Bartlett's Test Applied in a Latent Semantic Analysis of Parallel-Aligned Sentences in French and English

Katri A. Clodfelder
Indiana University
kclodfel@indiana.edu

Abstract

Bartlett's test, a simple test of statistical independence is used to inform a Latent Semantic Analysis of parallel-aligned sentences in French and English.

1 Credits

The texts used in this analysis were originally prepared for the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts and were taken from daily House journals of the Canadian Parliament. They were edited by Ulrich Germann. The LSA procedures were implemented in R, a system for statistical computation and graphics. Non-native R procedures were written by John C. Paolillo and Katri A. Clodfelder at Indiana University, Bloomington.

2 Introduction

Latent methods, and in particular Latent Semantic Analysis (LSA), have enjoyed a large popularity in recent years in numerous fields of research (Paolillo, 2004; Landauer, Foltz, and Laham, 1998; Landauer and Dumais, 1997). As used in Information Retrieval, the LSA model typically focuses on correspondences between texts of at least a certain size, whose syntactic information has not been retained. The loss of syntactic information includes word order, and very often, high-frequency, semantically-light terms and other lexical items which often serve as syntactic markers of meaning such as conjunctions, prepositions, modals, and pronominals. The latter are often considered to act as "noise" in the LSA model since they seemingly contribute little semantic information to the lower-frequency, semantically-heavy content terms that are typically the objects of Information Retrieval queries. The success of the LSA model extends to cross-linguistic Information Retrieval where it has been shown to correctly correlate relevant documents in response to natural language queries in as many as three languages (Rehder, Littman, Dumais, and Landauer, 1997).

The model has also been shown to correlate among linguistically-relevant data in at least two ways: syntagmatically when word order is accounted for in the model (Paolillo, 2004); and phonetically in the case of tongue shape representations relative to vowel pronunciation (Harshman, Ladefoged, and Goldstein, 1977).[1]

The above successes make the LSA model, and other latent methods such as principal components and factor analysis, appealing models for exploring cross-linguistic relations. While the model clearly correlates among contextually-related documents in a collection of multilingual texts (Rehder, Littman, Dumais, and Landauer, 1997), in this paper we are concerned with its potential as a statistical model of language (Paolillo 2004). As such, we begin to explore the model from that perspective using a small collection of 500 parallel-aligned sentences in French and English.

Our starting point deviates from the usual application of the LSA model in several ways. We use very small texts of sentence length only, retaining all lexical items; however, we do not explicitly represent word order as does Paolillo.[2] Sentence-length data is used because it permits us to better observe the influences of syntagmatically-associated terms, monolingually, and how they correlate, cross-lingually, when input as a collection of duolingual, collocated terms.

We do not perform the linear transformation known as TFIDF[3]. This decision is based on two

[1] Note that the Tongue Factor Analysis performed by Harshman, Ladefoged, and Goldstein (1977) is, strictly speaking, not the LSA model since it is not performed on textual data. However, inasmuch as the methods of LSA and Factor Analysis are born of the same mathematical precepts, we include it here.

[2] Note that Paolillo's (2004) analysis is limited to bigramically-represented data in the input.

[3] The Term-Frequency/Inverse Document Frequency (TFIDF) ratio is a linear transformation presumed to better represent the "informative" relations among the terms and

factors: our use of sentence-length texts; and our imposed perception of LSA as a model of language. In the standard application of LSA as an Information Retrieval model, text size is substantially larger than a single sentence and frequently-occurring, semantically-light terms assume less importance relative to less frequently-occurring, semantically-rich terms. This reality is captured by weighting each term frequency (TF) with the number of documents in which it occurs (IDF). Alternatively, high-frequency terms often provide informative linguistic value since they frequently participate in structural regularities (e.g., heads of phrases) or syntactic constructions (e.g., compound tenses) which are not important in document retrieval. Although we do not model these structural regularities or constructions in this analysis, we are interested in observing the influence of these terms in the model.

