

## Automatic Coherence Profile in Public Speeches of Three Latin American Heads-of-State

**René Venegas**

Pontificia Universidad Católica de Valparaíso  
rene.venegas@ucv.cl

### Abstract

Different studies provide evidence that the computational psycholinguistic algorithm called Latent Semantic Analysis (LSA) allows measuring local and global coherence in texts similarly to human evaluation (Foltz, Kintsch, Landauer 1998; McNamara, Cai & Louwerse 2007; McCarthy, Briner, Rus, & McNamara, 2007; McNamara, Louwerse & Jeuniaux 2009; Louwerse, McCarthy & Graesser 2010). The texts used in all these studies are written in English and correspond to scientific and literary texts. In Spanish, there are some studies using LSA that measure the semantic similarity between texts in automatic summary assessment (Pérez, Alfonseca, Rodríguez, Gliozzo, Strapparava & Magnini 2005; León, Olmos, Escudero, Cañas & Salmerón 2006; Venegas 2007, 2009, 2011); however, automatic measurement of coherence in Spanish has not yet been sufficiently investigated. The present study aimed at identifying a global and local coherence profile in a corpus of speeches in Spanish of three Latin American Heads of States (Perón, Castro and Pinochet), using Latent Semantic Analysis. Local coherence is calculated through the measurement of implicit semantic similarity between adjacent sentences and global coherence through the measurement of the similarity among the semantic content of the paragraphs. The corpus under analysis corresponds to a sample of 107 speeches. The semantic space was built using a multi register corpus and it is available through the "Interface for the measurement of lexical semantic similarity" in the El Grial interface ([www.elgrial.cl](http://www.elgrial.cl)). Results showed a systematic difference between the speeches of the Heads of State in terms of both local and global coherence. The Bonferroni analysis established an effect that distinguishes Perón's speeches from Pinochet's and Castro's speeches. This results show that Perón's speeches are more topically related than the other leaders', probably due to a discourse strategy to persuade voters. The identification of a profile of coherence might be relevant to predict cues of government discourse styles.

### Introduction

Cohesion and coherence are two main features in discourse comprehension and production. Cohesion is defined as "the use of explicit linguistic devices to signal relations between sentences and parts of texts" (Connor 1996:83). These cohesive devices are words or phrases that help the reader associate previous propositions with subsequent ones. Halliday and Hasan (1976) established that cohesion refers to the range of possibilities that exist for linking something with what has gone before, i.e. a meaning relation. There is always a possibility that a text be cohesive without necessarily being coherent: cohesion does not guarantee coherence: "cohesion is determined by lexically and grammatically overt intersentential relationships, whereas coherence is based on semantic relationships" (Connor 1996:83). Coherence then refers to how a text is organized, how the ideas are developed, and how the contents are related. In sum, coherent texts make sense to the reader. Despite the possible controversy in the conceptualization of both terms (Sanders & Pander 2006), in the present study we conceptualized coherence as a representational relationship and cohesion as the textual indications that coherent representation should be built on (Louwerse 2004). More specifically, we focused in local and global coherence.

Both local and global coherence are interesting to be studied in linguistic and psycholinguistic approaches to discourse (Sanders & Spooren 2009; Louwerse & Jeuniaux 2009). The measurement of coherence relations has been explored computationally using different techniques. This line of research has been supported by studies developed by researchers like Foltz, Kintsch, and Landauer (1998), McNamara, Cai, and Louwerse (2007), McCarthy, Briner, Rus, and McNamara (2007); McNamara, Louwerse, and Jeuniaux (2009), Louwerse, McCarthy, and Graesser

(2010) and McNamara, Louwerse, McCarthy, and Graesser (2011). All these studies are oriented by the so called statistical symbolism approach, inspired by mathematical analysis of texts that account for developmental data, inductive learning, and such psycholinguistic phenomena as word frequency effects, similarity, and semantic priming (De Vega, Graesser & Glenberg 2008). In order to study the cohesion and coherence in texts, Graesser, McNamara, Louwerse, and Cai (2004) developed the computational tool Coh-Metrix.

Coh-Metrix provides a wide range of computational linguistic indexes to meet the comprehensive and automatic text analysis (Graesser, McNamara, Louwerse, and Cai 2004: 294).

