A Bilingual Analysis of Cohesion in a Corpus of Leader Speeches

Lubna Shala, Vasile Rus, and Arthur C. Graesser[1]
Department of Computer Science
[1]Department of Psychology
The University of Memphis
Memphis, TN 38138
vrus@memphis.edu

Abstract

We study in this paper the cohesion of a leader's speeches over time. This is part of a larger project that aims at investigating the language of leaders and how their language changes over their stay in power. Here, we analyze the speeches of a leader who stayed in power for a long period of time, i.e. more than 30 years. We measure cohesion of speeches in the original language, which is Arabic in our case, as well as in English, based on human translations of the original speeches. The cohesion is measured in two different ways: using word overlap and Latent Semantic Analysis. Because of the morphological complexity of Arabic, the word overlap measure of cohesion becomes challenging in Arabic. Latent Semantic Analysis is totally unsupervised is applied similarly for Arabic and English. The results show that cohesion has a general down trend over time and that during and after major crises the leader's speeches exhibit an increase in cohesion which can be explained as an attempt on leader's behalf to make his policies more clear, most likely as a form of post-crisis management.

Introduction

In this paper, we present a linguistic analysis of the speeches of Egyptian leader Hosni Mubarak who stayed in power for more than 30 years from 1981 to 2012. We link the linguistic analysis to other personal and contextual factors in order to understand whether these factors are reflected in his speeches. The work presented in this paper is part of a larger project that aims at studying leaders' language.

We present here our work on analyzing how cohesion of leaders' speeches change over time. We measured cohesion of speeches in two ways: word overlap and Latent Semantic Analysis (LSA; Landauer et al., 2007). The analysis was conducted both in the original Arabic language and in English. The LSA method is directly applicable to both Arabic and English because it is totally unsupervised. Large collections of texts is all that is needed as input to LSA. We derived LSA spaces for Arabic using two different collections of texts: a collection of 72 MB of Arabic texts collected from online Arabic sources that includes books (history, novels, philosophy, politics, and studies) and news articles (economy, entertainment, health, politics, sports, and others) and a collection using Arabic Wikipedia. Due to major morphological differences between English and Arabic there are challenges when it comes to computing cohesion using word overlap. This problem is exacerbated by the fact that Arabic is a resource poor language, for the time being, when it comes to automated language processing tools.

Experiments were conducted on a corpus of 902 speeches in Arabic spanning Mubarak's tenure from 1981 to 2011 and 306 English translations of the speeches from the 1996-2011 period (English translations were only available for this period). Results show that cohesion goes down over time and that cohesion goes up after major events such as opposition challenges, wars, or economic crises. Furthermore, the cohesion of the speeches has an upward tendency in the first half of this tenure (up to early 1990s) followed by a sharp decline afterwards,.

The rest of the paper is organized as in the followings. The next section provides an overview of related work followed by a description of the data, i.e. the corpus of speeches. Then, we describe the major methods we used to measure cohesion in Arabic (for English we use standard methods which can be found in the works cited in the Previous Work section) and the results obtained. We conclude the paper with Discussion and Conclusions.

Previous Work

Discourse cohesion refers to forming connections among parts of the text using its surface elements (Tanskanen, 2006). Cohesion can also be described based on “how repetition is manifested, in numerous ways, across pairs of sentences” in a discourse (Boguraev & Neff, 2000).

Therefore, the degree of discourse cohesion can be computed by measuring the overlap between adjacent sentences in a discourse (Graesser et al., 2004). The overlap can simply refer to the proportion of overlapped words between two adjacent sentences, or can be extended to include semantic or conceptual overlap in which case more sophisticated methods, such as Latent Semantic Analysis (LSA; Landauer et al., 2007) that can automatically discover latent concepts underlying the meaning of texts, may be used. Cohesion between adjacent sentences is called local cohesion while cohesion between larger and more distant fragments of texts is called global cohesion. A gap in cohesion forces the reader to make inferences which, if successful, helps comprehension (McNamara, Cai, Louwerse, 2007). However, readers often lack the knowledge or skills to make such inferences and therefore cohesion facilitates comprehension. We show in this paper that leaders tend to increase the cohesion of their speeches during and immediately after major crises, as a way to make their policies be easily comprehended, which serve the ultimate purpose of autocratic leaders such as Mubarak to extend their tenure in power. Our use of LSA to measure cohesion in Arabic speeches which, to the best of our knowledge, has not been done before.