We perform the singular value decomposition procedure against the term-correlation matrix rather than the term-document matrix. This decision is based on the singular value decomposition theorem. When the latent roots and vectors of the correlation matrix are known, the original matrix can be decomposed (or expanded) in these terms (Basilevsky, 1994).[4]

If LSA is perceived as a statistical model of language,[5] then standard statistical methods can be utilized to confirm the significance of the computed components and their roots. To the degree that components may be identified with linguistic phenomena,[6] statistical testing may be useful in identifying those components that are most crucial to modelling the underlying relationships.

In this analysis, Bartlett's test, a simple test of statistical independence, is used to confirm initial dependency of the data and to inform decision-making with respect to component retention. Presented in Section 3, Bartlett's test permits the discovery of those components which represent a systematic variance structure.

Finally, in this analysis, the retained components undergo a secondary transformation known as oblique rotations. These methods are discussed in the following pages and are followed by a brief discussion of several components and some concluding remarks.

3 Bartlett's Test of Independence

The primary purpose of latent methods, including LSA,[7] is to re-organize underlying linear and non-linear, dependent associations among the data into a more optimal, linearly-independent, vector space representation (Basilevsky, 1994). The number of dimensions necessary to represent the initial data is reduced such that the original variables can be projected onto this smaller dimensional space, while losing only a minimum of information.[8] This smaller dimensional space is presumed to be some optimum number of components, or factors, necessary to explain the variance found in the original sample space.

Two questions that arise immediately concern the actual undertaking of the analysis to begin with and determining the number of components to retain once the decomposition has been performed. The first question has to do with the linear independence of the original variables. The second question has to do with selecting the number of components that may be discarded from the analysis without losing the ability to recreate the original data matrix. The latter question is not easily answered and is often determined without regard to statistical significance testing (Paolillo, 2004; Landauer and Dumais, 1997).

3.1 Linear Independence of the Variables

In LSA, the language data are initially represented on a term-document matrix where each cell entry contains the number of times a given term occurs in a given document. The interrelationships which exist among the terms in a given collection of sentences translate as non-linear data dependencies in an initial, high-dimensional vector space, demonstrating a high degree of correlation. Since the objective of

documents than the actual co-occurrence of terms as measured by the raw data (Landuaer and Dumais, 1997).

[4] Statistically speaking, the latent roots and vectors computed for the sample data may be conceived of as estimators of the population variables.

[5] As Paolillo (2004) observes, there are several reasons for why LSA fails at being a statistical model, including that the terms, which we desire to view as random variables, are not true random variables. They are predetermined by the choice of texts we choose to include in the analysis.

[6] Since components are only mathematical constructs, we must determine whether they can be identified with real-world phenomena (Basilevsky, 1994).

[7] Note that since LSA is well documented (see Landauer et al, 1998; Paolillo, 2004; Rehder et al, 1998), we dispense with an overview of the model.

[8] Theorem 3.9(iii), Optimality properties of the Principle Components model - PC's maximize the information content of the variables (Basilevsky 1994).

latent methods is to re-organize these underlying non-linear, dependent associations into a more optimal, linearly-independent, vector space representation, the starting point for LSA is to ensure that the variables in the initial vector space are not already linearly independent. If so, there is little point in continuing with the LSA procedures−it would be meaningless (Basilevsky, 1994).

Bartlett's test of statistical independence can be applied to the correlation matrix of an $n\ x\ p$ data matrix to test whether the distribution of the variables is multi-variate normal. When the random variables are linearly independent, the corelation matrix is an identity matrix I, whose leading diagonal elements contain only ones and whose off-diagonal elements contain only zeros; that is, all elements $a_{ij} = 1$, for $i = j$, and zero otherwise (Turnbull, 1961).

Equation (1) (where $|\mathbf{R}|$ is the determinant of the sample correlation matrix, p is equal to the number of sentences, and n is equal to the number of terms) is distributed as a chi-square distribution with $[(p/2)*(p-1)]$ degrees of freedom and can be used to test the null hypothesis that the population correlation matrix is an identity matrix H_0: $P=I$ (H_a: $P \neq I$) (Basilevsky, 1994). For large values of the X^2 statistic, we reject the null hypothesis that the population correlation matrix from which our 500 sentences were drawn is an identity matrix.

$$-[n-(2p+5)/6] \ln |\mathbf{R}| \qquad (1)$$

We applied the X^2 statistic given by Bartlett's test shown in equation (1) to the sample correlation matrix. Since the calculated value of the sample X^2 statistic approaches infinity, we reject the null hypothesis that the sample correlation matrix could be obtained from a population of random variables exhibiting linear independence and proceed with LSA.