Among the different indexes included, the computational psycholinguistic technique, Latent Semantic Analysis (LSA), compute the semantic similarities between words, sentences, and paragraphs by applying statistical computations to a large corpus of texts (Deerwester, Dumais, Furnas, Landauer, & Harshman 1990; Landauer & Dumais 1997; Landauer, McNamara, Dennis, & Kintsch 2007). In fact, different studies using LSA indexes demonstrated the validity of this technique to measure semantic relatedness like human grading in tasks like test grading, summary assessment, cohesion and coherence measurement, etc. Most of these studies were implemented using academic, scientific and/or literary texts written in English. An exception is the study conducted by Graesser, Dowell and Moldovan (2011) using literary and political texts. Comparatively, in Spanish, there are few studies that validate the use of LSA in relation to human assessment, measuring the semantic similarity between texts in automatic summary assessment (Pérez, Alfonseca, Rodríguez, Gliozzo, Straparava & Magnini 2005; León, Olmos, Escudero, Cañas & Salmerón 2006; Olmos, León, Escudero, & Botana 2009; Venegas 2007, 2009, 2011). Interestingly the study conducted by Hernández and Ferreira (2010) is probably the first investigation in Spanish that analyzed computationally coherence in police news. These researchers used six texts corresponding to Chilean journal news and selected eight evaluators. Four of them were experienced journalists and the other four were postgraduated Spanish teachers. The results showed that there is a high positive correlation between LSA and the journalists ( $r = .95$ ,  $p = .003$ ) and the teachers ( $r = .85$ ,  $p = .03$ ).

Based on the evidence that LSA is a reliable tool to measure coherence both in English and Spanish, the present study aimed at identifying a global and local coherence profile in a corpus of speeches in Spanish of three Latin American Head-of-States (Perón, Castro and Pinochet), using Latent Semantic Analysis.

### Latent Semantic Analysis

In a nutshell, LSA is a mathematical-statistical technique used for the extraction and representation of meaning

relations between words and paragraphs from a large number of texts. LSA extracts meaning representations from words and paragraphs exclusively based on mathematical-statistical analysis of the text. This idea of meaning representation without resorting to syntax complemented the notion that there are weak semantic interrelations between words and between paragraphs in these large amounts of corpora, which are empowered by the dimension reduction method known as Singular Value Decomposition (SVD) (Deerwester, Dumais, Furnas, Landauer & Harshman 1990). Singular Value Decomposition is used to construct a semantic space representative of the information required for one or several knowledge domains based on a text corpus. LSA thus enables the calculation of semantic similarities between words and paragraphs in texts, establishing measurements of vector representation by calculating the cosine of angles in a high vector space (Landauer, Foltz, & Laham 1998; Landauer, McNamara, Dennis & Kintsch 2007). The cosine values range from 1, for vectors running in the same direction (this means that what is measured is the same), to 0 for those orthogonal vectors (perpendicular in high dimensional space, meaning that what is measured is completely different).

LSA, as described before, allows to track the overlap and transitions of meaning as they move across written discourses by computing the semantic similarity of text segments. As discourse moves from one utterance to another, an LSA cosine can be assigned to each section of text. Discourse that is highly coherent is represented by higher cosine values as compared to less coherent discourse, which has lower cosine values (Crossley, Salisbury, McCarthy, & McNamara 2009).

## Methods

### Semantic Space

A 10,242,384-word corpus of diverse Spanish texts known as COTEGE (General Spanish Corpus) was employed for construction of the semantic space. This corpus comprises five multi-register corpora:

1. PUCV-2003 Corpus: A corpus collected by the FONDECYT 1020786 research team. This corpus is divided into 90 texts, equivalent to 1,466,744 words. In turn, this corpus is divided into three subcorpora (Technical-Scientific Corpus -CTC-, Written Latin American Literature Corpus -CLL-, and the Oral Interview Corpus -CEO-). This corpus is available on-line at [www.elgrial.cl](http://www.elgrial.cl)

2. Oral Spanish Corpus: A text corpus collected and transcribed by Universidad Autónoma de Madrid. The corpus contains a total 1,099,400 words and features 12 oral genres. This corpus is available at [www.llf.uam.es/corpus/corpus.html](http://www.llf.uam.es/corpus/corpus.html)

3. Contemporary Spanish Reference Corpus: A corpus collected by Universidad Autónoma de Madrid, which included the Reference Corpus for the Spanish Language in Chile and the Reference Corpus for the Spanish

Language in Argentina. This corpus comprises a total of 3,156,491 words with 10 different text genres. This corpus is available on-line at <ftp://ftp.uba.ar/pub/misc/corpus/>

4. Written Narrative Corpus: This corpus contains 86,3981 words and includes narrative texts. It is available on-line at [www.elgrial.cl](http://www.elgrial.cl)

5. ARTICO Corpus: The Corpus of Original Scientific Research Articles comprises 678 articles published in Scientific Electronic Library Online (SciELO) between 2000 and 2003. The corpus comprises a total 3,655,768 words (available on-line at [www.elgrial.cl](http://www.elgrial.cl)).

