Revista Latinoamericana de Psicología

www.editorial.konradlorenz.edu.co/rlp

ORIGINAL

A Text Mining Approach to the Text Difficulty of Latin American Peace Agreement

Juan C. Correa[a,*], María del Pilar García-Chitiva[b], Gustavo R. García-Vargas[a]

[a] *Facultad de Psicología Fundación Universitaria Konrad Lorenz, Colombia*
[b] *Facultad de Educación Universidad Pedagógica Nacional, Colombia*

Received 13 July 2017; accepted 19 January 2018

KEYWORDS
Colombian peace agreement, SMOG formula, Latin American peace accords, text linguistic difficulty, text mining, polysyllables

Abstract A peace agreement was recently subjected to a plebiscite as a solution to finish the Colombian armed conflict. With 62.57% of abstention, 18.44% of the Colombian electorate rejected this agreement. This paper aims to propose a methodological approach that shows how to linguistically analyze peace agreements as political products that are acceptable or not according to their text difficulty. Given the socio-political similarities of the armed conflicts of Colombia, Guatemala, and El Salvador, we scrutinized with sufficient computational detail these peace agreements. The results revealed that the text difficulty of these accords was more appropriate for a person with at least 19 years of education, suggesting that these sort of texts are not written for broader and less-educated audiences.

PALABRAS CLAVE
Acuerdo de paz en Colombia, Fórmula SMOG, Acuerdos de paz en Latinoamerica, Dificultad linguística del texto, minería de texto, polisílabas

"Una aproximación por minería de texto a la dificultad textual de los acuerdos de paz en Latinoamerica"

Resumen El gobierno de Colombia sometió recientemente a plebiscito un acuerdo de paz para dar por finalizado su conflicto armado. Para este plebiscito la abstención electoral fue del 62.57% y solo el 18.44% de los votantes rechazó la implementación del acuerdo. El objetivo de este trabajo es proponer una aproximación metodológica que muestre cómo analizar los acuerdos de paz como productos políticos que pueden ser aceptables o no según su dificultad textual. Dadas las semejanzas socio-políticas de los conflictos armados en Colombia, Guatemala y El Salvador, en este trabajo se muestra con suficiente detalle computacional el análisis por minería de texto de los acuerdos de paz celebrados en estos países. Los resultados revelaron que la dificultad textual de estos acuerdos exige un nivel educativo de

* Autor para correspondencia.
Correo electrónico: juanc.correan@konradlorenz.edu.co

http://dx.doi.org/10.14349/rlp.2018.v50.n1.6

al menos 19 años de educación formal, lo que sugiere que ese tipo de documentos no suelen redactarse para audiencias más numerosas con menores niveles de educación formal.

The Colombian contemporary history shows a series of violent events that have perpetuated armed conflicts among different groups of the society (McDougall, 2009). This conflict produced 220,000 casualties and six millions of forced displaced civilians (Centro Nacional de Memoria Histórica, 2013; Shultz et al., 2014). In 52 years several attempts were conducted for agreeing upon the terms of a peace process (McDonald, 1997). Due to this history some scholars claim that the Colombian conflict is among the longest ones worldwide (Maldonado, 2016). During September 2012 Juan Manuel Santos, the current Colombian President and Peace Nobel laureate, announced an agreement to start formal negotiations with the guerrilla movement involved in the conflict since 1964, *Las Fuerzas Armadas Revolucionarias de Colombia Ejército del Pueblo* (FARC-EP).

The publication of a final signed agreement took place on August 25, 2016. Colombians had the chance to know the terms of this agreement before deciding to support or reject its implementation in a referendum held on October 2, 2016. The political abstention in this plebiscite was 62.57% of the electorate, one of the greatest compared with previous elections (Correa & Camargo, 2017). The proposed implementation was officially rejected by 18.44% of the electorate. In fact, the votes rejecting the accord surpassed those supporting it by only 55,651, out of 34,899,945 registered voters.

Multiple reasons might explain these results and a method that unveils them might be of heuristic value. Of all registered conflicts by the Uppsala Conflict Data Program, those occurred in El Salvador and Guatemala are convenient for the sake of comparison. They are classified as three Latin American Intra-State armed conflicts that also involved active negotiations between their governments and guerrilla movements (Fisas, 2010). With the exception of El Salvador, the implementation of both the Colombian and the Guatemalan peace agreements were rejected in democratic plebiscites. By focusing on the peace deals of these nations, the distinctive ingredient of our approach is that it considers them as a sensitive unit of analysis which is neglected by previous research that relies on macroscopic data like the type of the conflict, the number of casualties, or their chronology as the usual reports of the Uppsala Conflict Data Program (Harbom, Högbladh, & Wallensteen, 2006; Kreutz, 2010) or the Ethnic Power Relations dataset (Wimmer, Cederman, & Min, 2009). Thus, the motivation of this paper consists of showing a complementary approach to analyze written peace agreements of Intra-State armed conflicts. Neither does this approach ignores the political or historical effects on the evolution of these conflicts (Matanock & García-Sánchez, 2017), nor the mass media effects to solve them (Serrano, 2015). However, as these effects have already been analyzed, we do not focus ourselves on these endeavors.