3.2 Root Equality

The heart of LSA is the Singular Value Decomposition (SVD) theorem. By it, and various other theorems related to transformations that reduce matrices to a simpler and more convenient shape, many desirable properties of matrices can be obtained, proved, or evaluated (Basilevsky, 1994; Turnbull, 1961). LSA exploits the SVD procedure in order to obtain latent vectors and roots of the term-document matrix representing the sample data. Each non-zero root provides a measure of the variance explained by the corresponding principal component. When some, or all, of the non-zero roots are equal, the components corresponding to those roots are said to be equally correlated.

Bartlett's test of independence given in equation (1) can be expressed in terms of the latent roots by replacing the ln $|\mathbf{R}|$ with the natural log of the product of the latent roots (ln $\prod_{i=1}^{p} \ell_i$). The modified equation given in (2) tests for equality of the latent roots of the correlation matrix, when the roots are already known. If all p roots are equal, H_0: $\lambda_1 = \lambda_2 = \cdots \lambda_p$ (H_a: all λ not equal), then the variation explained by the components is isotropic and can be explained by one component that is equally correlated to all p random variables (Basilevsky, 1994). The test was conducted and, again, as suspected based on the chi-square results of equation (1), the chi-square value approaches infinity, indicating that the latent roots are not equal and thus, at least two of the principal components differ significantly.

$$-[n-(2p+5)/6] \ln \prod_{i=1}^{p} \ell_i \qquad (2)$$

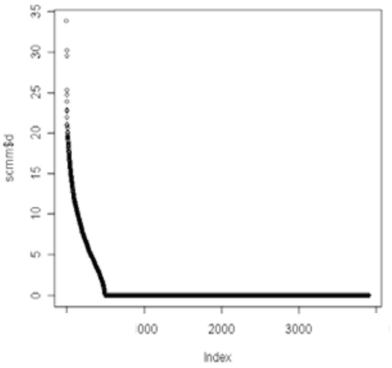


Figure 1: Latent roots of correlation matrix

The test of independence using equation (2) does not provide any information regarding which two components differ significantly or whether any subset of the components are equally correlated. The latter may occur when the last q roots are equal. Viewing the latent roots graphically (see Figure 1), it is apparent that the last q roots are very small. Several questions arise. Do the last q roots vary by sufficiently small amounts that they might actually be closer to being equal than not? Is the variance represented by them sufficiently unique qualitatively that they contribute virtually nothing to the overall systematic variance of the data set (captured in the first r components)? Is the total variance represented by them sufficiently small quantitatively that an adequate representation of the original data matrix could be reproduced without them? If a subset of q equally-correlated components exists, is the value of q large enough to substantially reduce the number r of remaining components so that interpretation of same is facilitated?

The first two of these questions are informed by yet another modification to Bartlett's test of independence. The third and fourth questions are often informed by value judgements of the analyst and in applications where interpretation of the components is not motivated, the number of r retained components is often heuristically-based (Landauer and Dumais, 1997). Bartlett's test provides a statistical basis for removing components from the analysis, thereby removing some of the subjectivity of component retention decisions. Equation (3) gives the computation of the statistical test for equality of the last q latent roots. In equation (3), q equals the number of latent roots tested for equality, $l\text{-}Bar_q = (1/q)\Sigma_{i=r+1}^{p} l_i$, the arithmetic mean of the last q latent roots, and r equals the number of retained components, with $(((q/2)*(q+1)) - 1)$ degrees of freedom. For large values of X^2, the null hypothesis H_0: $\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_p$ is rejected (Basilevsky 1994).

$$\begin{aligned} X^2 = & -[\, n - r - (1/6q)(2q^2 + q + 2) \\ & + \Sigma_{i=1}^{r} ((l\text{-}Bar_q)^2 / (l_i - l\text{-}Bar_q)^2) \,] \\ & \times [\, \Sigma_{i=r+1}^{p} \ln l_i - q \ln (1/q\ \Sigma_{i=r+1}^{p} l_i) \,] \qquad (3) \end{aligned}$$

The results of Equation (3), given in Table 1, indicate that the last $q=41$ latent roots are equally correlated at a significance level of 99.96% (we accept the null hypothesis). The last $q=41$ roots represent residual variation that is isotropic (unique) and which does not contribute to the systematic variance represented in the first r components. However, Basilevsky (1994) notes that even when the null hypothesis is rejected, it does not necessarily mean that the variance structure of the residual components is systematic, only that it is systematic at some significance level.