Following the application of SVD, COTEGE was transformed into a semantic space. This semantic space consisted of 297 dimensions and 99,966 unique words, with the assignment of their corresponding values and is available for use at [www.elgrial.cl](http://www.elgrial.cl).

## Research Corpus

The corpus analyzed in this research was conformed by the speeches of three Latin American Heads-of-State. The speeches under analysis were chosen considering that all the Heads-of-States participated on coups d'état and led different kinds of authoritarian regimens (Payne 1995; Hadenius & Teorell 2006; Meyer, 2010). Perón's regimen can be classified as a populist authoritarianism, Castro's regimen as one-party authoritarianism, and Pinochet's regimen as a military traditional authoritarianism.

Perón's speeches were collected from the public webpage: <http://www.pjmoreno.org.ar/documentos/discursosperon.aspx>. This webpage includes 21 speeches available in Spanish. The speeches are a sample of the speeches delivered by Perón between 1945 and 1974.

Castro's speeches were collected from the public webpage [www.cuba.cu/gobierno/discursos/](http://www.cuba.cu/gobierno/discursos/). This webpage includes 1115 speeches available in different languages. For this research, a sample of 71 speeches between 1959 and 2008 were selected.

Pinochet's speeches were collected from different libraries of governmental institutions (e.g. Congreso and Ministerio de Vivienda y Urbanismo) and libraries of private organizations (e.g. FLACSO, Cámara Chilena de la Construcción, Fundación Pinochet). The Pinochet subcorpus included 16 speeches corresponding to the yearly public account given each year between 1974 and 1989. It is worth noticing that in the Internet do not exist a webpage where all these speeches are accessible. Therefore, the speeches were digitalized.

Head of State	Speeches	Words	Sentences	Paragraphs
Perón	21	6,2370	3,170	1,137
Castro	70	48,2740	26,953	14,908
Pinochet	16	13,1148	5,386	4,146
Total	107	67,6258	35,509	20,191

Table 1. Details of speeches used in this study

The measurement of the cosine for local and global coherence was made using the "Interface for the measurement of lexical-semantic similarity" (<http://158.251.61.111/compareFilesFrase>). Using this interface, it is possible to compare several texts for different tasks. Local coherence was measured calculating the cosines values between each pair of adjacent sentences, segmented manually, in all the speeches. Global coherence was measured calculating the cosines values among the paragraphs, also segmented manually, of each speech.

## Results

### 1. Analysis of shallow variables for speeches

A first analysis considering some shallow variables of texts is presented. These variables have been found to be important indicators of style in texts (McCarthy, Lewis, Dufty, and McNamara 2006). For this study the information obtained comparing the number of words, sentences, paragraphs, and the ratio of words by sentences (WS), words by paragraphs (WP), and sentences by paragraphs (SP) can complement the local and global coherence analysis.

Head of State	Variable	N	MEAN	STD.DEV.
Perón	Words	21	2,970.000	4,057.607
	Sentences	21	150.952	173.750
	Paragraphs	21	54.143	101.949
	WS	21	18.717	5.893
	WP	21	65.379	24.726
	SP	21	3.805	1.756
Castro	Words	71	6,799.155	4,468.119
	Sentences	71	379.620	265.542
	Paragraphs	71	209.972	140.821
	WS	71	19.127	7.488
	WP	71	33.975	12.363
	SP	71	1.845	0.549
Pinochet	Words	16	8,196.751	5,398.947
	Sentences	16	336.625	206.442
	Paragraphs	16	259.125	177.881
	WS	16	23.614	3.768
	WP	16	35.097	7.803
	SP	16	1.526	0.461

Table 2. Shallow variables in the speeches of each Heads of State.

