-parallel corpora representation; we will then show, within parallel corpora, the examples of grammatically similar and dissimilar language combination. We will also talk about literal vs. non-literal translation. The majority of the research was done on mixed LSI spaces. However, the separate LSI space experiments produced interesting results that will be mentioned briefly in this section.

4.1 Mixed LSI Space

4.1.1 Parallel vs. non-parallel corpora

The results shown in this category are very important because they confirm the main hypothesis of this research – that the representation of a parallel corpus presents similarities between parallel parts of the corpus that are absent in a non-parallel corpus.

The documents in this collection come from “Smoke Bellew” by Jack London, namely the parts called “The Taste of the Meat” and “The Race for Number One”. In the case of a parallel corpus, the excerpts from both novels were combined into a single collection with their respective translations and the LSI space was created from them. In the case of a non-parallel corpus, the English part of the corpus was comprised of “The Race for Number One” excerpts and the Russian part from “The Taste of the Meat” excerpts.

The highest-value portion of the correlation coefficient matrix for the parallel corpus part of this experiment (see Table 2) shows that the values are well above the tentative threshold of 0.75 assumed in the course of the experiments. Also, the $S*V^T$ plot in Figure 2 shows some obvious similarities. For instance, the sub-vectors in dimensions 5 and 7 present opposite trends, whereas dimension 6 has similar patterns for both sub-vectors. The English sub-vector in dimension 4 is very similar to the Russian sub-vector in dimension 5. Thus, the results of this experiment support the hypothesis.

English sub-vector dimension	Russian sub-vector dimension	Correlation coefficient
3	3	0.8621
4	4	0.8793
4	5	0.9318
6	6	0.7890

Table 2. Correlation coefficients for parallel corpus from "Smoke Bellew".

The analysis of the parts of this corpus separately showed that “The Taste of the Meat” was less precisely translated into Russian than “The Race for Number One”. This might account for lower values for the former collection. Also, the patterns formed by each sub-corpus in this experiment remain distinct when the collections are combined.

The first 15 dimensions for this experiment involving the non-parallel collection from the same texts did not produce nearly as high similarity estimates as the previous experiment (see Table 3). However, there is still some similarity of behavior visible in dimensions 1 and 2, 3 and 4, and 7 and 8 of the $\Sigma*V^T$

matrix plot (see Figure 3). The reason may be that the novels were written by the same author, and on a similar topic. However, in this experiment, the similarities in the visual representation of the corpus are accompanied by low values of correlation coefficients.

English sub-vector dimension	Russian sub-vector dimension	Correlation coefficient
1	3	0.6368
7	17	0.6168
7	18	0.6173
1	2	0.5439
4	3	0.3629
8	7	0.2568

Table 3. Correlation coefficients for non-parallel corpus from "Smoke Bellew".