The relationship between status and language has been studied before. Broniatowski and Magee (2011) used meeting transcripts to automatically discovery status and leadership styles using Bayesian topic models. Analysis of leaders’ language in chatrooms and speeches has been studied before by Moldovan, Rus, and Graessser (2009), Rus et al., (2010), and Lubna, Rus, & Graesser (2010), who looked at the distribution of speech acts for the purpose of identifying status, i.e. leaders and followers. For instance, leaders would most likely have a higher distribution of speech acts such as commands. Leaders’ language has also been studied, for instance, to examine personalities and psychological states of the 2004 candidates for U.S. president and vice president (Slatcher et al., 2007). Rus and colleagues (2010) analyzed referential cohesion and word concreteness in a leader’s speech but they only did it for English. In our case, we measure cohesion in two languages, the original language which in our case is Arabic and also in English, using both LSA and word overlap.

The Data: Corpus of Speeches

We started by collecting a corpus consisting of 902 speeches delivered between 1981-2011 by Hosni Mubarak in Arabic. The speeches were collected from the Egypt State Information Service website at http://www.sis.gov.eg. A subset of 306 of those speeches was available in English as well. These latter speeches were translated by humans. All speeches were labeled with the speech delivery date. We present next a quick, shallow analysis of both the Arabic and English speeches which will shed some light on some differences between the two versions of the same speeches. The cohesion analysis will be presented later.

Arabic Speeches

Mubarak’s Arabic speeches vary in length from short speeches composed of only two sentences to speeches that are more than 720 sentences long. It should be noted that the short speeches in terms of number of sentences are not that short in terms of number of word occurrences, i.e. tokens. A word token is any string of characters between two spaces (punctuation separated as well). English tokens and Arabic tokens are different because, for instance, a pronoun followed by a verb are two tokens in English while they will be one token in Arabic as some pronouns are attached to words in Arabic (“I love you” is one word:أحبك).

There is a tendency in Arabic in general and in Mubarak’s speeches in particular for the sentences to be long. The total number of sentences in all Arabic speeches adds up to more than 52,000 sentences while the number of words totals to about 1.3 million words. Calculated as the average number of tokens (i.e. word occurrences excluding punctuation marks) per sentence (WPS), the mean sentence length of each of these speeches averages to about 33 words but reaches a maximum of almost 100 words per sentence.

English Speeches

Out of the 902 collected Mubarak speeches, only 306 speeches have their human English translations available on the Egypt State Information Service site. This smaller subset of speeches appears to have, in their original Arabic version, a higher sentence count average of 75 sentences and a significantly shorter sentence length average of about 25 WPS. By comparison, the complete set of 902 original speeches (in Arabic) has an average sentence count of 58 sentences per speech and the average sentence length is 33 WPS. The total number of sentences in the Arabic version of the 306 speeches sums up to almost 23,000 sentences while the total number of words is more than 390,000.

Interestingly, the sentence and word per sentence (WPS) counts vary between the English translations of the 306 speeches and their original Arabic equivalents. While both original Arabic version of these 306 speeches and their English translations are comparable in terms of average sentence length (about 25 WPS in Arabic and 26.5 WPS in English), the English translations tend to have a smaller sentence count average per speech of about 65 compared to 75 in the original Arabic speeches.

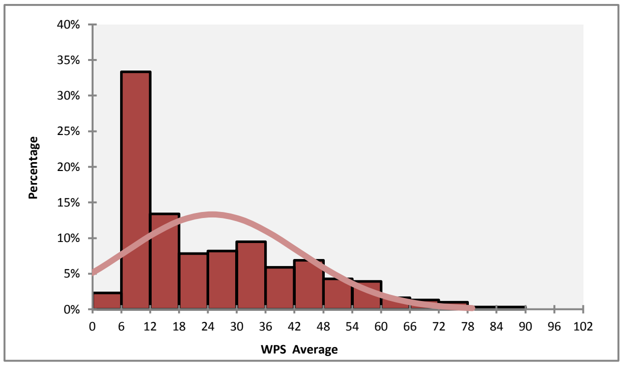

Figure 1. Distribution of Sentence Length for Arabic Speeches Subset with a Superimposed Normal Distribution Curve (N=306, $\mu = 24.99$, $\sigma = 17.99$)

Despite the fact that both subsets have similar WPS means, the two distributions look clearly different due mainly to the large difference in standard deviation values. The Arabic speeches' standard deviation (=.179) is about four times that of the English speeches (=.520). As a result, the WPS averages of the English speeches are clustered mostly around the mean value (26.54) while for the Arabic speeches the WPS averages are spread over a much larger range of values. Furthermore, the distribution of the Arabic speeches is more heavily skewed to the left indicating higher frequencies of shorter sentences in these speeches. About 57% of the speeches in the Arabic subset have WPS average of 24 sentences or less while only 27% of the English speeches have WPS average of 24 sentences or less.