Armed conflicts are connected to meaning and knowledge which are rooted in culture. Such relationships are evident by understanding enemies' motivations for attacking objects representing cultural heritage of their societies (Brosché, Legnér, Kreutz, & Ijla, 2016). Furthermore, "peace is the product of cultural practices maintained by the people in a verbal community" (Sacipa, Ballesteros, Cardozo, Novoa, & Tovar, 2006, p. 160). These statements show the crucial role of language for understanding armed conflicts. Thus, cultural practices of nations are susceptible of linguistic analyses. The most common form of analysis is to focus on text contents. An alternative analysis relies on the readability of peace agreements that quantifies nations' resolutions to finish internal armed conflicts while conveys a meaningful interpretation. Our proposal belongs to this second analytic category.

Existing research of armed conflicts in Latin America have noticed the relevance of analyzing textual materials resulting from negotiations between the actors of the conflicts. The typical approach relies on content analysis with qualitative results. Social psychologists (Borja-Orozco, Barreto, Sabucedo, & López, 2009), linguistists (Rosell, 2009; Rodríguez-Rodríguez, 2011), rhetoric philosophers (Olave, 2014), mass media communicators (Gómez-Giraldo, 2014) and political scientists (Call, 2003; Pérez, 2003) share this approach to understand some aspects of Latin American conflicts and their resolution. However, such an approach lacks the notion of "reproducibility". Reproducibility stands for the possibility that any independent scientist can obtain thesame results that are published in scientific journals, by applying the same (statistical) procedures to the same datasets analyzed in published papers (Mair, 2016). This sort of research, known as "reproducible research" (Gandrud, 2015), is well-suited for cultural studies that focus on quantitative analyses of texts and use R as their analytic platform (Mizumoto & Plonsky, 2015). Based on these ideas, the contribution of our work is twofold. On one hand, this is the first work that illustrates with sufficient computational detail a text mining method to scrutinize Latin American peace accords in terms of their ease of comprehension by the ordinary citizen. On the other hand, it provides the theoretical basis by which this method can be used to understand text difficulty of peace agreements.

Accordingly, we organize the rest of this paper as follows. In the next section, we provide a contextual description of relevant Latin American peace agreements. Then, we illustrate the theoretical and methodological connections between Latin American peace agreements and their analyses through text mining techniques. We proceed by describing our method and the results. We conclude by discussing the implications of our approach for further studies.

A Contextual Description of Selected Latin American Peace Agreements

The United Nations "Peace Agreements Database" presents a total of 104 documents that can be understood as peace accords resulting from Intra-State conflicts in the Americas. With

the exception of Canada, 102 manuscripts correspond to nine Latin American countries. Among these nations, only three of them have gone through internal armed conflicts involving fights between guerrilla movements and local governments and have been solved with signed agreements by both parties. These countries are Colombia (1964-2016), Guatemala (1960-1996) and El Salvador (1980-1994). A common motivation that promoted these conflicts was the failure of the homeland security to provide protection to civilians who were chronic victims of crime and social inequalities (Call, 2003; Pérez, 2003). Under these conditions, several groups emerged as adversaries: disarmed civilians, paramilitary groups, guerrilla armies and the military forces of official governments. These dynamics have also been present in Colombia, where indicators of ideology-discourse, attitudes, and emotional responses to social inequalities, all played significant roles for conforming armed groups (Ugarriza & Craig, 2013). For example, Villegas de Posada (2009) have noticed that people decided to enlist or join armed groups for fun and adventure, economic safety, retaliation, and promises, while demobilization was more associated with survival, physical-psychological safety, civilian safety, justice,self-determination, and belongingness.

These socio-political similarities go along with a prevalent cultural feature that assigns an influential role of language use on individuals' behavior. One example is the well-known "culture-of-peace" in the Guatemalan case. According to Oglesby (2007) the culture-of-peace narrative promoted by UNESCO was used because of its general appealing to the population, but it was also intended to re-express the idea of reconciliation which was found controversial in UNESCO's surveys. Other analyses also emphasize the role of language in armed conflicts. In the Colombian case, Olave (2014) proposed the utility of rhetoric to understand the conflict by splitting adversaries' speeches into three domains; the narrator, the audience and the language. From the narrator domain, it was observed that the counterparts of the conflict tried to legitimate their own speeches through the use of subjective mechanisms like appealing to moral standards followed by the members of its own group; a fact that can be understood as ideological consumerism, since "its linguistic manifestation goes along with consistent behaviors that allow the identification of group's beliefs in the real-world" (Correa & Camargo, 2017, p.37). From the audience domain, the counterparts of the conflict attacked their adversaries, judged their actions and blamed them for any failing in solving the conflict. And from the language domain, the discussions moved beyond the theoretical limits of legitimacy and legality for examining the relationship between government, insurgents and civilians, given the complicated relationships between the actors of the conflict and issues like social consensus and recognition.

In a similar vein, the role of communication in mass media has suggested that in Colombia there was no connection between citizens' aspirations and the agenda of conflict negotiators. While the latter were discussing the terms for incorporating the guerrilla leaders into local politics, ordinary citizens were against this negotiation (Gómez-Giraldo, 2014). Not least important is the psychological perspective that highlights how counterparts of the conflict interact to legitimate the conflict itself (Borja-Orozco, Barreto, Alzate, Sabucedo, & López, 2009; Borja-Orozco et al., 2009; Rico, Alzate & Sabucedo, 2017). Usually, these interactions can be described as debates involving interlocutors with ideologically opposing views, that tend to become flexible as time goes by and economic resources became scarce to sustain armed conflicts. Once these attitudes became flexible or less radical, peace accords like those of Guatemala and El Salvador could finally took place (Fisas, 2010). We posit the hypothesis that the characteristics of these conflicts are linguistically reflected in their peace agreements. The historic, political and socio-economic particularities of these countries play no role in determining the text difficulty of these sort of texts. As these documents describe significant changes for developing a peaceful society, they are a valid object of study for social and political scientists that use computerized methods.