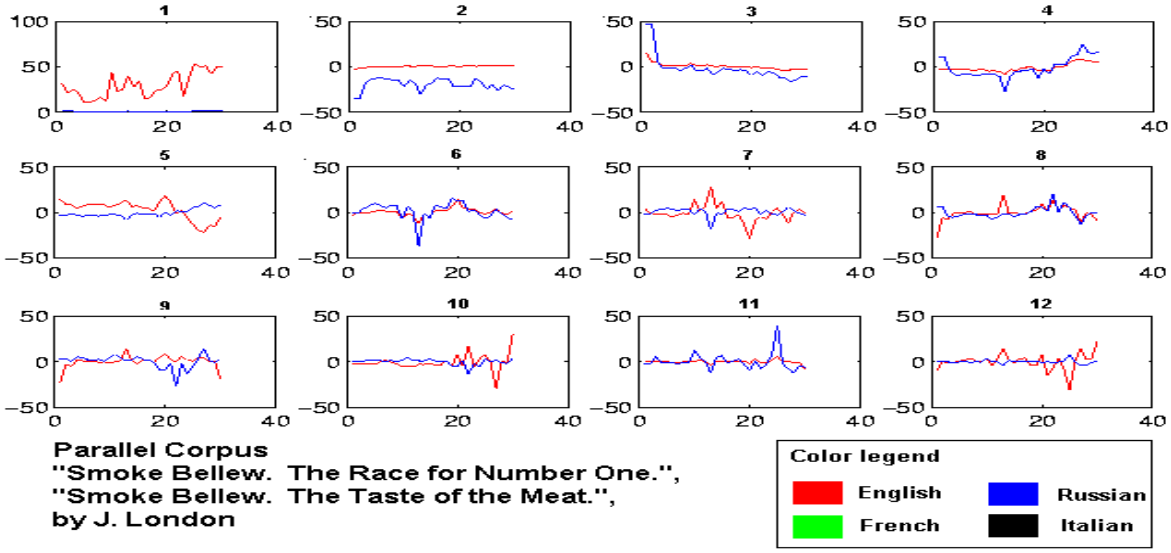


Figure 2. Parallel corpus example, English and Russian.

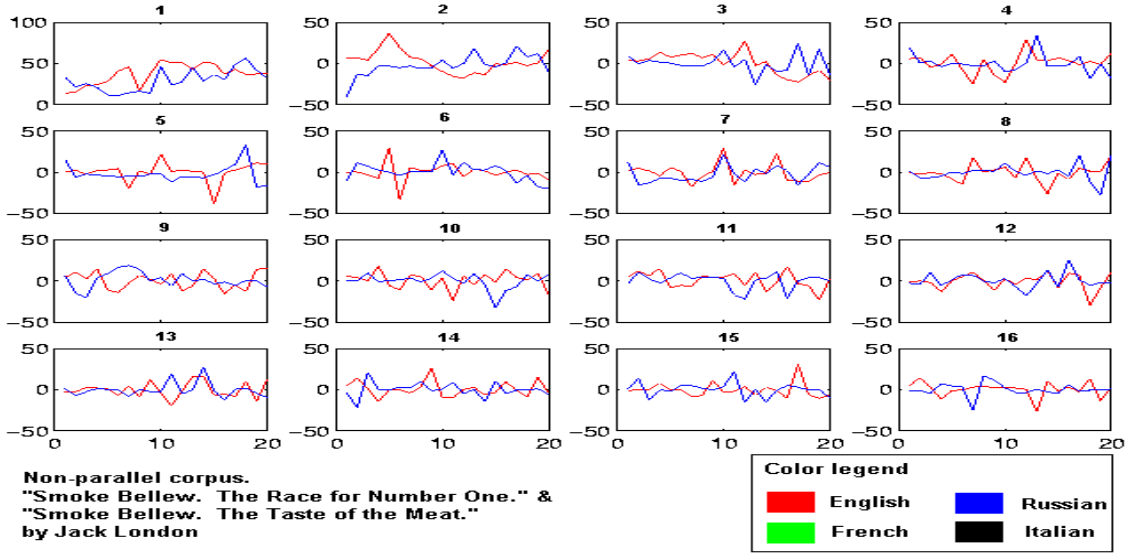


Figure 3. Non-parallel corpus example, English and Russian.

We can see that the highest correlation coefficient for the non-parallel corpus is 0.64 and the lowest correlation coefficient for the parallel corpus shown above is 0.78. We therefore set a tentative threshold of 0.75 for the correlation coefficients in judging a corpus as parallel. The data for the non-parallel corpora that we have tested with so far has proven consistent in not providing correlation coefficient measures above 0.70 (usually it was much lower).

4.1.2 Similar vs. dissimilar languages

The texts examined in this category are the French original of the "The Little Prince" by Antoine de Saint-Exupery with its Russian and English translations.

The analysis procedure for this category is similar to the one demonstrated in Section 4.1.1. In this series of experiments, the range of values in grammatically similar and grammatically dissimilar languages are not significantly different. The highest correlation coefficient for the English/French pair is 0.99, and 0.98 for the French/Russian pair. In other words, both grammatically more and less similar languages participating in parallel corpora produce high correlation values, and don't produce them in a non-parallel corpus. However, depending on the style of the translation, the difference between these values may be more significant, although the experiments done so far show that these values still fall within the 0.75-1.0 range.

We made the following observations about the Σ^*V^T matrix. In the French/English collection of excerpts from "The Little Prince" we saw that starting with dimension 1, the similar patterns are most often observed within the same dimension, e.g., dimension 1, dimension 3, etc. In dimension 4 we saw an effect of opposing patterns, where the English pattern is the same as in the previous dimension, but the French is the opposite. Thus, the idea of "mate" dimensions converging towards the same dimension seems to be supported.