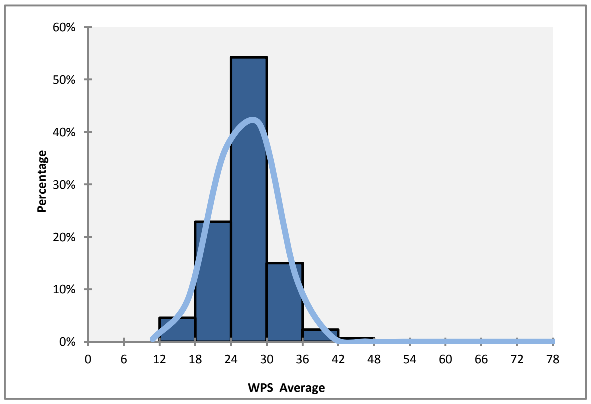

Figure 2. Distribution of Sentence Length for English Speeches Subset with a Superimposed Normal Distribution Curve (N=306, $\mu = 26.54$, $\sigma = 5.20$).

The difference in sentence count per speech and WPS and also the opposite trends of these two indices for the original Arabic speeches and their English translations can only be explained by a translation effect. Translators seem to reduce the number of sentences in a speech when translating compared to the original version of the speech. In order to check whether the information content is the same, i.e. there are no major content differences among the original and translated version, further analysis of the distribution of content words as well as function words is necessary. We plan to investigate this in future work. However, given that the translations were produced by a government-accredited organization it is safe to assume the information content may be mostly preserved.

It is important to note the difference in terms of WPS between the overall set of 902 speeches and the 306 speeches with English translations, which are from the later period of Mubarak's tenure (the 306 speeches are from the 1996-2011 period). It could be that later in his life, Mubarak preferred shorter sentences which could be a result of aging or a deliberate strategy to make his policies clearer (or both). Indeed, as people age their speech and writing becomes simpler in syntax and less dense in information (Kemper, 1987; Norman, Kemper, and Kynette, 1992). In other words, less complex syntax could be due partly to declines in working memory (Norman et al., 1992) but also may reflect awareness that simpler syntax is easier for listeners or readers to understand.

Measuring Cohesion Using Latent Semantic Analysis

Latent Semantic Analysis is a method for extracting and representing human conceptual knowledge by applying statistical computations to a large corpus of text (Landauer and Dumais, 1997). The fundamental concept behind LSA is that the meaning of a word is captured by "the company it keeps", i.e. by the words around it in natural texts. That is, the similarity between two words is defined by their likelihood to occur in the same contexts. LSA represents the meaning of individual words using a vector-based representation. To generate this representation, the LSA method starts with deriving word co-occurrence statistics from a large collection of documents capturing which words occur in which documents in the collection (i.e., a term-by-document matrix is generated at this step). Words co-occurring in the same documents will have more similar vector representations, which translates into the corresponding vectors being closer to each other in the space generated by the documents. As the number of documents from which the co-occurrence information must be derived is usually large, LSA relies on a mathematical procedure, called Singular Value Decomposition (SVD), to