An analysis of variance (ANOVA) was conducted on these variables to give an overview of which variables can account for differences among the speeches. The results show statistical differences in almost all these shallow variables. All variables produce a significant effect at the .05 level or higher. The variables are ranked by F value: a) SP ( $F=42.93$ ;  $p=0.000$ ); WP ( $F=36.29$ ;  $p=0.000$ ); paragraphs ( $F=12.45$ ;  $p=0.000$ ); words ( $F=7.47$ ;  $p=0.000$ ); sentences ( $F=7.20$ ;  $p=0.001$ ); WS ( $F=3.13$ ;  $p=0.047$ ). According to these results, all the variables distinguish among the speeches. The Bonferroni test indicates that Perón's speeches present a larger SP ratio

(Mean=3.8;  $p < .05$ ) compared with Castro's and Pinochet's speeches. The same result is observed by WP ratio (Perón > Castro = Pinochet;  $p < .05$ ). According to the number of paragraphs, the speeches with more paragraphs are Castro's and Pinochet's speeches (Mean<sub>castro</sub>=209.97; Mean<sub>pinochet</sub>=259.12;  $p = .626$ ). In terms of number of words, the larger are those of Castro's and Pinochet's speeches (Mean<sub>castro</sub>= 6799.15; Mean<sub>pinochet</sub>=8196.75;  $p = .805$ ). Similar results were found in regards to the number of sentences, i.e. the number of sentences is higher in Castro's and Pinochet's speeches (Mean<sub>castro</sub>=379,62; Mean<sub>pinochet</sub>=336,62;  $p = 1$ ). The WS ratio does not distinguish between the speeches of the three Heads-of-State (Mean<sub>perón</sub>=18,717; Mean<sub>castro</sub>=19,127; Mean<sub>pinochet</sub>=23,61). As we can observe, on the one hand, Perón's speeches are different from Castro's and Pinochet speeches in terms of almost all the variables. The sentence by paragraph ratio and the word by paragraph ratio are higher in Perón's speeches than Castro's and Pinochet's speeches. On the other hand, there is no difference between Castros's and Pinochet's speeches.

## 2. Local and global coherence

The results in this section examine whether LSA detected differences in the measurement of the implicit semantic similarities among the adjacent sentences, i.e. local coherence (LSAloc) and the implicit semantic similarities among the paragraphs in the speeches, i.e. global coherence (LSAglo).

Heads of State	Coherence level	MIN	MAX	MEAN	STD. DEV.
Perón	LSAloc	.12	.22	.17	.02
	LSAglo	.23	.45	.30	.05
Castro	LSAloc	.03	.23	.13	.04
	LSAglo	.10	.27	.20	.03
Pinochet	LSAloc	.13	.26	.20	.03
	LSAglo	.21	.33	.26	.02

Table 3. Local and global coherence measurement in the speeches of each Head-of-State.

The Figures 1, 2, and 3 show the distribution of the local and global coherence values across time.

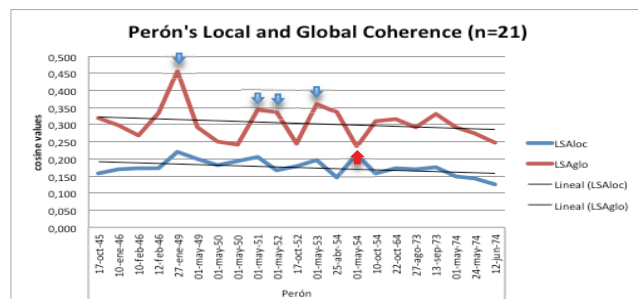


Figure 1. Perón's local and global coherence values

In the 21 Perón's speeches we observe heterogeneity among the global coherence values during the time line. Some peaks are evident (speeches pronounced at 27th January 1949, 1st May 1951, 1st May 1952, and 1st May 1953). The 1949 speech was delivered to the Constituent Reformer Assembly, which is a prepared and a very formal speech, the 1st May speeches were delivered in commemoration of the Labor Day at the plaza de Mayo in Buenos Aires. These speeches are characterized by the presence of oral features and by a strong appeal to the audience. The most important valley in Figure 1 corresponds to the speech pronounced on 1st May 1954. This speech is considerably shorter in comparison with the other 1st May speeches and some paragraphs are made up of, for example, by 2 lines or a list of names. The relation between the local and global coherence values is positive medium (pearson,  $r = .29$ ). In Perón's speeches, there is, through years, a linear tendency to be less coherent both in local and global level.