Scrutinizing Peace Accords through Text Mining Techniques: A Theoretical Perspective

The term "text mining" refers to an interdisciplinary field of research where data mining, linguistics, computational statistics, and computer science converge with different conceptual and methodological approaches to extract useful information from document collections. The general idea behind the use of text mining consists of transforming any written text into a structured format based on term frequencies and subsequently apply standard data mining techniques, such as "bag-of-words", text clustering and/or text classification (Feldman & Sanger, 2007). This logic can be applied to analyze political texts in general (Lucas et al., 2015). Our perspective shows the theoretical relevance of text mining techniques to evaluate the readability of peace agreements as political products of nations committed to stopping their own armed conflicts.

The process of transforming the text into a structured format begins with reformatting the text itself (this is also known as preprocessing in computer sciences). The goal is to leave the text with only relevant semantic information. This step typically requires removing all extra white spaces, page numbers and "stopwords". Stopwords are words that are so common in a language (e.g., articles, conjunctions, prepositions, etc.) that their semantic information value is almost zero (Feldman & Sanger, 2007). In preprocessing, it is also important to use the stemming procedure. This consists of erasing word suffixes to retrieve their radicals; thus, the Spanish words *gobierno*, *gobiernos*, *gobiernan* and *gobiernas* all become *gobiern*. This is quite common as it reduces complexity without any severe loss of information for standard applications. The resulting text after the preprocessing stage is also known as a "corpus" and this is usually decomposed in its staple linguistic components (or tokens) allowing the analysis by means of term frequencies, correlations between terms and so on.

Among the available methods for text analysis, the so-called "readability measures" are particularly useful. For example, the so-called Flesch-Kincaid grade level formula is well-known as it is included in commercial word-processing packages, despite its limited accuracy (Fitzsimmons, Michael, Hulley, & Scot, 2010). Interested readers in a brief history of these measures can consult Contreras, García-Alonso, Echenique, and Daye-Contreras (1999) or Benjamin (2012). Psychologists and educators were the main developers of these metrics since 1930. However, major developments

took place around World War II when Dale and Chall (1948) found that the words used in a text and the average sentence length were two major causes of reading difficulty. Based on these findings, Fry (1968) developed a formula for predicting text difficulty with the claim of saving time and then Mc Laughlin (1969) exploited these efforts to propose the well-known "SMOG grading" readability formula:

$$SMOG = 1.043\sqrt{(30 - S)} \times \frac{P}{S} + P + 3.1291 \qquad (1)$$

This metric indicates "the reading grade that a person must have reached if he is to understand fully the text assessed" (Mc Laughlin, 1969, p. 639). The formula considers two basic linguistic elements: P , the number of polysyllables (i.e., words of three or more syllables) and S, the average sentences length of the text (i.e., any string of words ending with a period, question mark or exclamation point). Although the SMOG formula was originally considered as a proxy of text complexity (Mc Laughlin, 1969), recent evidence shows that text complexity might be better regarded as a balance between the "disorder" of its information (i.e., the probability of word appearance), quantified by a metric known as "emergency", and the order conveyed in a text, quantified by entropy (Febres, Jaffé, & Gershenson, 2015). According to Febres and Jaffé (2017), these new complexity measures are not related to traditional readability measures of the text, but rather to literary quality, suggesting that readability is not necessarily related to complexity (nor, maybe, to literary quality).

The scientific foundation of the SMOG formula is based on the obvious fact that the use of words in written language is richer in adults than in children; that is, the higher the schooling years of a person the wider his written language; quantified by a metric called lexical diversity (Durán, Malvern, Richards, & Chipere, 2004). Lexical diversity is traditionally calculated through the "Type-Token Ratio"; that is, the number of different words (types) divided by the total number of words (tokens) in a document. Since the Type-Token Ratio varies in accordance with the length of the text, better metrics exist, being the SMOG formula particularly convenient for Spanish texts (Contreras et al., 1999) which have been honestly written for its freely reading (Bruce, Rubin, & Starr, 1981). Even though the SMOG formula has been applied to both English and Spanish texts (Berland et al., 2001), its application to peace agreements remains unknown to us.

The calculus of the SMOG formula is quite straightforward due to recent software developments (Meyer, Hornik, & Feinerer, 2008; Michalke, 2015; Rinker, 2013). The relevance of the SMOG formula for political communication is twofold. On one hand, it permits to evaluate the readability of peace accords by inhabitants of nations with internal armed conflicts. On the other hand, it features a procedure that can be used to evaluate the adequacy of political manuscripts ruling foreseen changes for peace-building purposes.

Materials and Methods

Our analyses were based on the official peace agreements signed by the governments of Colombia, Guatemala and El Salvador and their corresponding guerrilla movements. These accords were officially written in Spanish and so were their analyses. Data collection was as follows. All agreements, except the Colombian one, were downloaded from the "United Nations Peacemaker Database" (UNPD) which is freely available at http://peacemaker.un.org/document-search. The Colombian agreement that we used was the one subjected to plebiscite (signed in August 25, 2016) that can be found at https://www.mesadeconversaciones.com.co/sites/default/files/acuerdo-final-1473286288.pdf.