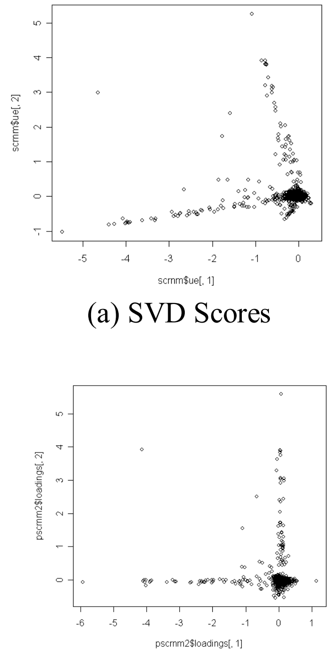
X^2:	1004
No. Roots Tested:	41
Degrees of Freedom:	860
Significance Level:	99.96%

Table 1: Bartlett's X^2 Statistic for Testing Equality of last $q=41$ Latent Roots

Based on the above test and that the last $q=41$ components account for only slightly greater than one percent of the total variance, the last 41 components are removed from the analysis. The remaining 456 components are retained and undergo a secondary transformation discussed in Section 4.

Other statistical tests are available that can be performed on the roots, vectors, and vector elements; however, they are more complicated to perform and the scope of this paper does not include them. Bartlett's test of independence is relatively simple, can be quickly computed, and provides an objective, statistical measure for removing at least q components from the analysis.

4 Secondary Transformations of the Components



(a) SVD Scores

(b) Rotated Scores

Figure 2: Scores for Components 1 and 2

Retaining $r=456$ components, we consider the estimated loadings (scores) of the terms. With such a large number of components, matchingthem to linguistic phenomena is difficult. Secondary transformations can be of assistance in interpreting the components because they attempt to polarize the score values for each component (Basilevsky 1994, Abdi 2003). While the rotation procedure does not reduce the number of components, component interpretation is facilitated because the total variance of the r rotated components is re-distributed in such a way that each component will have only a few very large values but many very small values. By observing the terms associated with the very large values, we can begin to make some inferential interpretations of what the components most likely represent.

The scatterplots of Figure 2 depict the SVD and rotation scores of Components 1 and 2 graphically, showing that the overall effect of the rotations is to polarize the scores so that similar and near-similar scores can more easily cluster together.

5 Results and Interpretation of the Components

5.1 The Language of the Input Data

The 500 sentences used in this analysis were culled from transcripts of Canadian parliamentary proceedings on five separate days. The language of the texts covers a broad range of topics including politics, the business of the state, the governance of the territories and provinces, the implementation of social and health-related programs, legislative proposals, the infrastructure of the country, national resources both public and private, all industries including manufacturing, agriculture, oil, communications, fishing, banking, and other commercial enterprises, military operations, and international relations. Nor is this list exhaustive since it is hardly possible to enumerate all the potential topics that may be discussed in Parliament on any given day.

We must also note that the speakers who actually address Parliament can be fairly diverse. It is hardly reasonable to expect consistency of style or manner of speaking across all speakers. Some may speak eloquently and with a more elevated speech pattern while others may speak rather plainly and in much shorter sentences. On the other hand, while not absolute, we should expect a certain level of discourse since it is more likely that individuals who address Parliament have attained some level of education and world experience that accords them this somewhat rare privilege (rare to the majority of the population whose daily affairs do not concern matters of the state).

5.2 Interpreting the Components

Due to space limitations, only the terms from four of the components are provided in Tables 3 and 4. On each component, the first grouping contains the most highly correlated terms while the second group contains terms whose scores cover a range of values. The terms in this group are arranged from more correlated to less correlated. Before discussing the individual components, some general observations are made.