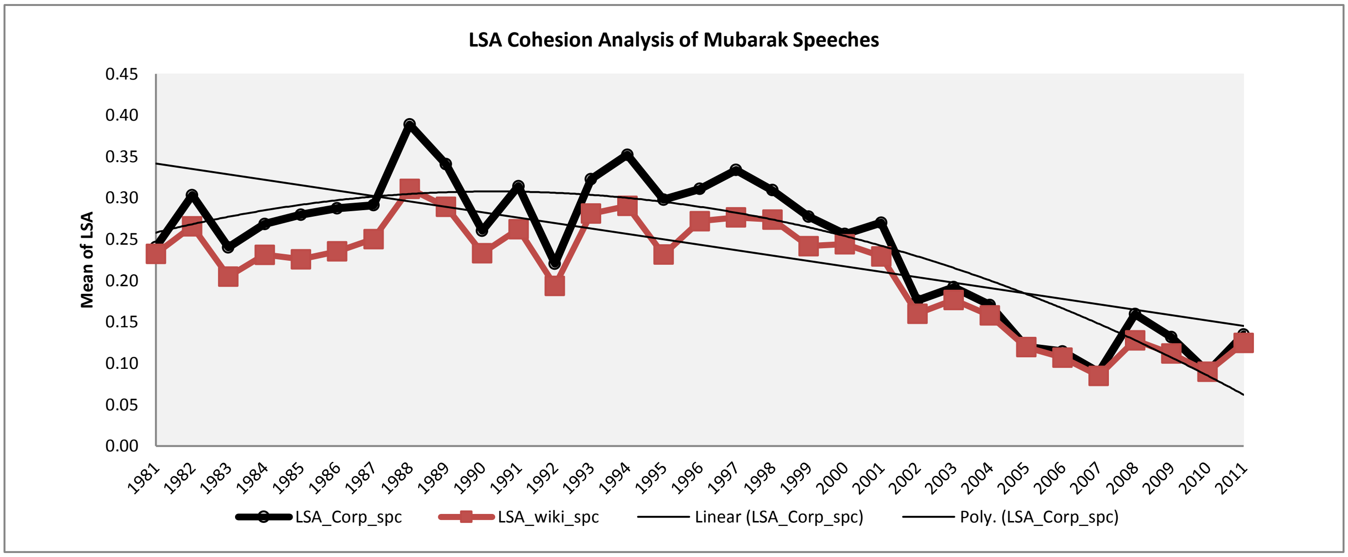

Figure 3. LSA cohesion trends for the Arabic speeches for the entire 1981-2011 period.

reduce the number of dimensions for the vector representation to about 300-500 dimensions. Both words and documents are represented in this reduced LSA space of latent dimensions. The latent dimensions can be considered as representing latent concepts underlying the text from which the LSA space was derived.

Given such a vector representation for words, the similarity of two words can be computed as the normalized dot-product, called cosine, between corresponding vectors. Due to additivity of vectorial representations, the meaning of a whole sentence (or large texts for that matter) can be represented by a vector itself by simply adding the individual words' corresponding LSA vectors. Thus, the semantic or conceptual similarity of two sentences could be quantified the same way as for words, i.e. by computing the normalized dot-product of the two sentences' LSA vectors. This natural extension of LSA to handle sentences is being exploited to analyze texts at discourse level, as explained next.

When used to model discourse cohesion, LSA tracks the overlap and transitions of meaning as they move across discourse by computing the semantic similarity of text segments (Crossley, Salsbury, McCarthy, & McNamara 2008). In our case, we will use LSA to measure the cohesion of Mubarak's speeches by measuring the degree of semantic overlap between adjacent sentences in his speeches. We will need to first obtain the LSA representation for each sentence in a speech, for which we need to derive an LSA space, and then find the semantic overlap, as a measure of cohesion, by computing the cosine between pairs of adjacent sentences. The cohesion of an entire speech is simply the average of the cohesion scores for all adjacent sentence pairs.

In order to build an LSA semantic space, a large corpus of Arabic texts was collected. About 72 MB of Arabic texts, composed of books (history, novels, philosophy, politics, and studies) and news articles (economy, entertainment, health, politics, sports, and others), was collected from online sources. However, Arabic sources are harder to collect compared to other languages and the resulting corpus is unbalanced in terms text genres as it is mostly composed of novels and history books. This unbalance raised a concern of the validity of the LSA space created from this corpus. As a result, a second, much larger corpus of 225 MB of Arabic texts that is composed of over 500 thousand Arabic Wikipedia articles was obtained. The downside of the new corpus is that there is no guarantee that Wikipedia articles are originally written in Arabic as some of them might have been translated from other languages such as English.

Two 300-dimension LSA spaces were built from the two corpora. The number of 300 dimensions which we used for the LSA spaces has been empirically found to yield good results by many research groups over many semantic processing tasks. Both LSA spaces led to comparable cohesion scores when used on Mubarak 902 speeches as can be seen in Figure 3. In fact, the yearly average LSA scores for Mubarak speeches using both spaces have a correlation of 0.998, which validates both LSA spaces. Since both spaces were valid, we have chosen to use the Arabic Wikipedia space in all consequent LSA analyses presented in this paper. The average LSA score over the 1981-2011 period is 0.227.

The LSA cohesion of Mubarak speeches is shown to have generally decreased over time: the year-LSA score correlation is -0.72.