The preprocessing of these documents involved their conversion from pdf files to UTF-8 text-formatted data without headers and footers since these items present no semantic relevant information. Additional non-semantic features like badges, scanned signatures, vignettes, tables and the like were all deleted. In addition, we removed all numbers, unnecessary whitespaces, punctuation and special characters as well as Spanish stopwords. All these tasks were conducted with the aid of the "tm" package which was developed for conducting text mining tasks inside the R environment (Meyer et al., 2008). We tokenized the contents of the manuscripts with the in-built koRpus tokenizer for Spanish (Michalke, 2015), the detection and frequency analysis of polysyllables was conducted with the aid of the "qdap" package (Rinker, 2013). The computation of the SMOG formula was performed with the in-built koRpus algorithm since this performs the operations on the tokenized corpus rather than using the direct counting employed by the in-built qdap algorithm.

The resulting corpus for each of these documents varied depending on the number of sub-documents that composed these accords. In the case of the Colombian peace agreement the whole manuscript was split into seven independent sub-documents. Each point of the agreement was treated as a single document and the seventh one contains the set of protocols and recommendations that were specified as addendum of the agreement. The official agreement for El Salvador was split into 12 sub-documents: a document containing previous acts in New York city, the introduction, its nine agreement points, and the final statement. The peace agreement of Guatemala was split into 13 sub-documents: 12 chapters and a protocol. Thus, our text sample was composed by32 Spanish-written documents. The frequency of appearance for every single word was assembled in a term-document matrix, where words are arranged as rows and the 32 documents of our sample are arranged as columns. The most frequent shared words across nations' peace agreements are those with higher frequencies in all or almost all documents. As a supplementary material we have developed an easy-to-use R script that allows the exact reproduction of these procedures and its corresponding results. All preprocessed peace agreements, as well as their structured dataframes are available as supplementary materials. These steps are actually documented in the following protocol https://www.protocols.io/view/the-colombian-signed-peace-agreement-a-text-mining- h8db9s6

Results

We began our analysis by calculating the SMOG index for each component document of the peace accords for Colombia, Guatemala and Salvador. Table 1 shows that, regardless their length, these agreements are linguistically consistent,

reflecting the fact that they are similar products in terms of their readability.

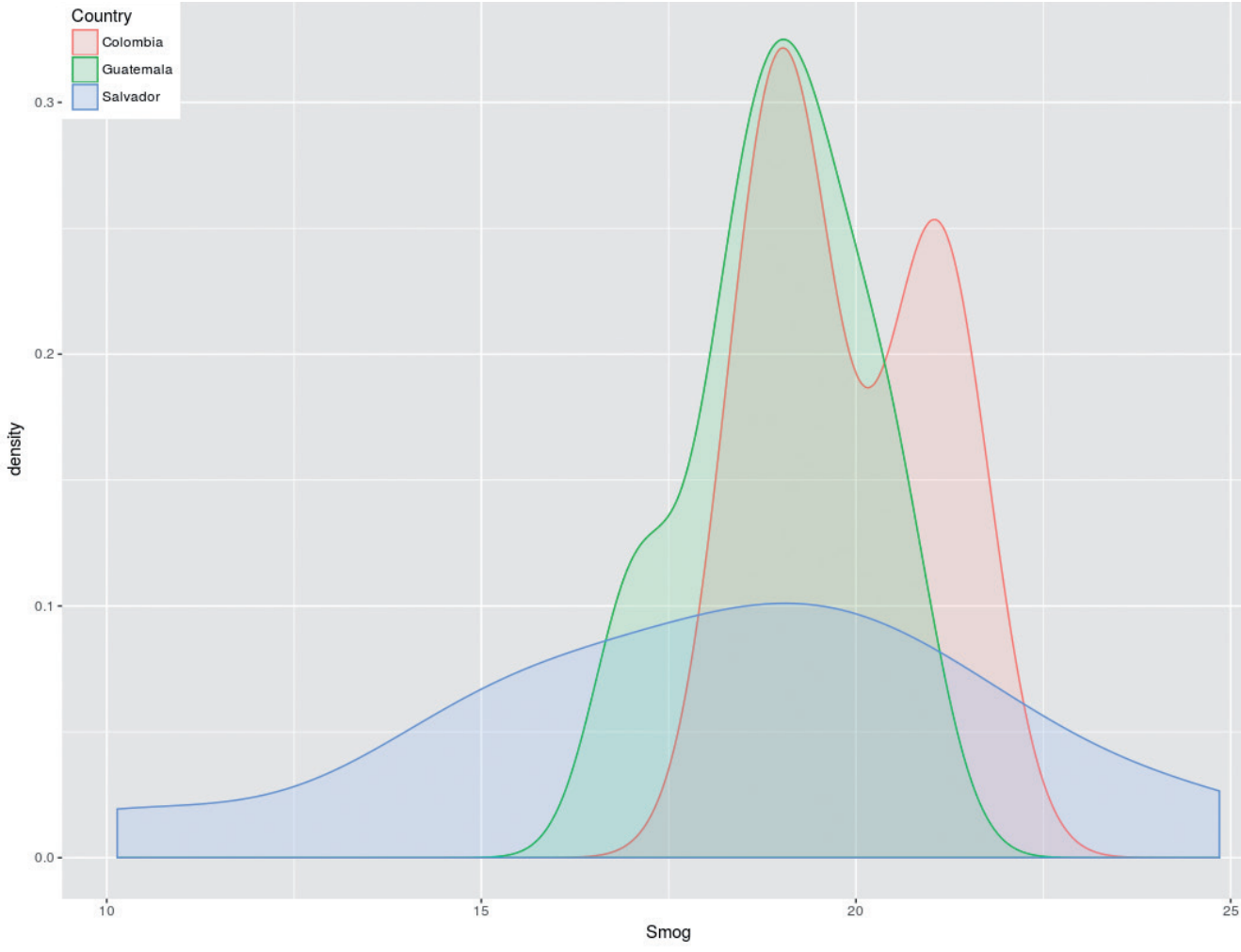
Figure 1 shows the statistical distribution of the readability for the Latin American peace agreements. The shape of these distributions as well as their location on the abscissa reveal interesting patterns. Their peaks show that the average readability of these accords is around 19 schooling years. Put it in terms of an educational profile, these distributions reveal that the content of these accords could be fully understood by a person who began studying at 5 years old and never stopped until completing a master degree. The width of these distributions reveals the variance of the readability for these accords. Both the Colombian and the Guatemalan accords show similar variances (i.e., similar readabilities), which contrasts with the deal for El Salvador whose readability is more heterogeneous (some parts of the documents are highly difficult to understand while other parts are easier). The distribution of the readability for the Colombian accord is more located to the right, which indicates that this accord is more difficult to understand compared with the other manuscripts. These differences, proved to be not significant (F = 1.046; df = 2; p = .364); a fact that reveals the socio-cultural similarities among the nations that produced these documents to formalize the end of their conflicts.