First, the majority of high-frequency, semantically-light lexical items do not generally correlate to a specific component as their scores typically fall somewhere near zero on most components. Second, while word order is expressly omitted from the input data, terms that generally co-occur as bigrams or trigrams appear to correlate similarly or near-similarly on a given component (e.g., *red tape*, *sexually transmitted diseases*). Third, because we used an oblique rotation, complete orthogonality (independence) between the components is not necessarily retained. Some terms may, therefore, correlate to more than one component.

ACCABLE	AFFRANCHISSEMENT
ANTICIPATION_X	ADMINISTRATIVES
DEPUIS	DETAIL_X
DÉCOULER	DÉTAIL
EMPÊCHAIT	ENTHOUSIASME
ENVISAGER	EXPRIMER
FREEDOM_X	INQUIÉTUDE
MANQUERAIENT	POSSIBILITÉS
RED	PRIVAIT
TAPE_X	TRACASSERIES
WRAPPED_X	

Table 2: Most strongly correlated terms on Component 2 - Score = 5.597

Finally, the negative and positive correlations of the terms are purely a function of the mathematics of the procedure. We cannot make any inferences with respect to the negative or positive connotations that humans may associate with these topics until the clusters are examined. For example, the terms on Component 2 correlate positively (see Table 4) and yet, we note that included among these terms is the bigram *red tape (tracasseries administratives)*. The inclusion of *red tape* in this cluster of terms suggests that, semantically, the component relates to some type of bureaucratic process, which is rarely viewed favorably.

5.2.1 Component 1

Semantically, Component 1 appears to correlate limited manufacturing opportunities in the south with transportation difficulties due to a lack of infrastructure development. This determination is made based on the syntagmatically possible nominal compounds which correlate on the component. For example, the English *n*-gram possibilities include *relatively limited manufacturing opportunities* and *manufacturing opportunities (are) relatively limited* both of which could potentially correspond to the French *n*-gram *occasions (de) manufacturier relativement limitées*.

Component 1	***Component 4***
-5.93	***-5.015***
ACQUIS	ABORTION_X
DIFFICULTIES_X	ADOLESCENTS
DUE_X	ADOLESCENTS_X
EXAMPLE_X	AVORTEMENTS
EXEMPLE	BEHAVIOUR_X
EXTRÊMEMENT	DANGEREUX
INFRASTRUCTURE	DISEASES_X
KINDS_X	DÉSIRÉES
LIMITED_X	GROSSESSE
LIMITÉES	MALADIES
MANUFACTURIER	PREGNANCY_X
MANUFACTURING_X	PROMOTE_X
NATURE	PROMOUVOIR
OBTAIN_X	SEXUALLY_X
OCCASIONS	SEXUAL_X
RELATIVELY_X	SEXUEL
RELATIVEMENT	SEXUELLEMENT
SOUTHERN_X	TRANSMISES
SUD	TRANSMITTED_X
TRANSPORT	UNSAFE_X
TRANSPORTATION_X	UNWANTED_X
-4.017 to -2.222	***-3.486 to -0.6724***
OPPORTUNITIES_X	VISENT
EXTREMELY_X	HOPED_X
DÉPENDANT	MEASURES_X
GRANTED_X	COMPORTEMENT
TIRER	RESPONSABLE
NATURE_X	PROTÉGER
REVENUS	PROTECT_X
ESPECIALLY_X	RESPONSIBLE_X
INFRASTRUCTURE_X	CONTRE
REPRÉSENTE	NON
DEPENDENT_X	MESURES
PROBLÈMES	THESE_X
MANQUE	BY_X
LACK_X	FROM_X
RAISON	CES
GENRE	
INCOME_X	
SURTOUT	
APPUYER	
SECTEUR	
EXPLOITATION	
REVENU	

Table 3: Rotated Scores of Components 1 and 4

Also of interest on Component 1 are the cross-linguistic correspondences that correlate together. For example, adverbials correspond to adverbials, nominals to nominals, and plurals to plurals. However, not all of what we would consider to be rather literal correspondences correlate exactly, nor do all terms have a cross-linguistic correpondent. English *infrastructure* and *income* do not correlate precisely to French *infrastructure* and *revenu*. English *due*, which often collocates with *difficulties*[9], appears to have no correlating correspondent at all. What this suggests is that the model correlates among equivalent semantic constituents cross-linguistically, even when they are expressed quite differently in the two languages. This is seen more clearly in interpreting Component 4.