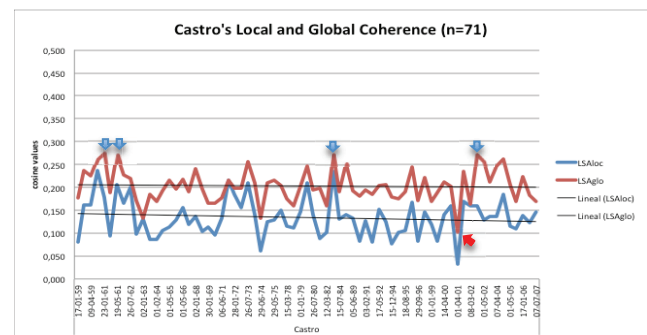


Figure 2. Castro's local and global coherence values

Overtime Castro's speeches show some salient high values, particularly during the year 1961 (23-01-1961 and 19-05-61) and some in the years 1983 and 2002. The speech of January 1961 was pronounced in the graduation ceremony of volunteer teachers. The speech of May 1961 was pronounced in the context of the reception of the "Lenin for the Peace" award. The speech of April 1983 was pronounced at the VII Conference of the Movement of Non-Aligned Countries and the speech of March 2002 was delivered at the ceremony for the 45<sup>th</sup> anniversary of the attack on the presidential palace. The first three speeches are longer than the mean. The style is probably influenced by the formal context of the official ceremonies; this might explain the higher local and global coherence scores. Unlike the other speeches, Castro's speech of 2002 is a recount of the events leading to the death of his comrades and their heroic deeds. It is in comparison a shorter speech but with a higher word by paragraph ratio. The most conspicuous local and global coherence valley is in the speech of 1<sup>st</sup> April 2001. This speech was given at the 105<sup>th</sup> Conference of the Interparliamentarian Union. It is the shortest speech and it appears less elaborated in comparison with his most coherent speeches. Castro's speeches are radically different to Perón's and Pinochet's, as we focus on the relation between the local and global



coherence (pearson,  $r = .74$ ). Finally, a linear tendency to be less coherent through the years is observable in the distribution of the data.

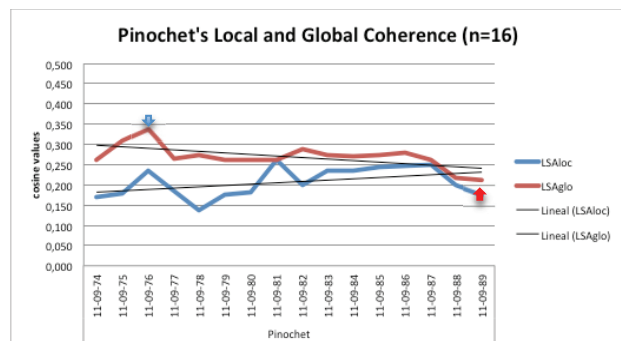


Figure 3. Pinochet's local and global coherence values

Pinochet's speeches show a high global coherence peak in the speech given at the 11th September 1976. It is a short speech with one of the highest word by paragraphs and sentence by paragraph ratio. It is clearly a prepared speech, including topics like the establishing of the new Chilean institution and the promise of prosperity for the country. The least globally coherent speech was delivered at 11<sup>th</sup> September 1989. This speech is longer than the mean, but the word by paragraph and sentence by paragraph ratio is lower in comparison with mean. In the speech Pinochet highlights the Chilean context before the coup d'état, summarizes the achievements of his government, and argues against the center-right wings and left wings parties in relation to the political and constitutional charges presented by these parties. Something relevant to this speech is the inclusion of other voices, representing the opponents of his regimen. It is worth noticing, that the relation between local and global coherence scores is very low related (pearson,  $r = 0.19$ )

The results given by the ANOVA analysis at the local level of coherence allow us to establish a statistical difference among the three Heads-of-State's speeches ( $F = 28.74$ ;  $p = .000$ ), whereas the Bonferroni analysis identifies differences among the three group of speeches ( $\text{Mean}_{\text{perón}} = .175 - \text{Mean}_{\text{castro}} = .134$ ,  $p = .000$ ;  $\text{Mean}_{\text{perón}} = .175 - \text{Mean}_{\text{pinochet}} = .207$ ,  $p = .036$ ;  $\text{Mean}_{\text{castro}} = .134 - \text{Mean}_{\text{pinochet}} = .207$ ,  $p = .000$ ). The results indicate that Pinochet's speeches are the most coherent at the local level, followed by Perón's speeches and finally by Castro's speeches.