Table 1 Descriptive Statistics of Latin American Peace Accords

Country	Documents	Smog		Age	
		Mean	SD	Mean	SD
Colombia	7	19.87	1.12	24.89	1.12
Guatemala	13	18.98	1.14	23.98	1.14
El Salvador	12	18.15	3.87	23.15	3.87
Total	32	18.87	2.55	23.87	2.55

A closer scrutiny to the average word length of these texts also reveals interesting patterns: 78% of the Colombian and Guatemalan agreements was composed by polysyllables. The Salvadoran text, in contrast, was simpler with 44% of polysyllables. The richest text was the Colombian one with a total of 6,742 types, followed by the Guatemalan document with 4,655 types and the Salvadoran manuscript with 2,622 different words. The linguistic composition of these texts was far above everyday texts such as news, propaganda, advertisement messages or songs that have shown an average of 661 words (Ramírez-Esparza, Pennebaker, García, & Suriá, 2007). As a matter of fact, these Latin American peace deals show a linguistic composition which

Figure 1. Statistical distribution of the readability for Latin American peace agreements.



is as similar as the novels of "*Pedro Páramo*" (with 4,820 different words) produced by the Mexican writer Juan Rulfo or the universal masterpiece "*Don Quixote of the Mancha*" (with 6,684 different words) conceived by the Spaniard Miguel de Cervantes Saavedra (Sánchez & Cantos, 1997).

Figure 2 depicts a wordcloud with a subset of the most frequent shared words among the three Latin American peace agreements. Words with largest font sizes are those more frequently used throughout the documents. The most frequent word was "will" (1,693) as the auxiliary verb used to express desire, choice, willingness or consent. The second most common word was "national" (902) which was followed by "agreement" (878), "government" (715), "rigths" (606), "peace" (594), "shall" (484), "social" (467), "implementation" (442), "development" (422), "political" (415), "public" (412), "participation" (405), "special" (387), "farcep"(370) and so on. This wordcloud captures the sensitive topics associated with peace-building processes with ideas of reconciliation (i.e. agreement, peace, rights, participation), ideas of security (i.e., security, forces, court, state, crimes, law, conflict) and economic stability (i.e., development, ensure, land, health). The emphasis of these ideas can be grasped by seeing their frequency of appearance throughout the texts. Excepting the word "farcep" which is the acronym for the Colombian guerrilla movement, the rest of the words in this wordcloud show that the emphasis of these ideas is consistent across nations.

Figure 3 shows in descending order of appearance the list of the most frequent words that corresponds to the upper 1 percentile of Latin American Peace Agreements treated as a large corpus. This list is of special interest for adapting the readability of the agreements to specific audiences, as it conveys which polysyllabic words could be replaced for other shortest words semantically similar. For example, the polysyllabic word "agreement" could be replaced by the word "deal"; likewise the word "implementation" could be replaced by "fulfillment" and so on. With lists of this sort, politicians could optimize the comprehension of their written texts by carefully shrinking the average length of words; an action that would avoid possible communication biases favoring elite groups of the society with higher levels of education.