We also conducted a cohesion analysis based on LSA using the English version of Mubarak's speeches. Similar to the analysis just described for the speeches in Arabic, we first derived an English LSA space using a large corpus of 37,520 English texts which contain 12,190,931 word tokens. The corpus, compiled by Touchstone Applied Science Associates (TASA), is an untagged collection of educational materials representing texts that typical high school students have encountered throughout their lifetimes from various genres (science, language arts, health, economics, social studies, business, and others).

Similar to the Arabic analysis, a 300-dimension LSA space was built from the TASA corpus and the yearly average LSA scores for all Mubarak's English translated speeches were calculated (average LSA score is 0.295). The LSA cohesion of the English speeches has also generally decreased over time with year-LSA Score correlation of -0.39 which is considerably smaller than the correlation for the Arabic speeches (-0.72) implying that the decline in LSA cohesion over time was not as noticeable in the English speeches is it was for the Arabic ones. There are a couple of points we would like to make that could explain these differences. English LSA cohesion scores are generally higher than the Arabic LSA scores. For instance, the average LSA score for Arabic speeches during the 1996-2011 period is 0.197 while for the English speeches for the same period the average LSA score is 0.295. This is due to structural differences between the Arabic and English speeches (e.g. the presence of extremely large sentences in Arabic speeches – discussed later) and morphological differences between English and Arabic (Arabic is a morphologically complex language). The structural differences between the English and Arabic speeches could be translation effects or cultural effects, e.g. there is a tendency in Arabic texts for extremely long sentences. Furthermore, it is important to keep in mind that the Arabic dataset is considerably larger than the English translated dataset which may affect the results. To eliminate this confound, the subset of the Arabic speeches with English translations was analyzed. The year-LSA correlation for the Arabic subset is -0.90 which is even higher than the entire Arabic speeches set correlation.

Measuring Cohesion Using Word Overlap

Word type overlap is computed by measuring the proportion of common word types (unique words) in each pair of adjacent sentences in the speeches. Overlap score is calculated using the formula:

$$WordOverlap = \frac{\#common\ word\ types}{average\ \#\ of\ types}$$

The overlap of all words (content and functional) can be considered as well as just content word overlap. The content words are the common nouns, verbs, adjectives, and adverbs. In order to compute content word overlap, functional words (such as prepositions) must be identified and removed from each speech before counting overlap between adjacent sentences.

There is a challenge when computing word overlap in Arabic due to morphological complexity of Arabic. Indeed, the morphological structure of words in Arabic is complex. An Arabic word is composed of a stem surrounded by affixes that inflect gender, number, and tense. Arabic is a clitic language which means determiners, conjunctions, prepositions, particles, and pronouns are often attached to the word stem as prefixes and suffixes. Due to this complex morphological structure of an Arabic word, word tokenization is a non-trivial step as one word can be tokenized into one or more tokens based on the selected scheme, which may have significant effects on overlap scores.

Given that affixes of a word are always tokenized into functional words (pronouns, prepositions, determiners, etc.), our cohesion measurements were conducted using the following four different schemes:

- All (content and functional) words overlap (no tokenization)
- Content word overlap with only prefixes separated
- Content word overlap with only suffixes separated
- Content word overlap with both prefixes and suffixes separated

Under each scheme, overlap scores are averaged for each speech and then the overall average for all speeches in a year is calculated. All 902 original Arabic speeches were first tokenized and tagged using AMIRA 2.0 Part of Speech Tagger for Arabic, a tool which offers several tokenization schemes for Arabic (Diab, 2009).

The average overlap scores for all four different schemes are considerably low: 0.0357, 0.0321, 0.0245, and .0293, respectively. Importantly, all overlap scores correlate highly with LSA scores and they all decrease over time. An interesting pattern that we noticed regards the percentages of adjacent sentences with at least one word overlap. This measure goes down over time from a maximum of 0.76 in 1988 to 0.12 in 2010, meaning that over time there are more and more adjacent sentences in Mubarak's speeches with no word overlap at all.

There are several possible reasons for the low overlap scores. One likely reason is the morphological complexity of Arabic words as the boundaries between derivation and inflection are not as precisely defined in Arabic as they are in English. This makes it very difficult for a morphological analyzer to separate all affixes. While AMIRA 2.0 does a decent job at tokenizing and assigning part of speech tags to words in the Arabic texts, many words are not tokenized properly resulting in variations of the same stem (I give, you give, we give…) to be marked as different words which decreases the word overlap score. LSA might address this issue as it somehow computes cohesion based on the underlying semantics rather than the surface level of words.