At the global coherence level, the analysis of variance shows, like the local coherence level, a statistical difference among the speeches ( $F = 65.13$ ,  $p = .000$ ). The Bonferroni analyses differentiate statistically among the three groups of speeches ( $\text{Mean}_{\text{perón}} = .304 - \text{Mean}_{\text{castro}} = .203$ ,  $p = .000$ ;  $\text{Mean}_{\text{perón}} = .304 - \text{Mean}_{\text{pinochet}} = .269$ ,  $p = .022$ ;  $\text{Mean}_{\text{castro}} = .203 - \text{Mean}_{\text{pinochet}} = .269$ ,  $p = .000$ ). Therefore, Perón's speeches show a higher global coherence mean value, followed by Pinochet's speeches and finally by Castro's speeches.

The shallow variables are not useful to distinguish between Castro's and Pinochet's speeches. On the contrary, the local and global coherence values allow distinguishing among the three groups of speeches. This result provides more evidence in favor of the use of semantic similarity measures with LSA to complement these shallow variables for comparing discourse styles based on the coherence in texts.

According to the results presented before, it is possible to state that Perón's speeches present a higher sentence by paragraph ratio and higher word by paragraph ratio. These results might be interpreted as a higher textual density and therefore a more cohesive speech style. This interpretation can be related to the results obtained by the global coherence measurement, allowing to state that Perón's speeches are not only more cohesive, but also more globally coherent than the other speeches.

## Conclusions

We have shown that there are differences between the speeches of the Heads-of-State. According to the results, Perón's speeches cohere more than Pinochet's and Castro's. This difference can be explained due to the higher values in SP and WP ratio (indicating higher information density) and the global coherence scores. An interpretation for this result is that the Argentinian populist authoritarianism represented by Perón sustained a strong discourse strategy associated to the persuasion of the working class and the less favored population; therefore the speeches are repetitive and topically organized to persuade voters. It is worth noticing that Perón was three times elected in Argentina after the coup d'état, i.e. he lead a legitimated government. As we observed, the local coherence score is lower than Pinochet's, which could be explained in relation to the less prepared speeches. Like Castro's speeches, Peron's were oral speeches transcribed and therefore it is possible to notice some gaps in the relatedness between the adjacent sentences.

Castro's speeches are characterized in general terms by a low local and global coherence. There is also a high heterogeneity between the speeches over time. An interesting result is the high correlation between the local and global coherence scores, which can be associated to a particular oratory style. An interpretation of these results is that the relation between orality, interaction with the audience, and the size of the speeches affect the coherence in the texts. The topics of the speeches are more related to the maintenance of the ideal of the revolution trough narrative and anecdotic sequences and the opposition to the external economic and political forces, two characteristics that are associated to one-party authoritarianism.

Pinochet's speeches are best characterized by a high local coherence and less global coherence. This result might be explained in terms of higher degree of elaboration and formality in which these speeches were produced and delivered. The speeches are also as long as Castro's, something that indicates the presence of a wider range of

topics and therefore a smaller global coherence index. The relevant topics in the speeches are related to the recount of the events that took place before the coup d'état and the justification of the authoritarian military regime over a 16-year period.

In general terms, it is interesting that the deeper semantic relations measurements were more useful to distinguish the speeches than the shallow level variables.

These results provide evidence for the importance of, at least, complementing both kinds of variables to characterize the coherence in texts and authors discourse styles. These findings are consistent with some results obtained in the analysis of other authoritarian leader discourse, like Saddam Hussein, using the paradigm of social language processing (Hancock, Beaver, Chung, Frazee, Pennebaker, Graesser and Cai 2010).

Finally, a systematic decrease in global coherence is found in all the latter speeches of each Heads-of-State, probably due to an increase in the common ground.

A deeper qualitative and quantitative analysis should be necessary to better characterize these speeches, but these preliminary results suggest that the identification of a profile of coherence might be relevant to predict cues of government discourse styles.