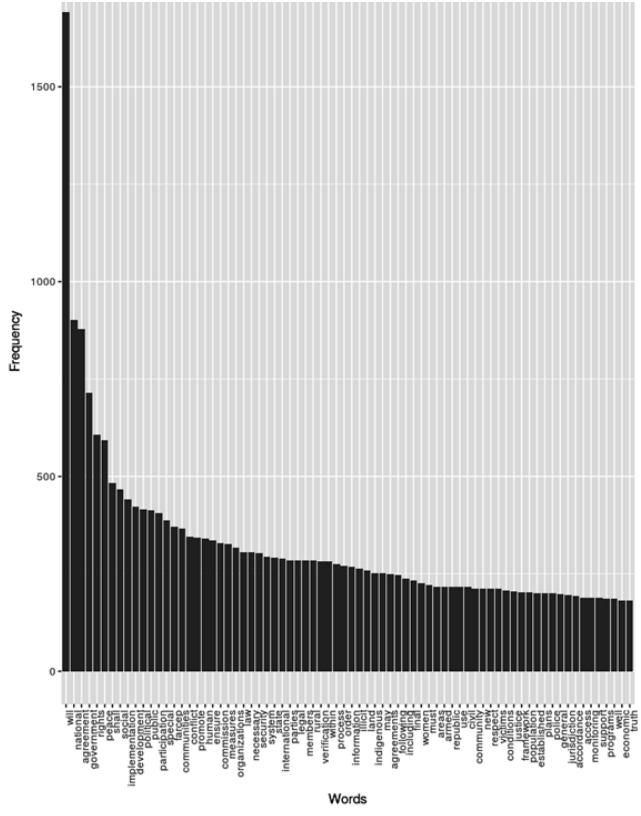
Figure 2. Wordclouds for Latin American Peace Agreements.

Discussion

We posited the idea that peace agreements resulting from Intra-State armed conflicts can be understood as political products that are susceptible of linguistic scrutiny. As such, these texts represent a valid object of study in political communication focusing on understanding conflict and peace-building processes (Goodhand & Hulme, 1999; Call,2003). Since these processes are conducted by individuals and groups who share similar attitudes to promote peace (Harbom et al., 2006; Joshi, Melander, & Quinn, 2015), our aim was to develop a theoretical approach that showed how nations' peace agreements can be linguistically analyzed. An essential element of our approach is the contextual description of nations. This description serves as a criterion to select different countries with similar socio-political and cultural characteristics. We posited that these similarities are reflected in the Latin American peace agreements.

The theoretical value of analyzing the readability of peace agreements through the SMOG formula (Mc Laughlin, 1969) was the second element of our approach. Readability plays a crucial role for clarifying the scope of these texts

Figure 3. List of most frequent words for Latin American peace agreements corresponding to the upper 1.5 percentile



to ordinary citizens who are not familiarized with this sort of linguistic style (Charrow & Charrow, 1979). Such a role is even more relevant when the contents of these accords are subjected to plebiscites (such as the Colombian and the Guatemalan case). As pointed out by Beltrán-Oicatá and Sandoval-Escobar (2015) elections can be understood as decision-making processes by which individuals make judgments about political options (i.e., to support or reject the implementation of peace accords). If the contents of these agreements are not understandable by the electorate who is called to participate in a plebiscite, any attempt aimed to implement the contents of these accords is doomed to fail.

Our approach can be used as a guidance for political advising. Politicians might be benefited from readability analyses in two fronts. Firstly, They can easily identify which part of their documents are more difficult to understand. Secondly, they could adapt the contents of the documents for a general audience. By following this strategy, politicians might increase the odds of political support in terms of electoral results. This implication is in accordance with previous attempts aimed to making legal language understandable (Charrow & Charrow, 1979) and can be applied by social scientists interested in peace-building processes.

In scrutinizing Latin American peace agreements we showed that these manuscripts were not accessible to the general population. Their linguistic composition proved to be more appropriate for a person with at least 18 years of formal education. This result contrasts with the 4.6 years of Salvadorean formal education in 1995 (Rivas-Villatoro, 2013); the 5.6 years of formal education for the Guatemalan population in 1985 (Edwards, 2002) or the Colombian average educational attainment in 2015 which was between 9.9 and 11 years of formal education for half of the Colombian population (Perfetti del Corral, Prada-Lombo, & Freire-Delgado, 2016). It is important to recall that the Colombian electorate had only 38 days between the moment when the agreement was signed and the call for participation in the plebiscite to decide whether or not to support the proposed implementation of the peace agreement. The results of the Colombian plebiscite could have been largely different if parties of the conflict had conducted pilot tests to divulge adapted versions of the agreement to a less educated audience, diminishing in this way possible biases of searching democratic support in elite groups of the society who tend to be better educated (Matanock & García-Sánchez, 2017). This result is missing in more recent reviews addressing the positive lessons that the Colombian peace process has shown for other conflicts in the world (Maldonado, 2017).

Our study might be also understood as a relevant case for the automatic analysis of political texts (Grimmer & Stewart, 2013); a promising methodology that can be incorporated into the research methods toolbox of political analysts. We have shown how recent technology can be used in a straightforward manner to explore the readability of peace agreements. The R script that we have developed as supplementary material for this paper was conceived as an example that can be followed by other scientists interested in exploring the documents of Colombia, Guatemala and El Salvador where the need to understand the social mechanisms leading to permanent violence has been commonly highlighted (McDougall, 2009; Preti, 2002; Bourgois, 2001). Given the fact that peace agreements often include sensitive issues for reestablishing justice and governance (Wagner & Druckman, 2016), text mining techniques in general and readability evaluations in particular, are useful for assessing the difficulty of these texts and clarify the order of peace implementation that prove to be a key ingredient for reducing destabilizing effects of post-accord elections (Joshi et al., 2015).