The French/Russian representation also presents strong similarity trends, but here the "mate" dimensions are more dispersed, e.g. similar patterns occur in dimension 1 and 2, 3 and 4, 5 and 6, 8 and 12, etc. This also seemed to confirm the "mate" dimensions suggestion. The highest correlation coefficients for this series of experiments were between 0.99 and 0.88.

4.1.3 Literal vs. literary translation

A literary translation is intended for transmitting both the meaning and the spirit of the original work. This means that, considering that different languages have different imagery mechanisms, code words, and idioms, it is almost impossible to translate a literary text literally into another language and obtain a translation that's worth reading. Thus, the only claim that can be made towards the usability of LSI in application to literary translation is that it measures not so much the literality of translation, as its consistency; the style of the original and of the translation needs to have similar "texture". For instance, if the author of the original work uses a single imagery for describing the blue sky and the translator uses three different images for the same purpose, the consistency of the translation will suffer.

The examples that we worked with here are the excerpts from "Smoke Bellew", specifically, parts of "The Race for Number One" and "The Taste of the Meat". In this pair, "The Taste of the Meat" is the less literal translation. The correlation coefficient values for this experiment are not very high (the highest is 0.86), nor the dimensions in which they are high, very multiple. However, these values are located within the assumed threshold and the dimensions that possess these high similarity values are within the first 10-15 dimensions.

The analysis of the second element in the selected pair, "The Race for Number One" produced the following much higher correlation coefficient values (the highest 0.9) and there is a significant number of dimension pairs with similarity value above 0.75.

The separate LSI space approach presented similar results to those described for the mixed LSI space, which seems to support our hypothesis. For additional information on these and our other experiments, we refer the reader to the first author's thesis (Katsnelson 2000).

5 Conclusion

The application of the LSI method and the subsequent correlation analysis showed that parallel corpora produce larger correlation coefficient values for rows of the S^*V^T matrix of the term-document matrix. The row plots of this matrix are also visibly more similar between the parts of a parallel corpus than between the parts of a non-parallel corpus.

We analyzed several aspects of corpora comparison, such as parallel vs. non-parallel collections, literal vs. non-literal translations, and document collection pairs in languages with different degrees of grammatical similarity. The proposed threshold of 0.75 held for all types of parallel corpora. The

experiments also showed that visual analysis is as important for the application of our method to parallel corpora identification, as the statistical analysis.

The limitation of our experiments has been that the number of dimensions that we have worked with was low due to the computational expense of SVD, and the limited number of documents. In the future we are planning to address this issue by both increasing dimensionality of experimental LSI space and considering alternative methods of dimensionality reduction that are less computationally demanding. We also hope to extend this method into evaluating the quality of translation.

References

T. K. Landauer, M. L. Littman, "Computer information retrieval using latent semantic structure". U. S. Patent No. 4,839,853, Jun 13, 1989.

S. Deerwester, S. T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis. Journal of the Society for Information Science, 41(6), 391-407, 1990.

R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval", ACM Press, New York, 1999.

M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using Linear Algebra for Intelligent Information Retrieval. SIAM Review, 37(4), 1995, pp. 573-595, 1995.

M. L. Littman, S. T. Dumais, and T. K. Landauer. Automatic Cross-Language Retrieval Using Latent Semantic Indexing. SIGIR96 Workshop on Cross-Linguistic Information Retrieval, 1996.

S. T. Dumais, T. A. Letsche, M. L. Littman, and T. K. Landauer. "Automatic cross-language retrieval using Latent Semantic Indexing", AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, March 1997.

M. Damashek, "Gauging similarity with n-Grams: Language-Independent Categorization of Text", Science, February 1995, Vol. 267, pp. 843-848.

C. Nicholas and R. Dahlberg. Spotting Topics with the Singular Value Decomposition. PODDP'98, St. Malo, pp. 82-91,March 1998.

Y. Katsnelson, "Parallel Corpora Identification Using LSI", M.S. thesis, UMBC Department of Computer Science and Electrical Engineering, January 2000, http://www.cs.umbc.edu/~ykatsn1.