## References

- Connor, U. 1996. *Contrastive Rhetoric: Cross-cultural aspects of second-language writing*. Cambridge, CUP.
- de Vega, M., Graesser, A., and Glenberg, A. 2008. Reflecting on the debate. In M. de Vega, A. Glenberg, and A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 397-440). Oxford, U.K.: Oxford University Press.
- Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, 391-407.
- Foltz, P., Kintsch, W., and Landauer, T. 1998. The measurement of textual Coherence with Latent Semantic Analysis. *Discourse Processes*, 25, 285-307.
- Graesser, A., McNamara, D., Louwerse, M., and Cai, Z. 2004. Coh-matrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193-202.
- Graesser, A., Dowell, N. and Moldovan, C. 2011. A computer understanding of literature. *Scientific Studies of Literature*, 1(1), 24-33.
- Hancock, J., Beaver, D., Chung, C., Frazee, J., Pennebaker, J., Graesser, A. and Cai, Z. (2010). Social language processing: A framework for analyzing the communication of terrorists and authoritarian regimes. *Behavioral Sciences of Terrorism and Political Aggression*, 2(2), 108-132.
- Halliday, M. and Hasan, R. 1976. *Cohesion in English*. London: Longman UK group Limited.
- Hernández, S., and Ferreira, A. 2010. Evaluación automática de coherencia textual en noticias policiales utilizando Análisis Semántico Latente. *RLA. Revista de Lingüística Teórica y Aplicada*, 49(1), 51-68.
- Landauer, T., and Dumais, S. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211-240.
- Landauer, T., Foltz, P., and Laham, D. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.
- Landauer, T., McNamara, D., Dennis, S., and Kintsch, W. (Eds.). 2007. *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum
- León, I., Olmos, R., Escudero, I., Cañas, J., and Salmerón, L. 2006. Assessing short summaries with human judgments procedure and Latent Semantic Analysis in narrative and expository texts. *Behavior Research Methods, Instruments and Computers*, 38(4), 616-627.
- Louwerse, M. 2004. Un modelo conciso de cohesión en el texto y coherencia en la comprensión. *Revista Signos. Estudios de Lingüística*, 37(56), 41-58
- Louwerse, M., and Jeuniaux, P. 2009. Computational psycholinguistic techniques to measure cohesion in discourse. In J. Renkema (Ed.), *Discourse of course* (pp. 213-226). Amsterdam: Benjamins.
- Manning, C. and Schütze, H. 2003. *Foundations of statistical natural language processing*. Massachusetts: MIT Press.
- McCarthy, P., Briner, S., Rus, V., and McNamara, D. 2007. Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. In A. Kao, and S. Poteet (Eds.), *Natural language processing and text mining* (pp. 107-122). London: Springer-Verlag
- McCarthy, P., Lewis, G., Dufty, D., and McNamara, D. 2006. Analyzing writing styles with Coh-Matrix. In Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS) (pp. 764-770).
- McNamara, D., Cai, Z., and Louwerse, M. 2007. Comparing latent and non-latent 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
- McNamara, D., Louwerse, M., McCarthy, P., and Graesser, A. 2010. Coh-matrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4), 292-330.
- Olmos, R., León, J., Escudero, I. and Botana, J. 2009. Efectos sobre el tamaño y especificidad de los corpus en la evaluación de resúmenes mediante el LSA y jueces expertos. *Revista Signos. Estudios de Lingüística*, 41(69), 71-81
- Pérez, D., Gliozzo, A., Strapparava, C., Alfonseca, E., Rodríguez, P. and Magnini, B. 2005. About the effects of combining Latent Semantic Analysis with natural language processing techniques for free-text assessment. *Revista Signos. Estudios de Lingüística*, 38(59), 325-343.
- Sanders, T. and Pander, M. 2006. Cohesion and coherence: Linguistic Approaches. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (pp. 591-595). Amsterdam: Elsevier
- Sanders, T. and Spooren, W. 2009. Causal categories in discourse. Converging evidence from language use. In T. Sanders and E. Sweetser (Eds.), *Linguistic categories of causality in discourse* (pp. 197-225). Berlin: Mouton de Gruyter.
- Venegas, R. 2007. Using Latent Semantic Analysis in a Spanish research article corpus. In G. Parodi (Ed.), *Working with Spanish corpora* (pp. 195-216). Londres: Continuum.
- Venegas, R. 2009. Towards a method for assessing summaries in Spanish using LSA. In H. Chad Lane, and Hans W. (Eds.), Proceedings, of the twenty-Second International Florida Artificial Intelligence Research Society Conference (pp.113-115). Washington, DC: AAAI
- Venegas, R. (2011). Evaluation of Narrative and Expository Text Summaries Using Latent Semantic Analysis. In P. McCarthy and Ch. Boonthum-Denecke (Eds.) *Applied Natural Language Processing: Identification, Investigation and Resolutions* (531-544). Hershey: IGI Global.