5.2.2 Component 4

Component 4 correlates to adolescent sexual behavior and the negative results of that behavior. Again we see corresponding nominal compounds cross-linguistically (3) but we also see some evidence that this component correlates intra-sentential nominal-verbal constitutents cross-linguistically (4).

(3) sexually transmitted diseases
maladies sexuellement transmises

unsafe sexual behavior (of) adolescents
comportement sexuel dangereux (des) adolescents

In (4), English *hoped* and French *visent* correlate almost precisely the same and yet, it is not likely that one would consider *hope* to be an English equivalent of French *viser* (to aim). What (4) suggests is that LSA correlates cross-linguistic, semantically-equivalent, non-literal, language-specific syntagmatic associations.

(4) (it is) hoped (that) these measures (will) promote responsible sexual behavior

ces mesures visent (à) promouvoir (un) comportement sexuel responsable

[9] In English, *due* often collocates with *difficulties* in the expressions *difficulties due to..., due to difficulties with..., difficulties with (whatever) due to...* and so on.

5.2.3 Component 6

Component 6	***Component 8***
4.96	***-4.832***
ANTICIPATED_X	ALIKE_X
AURIONS	ATTIRER
CLOSER_X	AWARENESS_X
ENVOYÉES	DÉSIGNANT
FAIRNESS_X	DISEASE_X
FRANCHISE	ÉTENDUE
IMAGINE	FEMMES
INOCULATED_X	GRAVITÉ
INOCULATION_X	HOMMES
PRÉVOIR	INFORM_X
REFERRING_X	INITIATIVE_X
SUPPOSE_X	MAGNITUDE_X
TROOPS_X	MALADIE
TROUPES	OCTOBRE
VACCINER	SENSIBILISATION
	SEVERITY_X
3.458 to 1.626	***-3.486 to -0.3306***
SUBJECT_X	VOULAIT
LITTLE_X	MEN_X
ÉTRANGER	MONTH_X
ABORD	SOCIÉTÉ
ASSEZ	SOCIETY_X
SENT_X	BREAST_X
CONVAINCUS	CANCER
DÛ	SEIN
SHE_X	CANCER_X
CONCLUSION_X	ATTENTION
INTERVENTION	MOIS
PEACEKEEPING_X	WOMEN_X
MAINTIEN	CANADIENNE
NOS	TOUS
COME_X	AN_X
SHOULD_X	CANADIANS_X
SERONT	CANADIENS
INDEED_X	CANADIAN_X
PAIX	CETTE
BEFORE_X	LA
	OF_X
	SUR

Table 4: Rotated Scores of Components 6 and 8

Semantically, Component 6 (see Table 4) correlates to the deployment of troops in a foreign country (not indicated) whose purpose is "keeping the peace." The clusterings suggest that the main issue is vaccination of the troops, not their peacekeeping mission. On this component, we see some evidence that the model correlates among verbal compounds (5).

(5) (we) should (have) anticipated
(nous) aurions dû prévoir

(we) should (have) innoculated
(nous) aurions dû vacciner

5.2.4 Component 8

Component 8 can only correlate to what is called "Breast Cancer Awareness Month." Social programs related to public health, whether breast cancer or sexually transmitted diseases, are often funded by federal legislation and it is no surprise to find that related terms correlate to specific components. On Component 8, syntagmatic associations are confounded between the two languages. For example, the program is intended to sensitize *all Canadians, men (and) women alike* to the seriousness of breast cancer. The most likely French correspondent to the italicized phrase is *tous (les) canadiens, hommes (et) femmes* (all Canadians, men and women) which contains no correspondent to *alike*. That English *alike* correlates more strongly to the cluster containing French *hommes* and *femmes* rather than to English *men* and *women* suggests that the model correctly correlates the usage of a term in one language to appropriate correspondences in the second language.

6 Concluding Remarks

The previous discussions should make it clear that even in a relatively small experiment such as the one conducted here, the number of components can be fairly large, making the task of matching them to real-world phenomena extremely difficult. And, as shown above, the removal of a residual variance structure may not substantially reduce the total number of components. Since, presumably, the systematic variance structure is representative of the interrelationships existing among the data, we must begin to understand what the components actually represent. Do some components correlate to the semantic content of the collection of texts and others to more syntactically-driven relationships?