A final implication relates to the possibility of analyzing peace agreements of other Non-Spanish-Spoken cultures, as well as to failed peace accords elsewhere. In particular, our script can be adapted for English-spoken countries that have gone through peace processes such as Northern Ireland (1987-2008), South Africa (1989-1994), South Sudan (1998-2005) and Sierra Leone (1994-2002). This script can also be adapted for French-spoken countries such as Burundi (1998-2998), Mali (2012 - Present) and Central African Republic (2012 - Present). Yet, our method might be enhanced by linguists and computer scientists committed to the development of other computer algorithms targeting languages of other nations who have also went through peace processes such as Angola (1988-2002), Tajikistan (1992-1997), Indonesia (2000-2005), Nepal (2002-2006) or Sri Lanka (1983 - 2009).

References

Beltrán-Oicatá, C., & Sandoval-Escobar, M. (2015). Efecto de la capacitación y el diseño del tarjetón sobre la comprensión y la validez del voto. *Universitas Psychologica*,*14* (3), 1067-1076. http://dx.doi.org/10.11144/Javeriana.upsy14-3.ecdt

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24* (1), 63–88. http://dx.doi.org/10.1007/s10648-011-9181-8

Berland, G. K., Elliott, M. N., Morales, L. S., Algazy, J. I., Kravitz, R. L., Broder, M. S., . . . McGlynn, E. A. (2001). Health information on the internet: Accessibility, quality, and readability in English and Spanish. *Journal of the American Medical Association*,*285*(20), 2612–2621.

Borja-Orozco, H., Barreto, I., Alzate, M., Sabucedo, J. M., & López, W. (2009). Creencias sobre el adversario, violencia política y procesos de paz. *Psicothema*, *21*(4), 622–627.

Borja-Orozco, H., Barreto, I., Sabucedo, J. M., & López, W. (2009). Construcción del discurso deslegitimador del adversario: Gobierno y paramilitarismo en Colombia. *Universitas Psychologica*, *7*(2), 571–584.

Bourgois, P. (2001). The power of violence in war and peace: Post-cold war lessons from El Salvador. *Ethnography*, *2*(1), 5–34. http://dx.doi.org/10.1177/14661380122230803

Brosché, J., Legnér, M., Kreutz, J., & Ijla, A. (2016). Heritage under attack: Motives for targeting cultural property during armed conflict. *International Journal of Heritage Studies*, 1-13. http://dx.doi.org/10.1080/13527258.2016.1261918

Bruce, B., Rubin, A., & Starr, K. (1981). Why readability formulas fail. *IEEE Transactions on Professional Communication*, *PC-24* (1), 50–52. http://dx.doi.org/10.1109/TPC.1981.6447826

Call, C. T. (2003). Democratisation, war and state-building: Constructing the rule of law in El Salvador. *Journal of Latin American Studies*, *35*(04), 827–862. http://dx.doi.org/10.1017/S0022216X03007004

Centro Nacional de Memoria Histórica. (2013). *¡Basta ya! Colombia: Memorias de guerra y dignididad*. Bogotá, Colombia: Imprenta Nacional.

Charrow, R. P., & Charrow, V. R. (1979). Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia Law Review*, *79*(7), 1306–1374. http://dx.doi.org/10.2307/1121842

Contreras, A., García-Alonso, R., Echenique, M., & Daye-Contreras, F. (1999). The SOL Formulas for converting SMOG readability scores between health education materials written in Spanish, English, and French. *Journal of Health Communication*, *4*(1), 21–29. http://dx.doi.org/10.1080/108107399127066

Correa, J. C., & Camargo, J. E. (2017). Ideological consumerism in Colombian elections,2015: Links between political ideology, Twitter activity, and electoral results.*Cyberpsychology, Behavior and Social Networking*, *20*(1), 37-43. http://dx.doi.org/10.1089/cyber.2016.0402

Dale, E., & Chall, J. S. (1948). A formula for Predicting readability: Instructions.*Educational Research Bulletin*, 37–54.

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, *25*(2), 220–242. http://dx.doi.org/10.1093/applin/25.2.220

Edwards, J. (2002). *Education and poverty in Guatemala* (Tech. Rep.). Department of Economics, University of Tulane. Retrieved from http://datatopics.worldbank.org/hnp/files/edstats/GTMwp02.pdf

Febres, G., & Jaffé, K. (2017). Quantifying structure differences in literature using symbolic diversity and entropy criteria. *Journal of Quantitative Linguistics*, *24*(1), 16–53. http://dx.doi.org/10.1080/09296174.2016.1169847

Febres, G., Jaffé, K., & Gershenson, C. (2015). Complexity measurement of natural and artificial languages. *Complexity*, *20*(6), 25–48. http://dx.doi.org/10.1002/cplx.21529

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Fisas, V. (2010). Procesos de Paz Comparados. *Quaderns de Construcció de Pau, Catalunia, 14* , 8–12.

Fitzsimmons, P., Michael, B., Hulley, J., & Scot, G. (2010). A readability assessment of online Parkinson's disease information. *The Journal of the Royal College of Physicians of Edinbourgh*, *40*(4), 292-296.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading*, *11*(7), 513–578.

Gandrud, C. (2015). *Reproducible Research with R and Rstudio* (2nd ed.). Boca Raton, FL: CRC.