Another possible explanation for the low cohesion is that Arabic views sentence structure from a different perspective compared to English. Punctuations marks are not essential in Arabic writing and do not contribute to the sentence meaning as it is the case for English. Indeed, Arabic writers have a tendency to advance the discourse by introducing and elaborating on new topics in the same sentence using constructs such as conjugating conjunctions. The result is longer sentences and fewer sentences in a speech. The use of punctuation marks and the creation of new sentences as a way to introduce and elaborate on a new topic is less frequent. Instead, they prefer to combine related sentences in a long sentence (e.g. "I wanted to take a walk but it was raining so I drove to the library to work on my research and while I was there, the rain stopped and I decided to walk home and pick up my car later however after the first few steps I realized that I was carrying too many heavy books so I ended up driving home."). In other words, an Arabic sentence is more similar to an English paragraph than it is to an English sentence which means the sentence overlapping scores in Arabic may be comparable to word overlapping score between adjacent paragraphs in English, an interesting future research goal.

Discussion and Conclusions

The cohesion of Mubarak's speeches follows an up-down pattern over his tenure in power with a general tendency of going down. As can be seen in Figure 1, there is an upward trend in cohesion during the first half of his tenure (up to 1997) followed by a sharp decline. Indeed, the average LSA cohesion score during the first period is 0.255 while the average LSA cohesion score for the second period is 0.187, a significant drop. In particular, it can be observed from the figure that there are two periods in which there is a sharp increase in cohesion: 1986-1988 which coincides with the political changes in Eastern Europe and 1993-1997, a period of active opposition movements within Egypt by the major Muslim Brotherhood organization, a non-political organization at that time.

The general down trend for the cohesion of Mubarak's speeches can be explained by either or both of the following two hypotheses: (1) cohesion of speeches goes down with age – this is generally true for the whole population not only for leaders, (2) cohesion goes down the longer a leader stays in power; this is based on the "common ground" theory of discourse according to which the more a leader stays in power the more "common ground" there is between the leader and his followers which affords the leader being less cohesive without running the risk of his policies not being understood.

To find the exact cause, we plan to study other long-serving leaders' speech before they age (to eliminate the age confound) or before they stayed in power for too long (to eliminate the common ground confound).

References

Broniatowski, D.A. and Magee, C.L. (2011). Towards a Computational Analysis of Status and Leadership Styles on FDA Panels. Social Computing, Behavioral-Cultural Modeling and Prediction Lecture Notes in Computer Science Volume 6589, 2011, pp 212-218.

Boguraev, B.K. and Neff, M.S. (2000), Lexical Cohesion, Discourse Segmentation and Document Summarization, RIAO-2000.

Crossley, S. A., Salsbury, T., McCarthy, P. M., & McNamara, D. S. (2008) LSA as a measure of second language natural discourse. In V. Sloutsky, B. Love, and K. McRae (Eds.), Proceedings of the 30th annual conference of the Cognitive Science Society (pp. 1906-1911). Washington, D.C.: Cognitive Science Society.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., and Cai, Z. (2004). Coh-Metrix: Analysis of Text on Cohesion and Language. Behavioral Research Methods, Instruments and Computers, 36: 193–202.

Kemper, S. (1987). Life-span changes in syntactic complexity. Journal of Gerontology, 42, 323–328.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). Handbook of latent semantic analysis. Mahwah, New Jersey: Lawrence Erlbaum.

McNamara, D.S., Cai, Z., & Louwerse, M.M. (2007). Optimizing LSA measures of cohesion. In T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), Handbook of Latent Semantic Analysis (pp. 379-400). Mahwah, NJ: Erlbaum.

Norman, S., Kemper, S., & Kynette, D. (1992). Adults' reading comprehension: Effects of syntactic complexity and working memory. Journal of Gerontology: Psychological Sciences, 47, p. 258-265.

Shala, L., Rus, V. & Graesser, A. (2010). Automatic Speech Act Classification in Arabic. Subjetividad y Procesos Cognitivos Journal, Vol. 14, No. 2, 2010, pp. 284-292.

Rus, V., Shala, L., Graesser, A.C., Cai, Z., & Kaltner, J. (2010). Analysis of Leaders' Language and Discourse. International Conference of the Society for Text and Discourse, Chicago, IL, August 2010.

Slatcher, R.B., Chung, C.K., Pennebaker, J.W., & Stone, L.D. (2007). Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. Journal of Research in Personality, 41, 63-75.

Tanskanen, Sanna-Kaisa. (2006). Collaborating Towards Coherence: Lexical Cohesion in English Discourse. John Benjamins Pub Co.