Shown in Table 5 are the most strongly correlated terms on Component 12. Note that three of the French terms are verbs conjugated in first person plural (*-ons*). The first person plural pronouns *nous* and *we* also correlate to Component 12, although less strongly with scores of 1.694 and 1.116, respectively. Is it a coincidence that the individual scores of these terms are also highest on the same component whose highest scores correlate to verbs conjugated in the first person?

INTERROGEONS	AGIRAIT
RÉJOUISSONS	ASSUMED_X
SUPPOSIONS	INSISTANCE
INSISTENCE_X	WONDERING_X
PRÉSENT	SOUHAITÉ

Table 5: Most strongly correlated terms on Component 12 (*Score=4.676*)

Understanding and interpreting the components as was done here is dependent on cognitive and reasoning capabilities unavailable to the model, as well as our own personal knowledge of permissible constructions in each language. For this reason, the capability of the model to confirm cross-linguistic correspondences among terms that correlate precisely the same appears limited. For example, given a task to evaluate whether the phrase *sexually transmitted diseases* corresponds cross-linguistically to *maladies sexuellement transmises*, the best the model can do is to confirm that all the terms of each phrase correlate in the same cluster, on the same component.

For the same reason that the model cannot confirm cross-linguistic correspondents among terms that correlate precisely the same, it cannot reject non-corresponding, cross-linguistic pairings. Given the same evaluative task as before, and given the phrases *comportement sexuel dangereux* and *sexually transmitted diseases*, the model cannot deny that these two phrases are cross-linguistic correspondents (as it should do).

The above represents a preliminary excursion into the methods and practices of latent methods and the view of LSA as a statistical model of language rather than an Information Retrieval model. Many procedures for testing significance of components and roots have not yet been tried. Additionally, we note that the rotation procedures used in this analysis may contribute to the problem since it forces terms with small differences in individual scores to correlate as if those differences did not exist. It is possible that unrotated scores would improve the model's capability to distinguish among corresponding terms cross-linguistically.

As noted previously, our initial data representations do not include word order information. Paolillo (2004) illustrates that when such information is explicitly represented, latent methods appear to correlate syntagmatically-relevant data. Our own work in this analysis appears to confirm Paolillo's observation that although the LSA model is sensitive to both syntactic and semantic influences, it cannot distinguish between them.

Clearly, in going forward, there is much to consider. We believe that for purposes of linguistic- and language-related analyses, latent methods are promising and hope that the discussions presented here provide some insights into the difficulties and complexities of using and understanding the LSA methodology..

References

Abdi, Hervé, 2003. "Factor Rotations in Factor Analyses" In *Encyclopedia of Social Sciences Research Methods,* Eds. Lewis-Beck, M., Bryman, A., Futing T. Thousand Oaks, CA: Sage

Basilevsky, Alexander, 1994. *Statistical Factor Analysis and Related Methods: Theory and Applications*. John Wiley and Sons, Inc. New York.

Hanushek, Eric A.and John E. Jackson., 1977. *Statistical Methods for Social Scientists*. Academic Press, Inc., London.

Harshman, Richard, Ladefoged, Peter and Louis Goldstein, 1977. "Factor analysis of tongue shapes" In *Journal of Accoustical Society of America*, Vol 62, No. 3.

Landauer, T. K. and Dumais, S. T., 1997. "A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge." *Psychological Review,* 104, 211-240.

Landauer, T. K., Foltz, P. W., and Laham, D., 1998. "An Introduction to Latent Semantic Analysis." *Discourse Processses*, 25, 259-284.

Paolillo, John C., 2004. "Latent Structure Analysis: Semantic or Syntactic?" International Conference on Natural Language Processing. Hyderabad, India.

Rehder, Bob, Littman, Michael, Dumais, Susan and Thomas K. Landauer, 1998. "Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing." *Proceedings of TREC-6.*

Turnbull, H. W. and A. C. Aitken, 1961. *An Introduction to the Theory of Canonical Matrices*. Dover Publications, Inc., New York