Gómez-Giraldo, J. C. (2014). Los discursos en el proceso de paz en Colombia: Un análisis de la capacidad de los negociadores de permear a las audiencias. In *III Congreso Internacional de la Asociación Latinoamericana de Investigadores en Campañas Electorales*. Chía - Colombia. Retrieved from http://www.alice-comunicacionpolitica.com/abrir-ponencia.php?f=627-F54171b9a6271410

Goodhand, J., & Hulme, D. (1999). From wars to complex political emergencies: Understanding conflict and peace-building in the New World disorder. *Third World Quarterly*, *20*(1), 13-26. http://dx.doi.org/10.1080/01436599913893

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 267-297. http://dx.doi.org/10.1093/pan/mps028

Harbom, L., Högbladh, S., & Wallensteen, P. (2006). Armed conflict and peace agreements. *Journal of Peace Research*, *43*(5), 617–631. http://dx.doi.org/10.1177/0022343306067613

Joshi, M., Melander, E., & Quinn, J. M. (2015). Sequencing the peace: How the order of peace agreement implementation can reduce the destabilizing effects of post-accord elections. *Journal of Conflict Resolution*, *61*(1), 4–28. http://dx.doi.org/10.1177/0022002715576573

Kreutz, J. (2010). How and when armed conflicts end: Introducing the UCDP conflict termination dataset. *Journal of Peace Research*, *47*(2), 243–250. http://dx.doi.org/10.1177/0022343309353108

Lucas, C., Nielsen, R., Roberts, M., Stewart, B., Storer, A., & Tingley, D. (2015). Computer assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254-277. http://dx.doi.org/10.1093/pan/mpu019

Mair, P. (2016). Thou shalt be reproducible! A technology perspective. *Frontiers in Psychology*, *7*, 1079. http://dx.doi.org/10.3389/fpsyg.2016.01079

Maldonado, A. (2016). *Early lessons from the Colombian peace process*. The London School of Economics and Political Science Global South Unit. Retrieved from http://eprints.lse.ac.uk/65606/

Maldonado, A. (2017). What is the Colombian peace process teaching the world? *New England Journal of Public Policy*, *29*(1), 1-8.

Matanock, A. M., & García-Sánchez, M. (2017). The Colombian paradox: Peace processes, elite divisions and popular plebiscites. *Daedalus*, *146* (4), 152-166. http://dx.doi.org/10.1162/DAED_a_00466

McDonald, G. (1997). *Peacebuilding from below: Alternative perspectives on Colombia's peace process*. London: Catholic Institute for International Relations.

McDougall, A. (2009). State power and its implications for civil war Colombia. *Studies in Conflict & Terrorism*, *32*(4), 322-345. http://dx.doi.org/10.1080/10576100902743815

Mc Laughlin, G. H. (1969). SMOG grading- a new readability formula. *Journal of Reading*, *12*(8), 639–646.

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Michalke, M. (2015). *Korpus: A n R package for text analysis. [computer software]. https://cran.r-project.org/web/packages/koRpus/index.html*

Mizumoto, A., & Plonsky, L. (2015). R as a lingua franca: Advantages of using R for quantitative research in applied linguistics. *Applied Linguistics*, *37*(2), 284-291. http://dx.doi.org/10.1093/applin/amv025

Oglesby, E. (2007). Educating citizens in postwar Guatemala: Historical memory, genocide, and the culture of peace. *Radical History Review*, *2007*(97), 77–98. http://dx.doi.org/10.1215/01636545-2006-013

Olave, G. (2014). Aproximaciones retóricas al conflicto armado Colombiano: Una revisión bibliográfica. *Forma y Función*, *27*(1), 155–197.

Pérez, O. J. (2003). Democratic legitimacy and public insecurity: Crime and democracy in El Salvador and Guatemala. *Political Science Quarterly*, *118*(4), 627–644. http://dx.doi.org/10.1002/j.1538-165X.2003.tb00408.x

Perfetti del Corral, M., Prada-Lombo, C. F., & Freire-Delgado, E. E. (2016, March). *Encuesta Nacional de Calidad de Vida* (Tech. Rep. No. ECV-2015). Bogotá, Colombia: DANE. Retrieved from http://www.dane.gov.co/files/investigaciones/condiciones_vida/calidad_vida/Boletin_

Preti, A. (2002). Guatemala: Violence in peacetime: A critical analysis of the armed conflict and the peace process. *Disasters*, *26*(2), 99–119. http://dx.doi.org/10.1111/1467-7717.00195

Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., & Suriá, R. (2007). La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista Mexicana de Psicología*, *24*(1), 85–99.

Rico, D., Alzate, M. & Sabucedo, J. M. (2017). El papel de la identidad, la eficacia y las emociones positivas en las acciones colectivas de resistencia pacífica en contextos violentos. *Revista Latinoamericana de Psicología, 49*(1), 28-35. http://dx.doi.org/10.1016/j.rlp.2015.09.013

Rinker, T. W. (2013). qdap: Quantitative discourse analysis package [Computer software manual]. Buffalo, NY: Retrieved from http://github.com/trinker/qdap ((Computer Software, Version 2.2.5))

Rivas-Villatoro, F. (2013). *El financiamiento de la educación en El Salvador* (1st ed.). San Salvador: Ediciones Centroamericanas. Retrieved from https://www.unicef.org/elsalvador/El_financiamiento_de_la_Educacion_en_El_Salvador.

Rodríguez-Rodríguez, C. (2011). ¿Conflicto armado interno en Colombia? Más allá de la guerra de las palabras. *Magistro*, *4*(7), 111–125.

Rosell, C. (2009). *El poder de la palabra: Un análisis de 21 extractos del discurso presidencial colombiano con respecto a las FARC* (Undergraduate thesis) Retrieved from https://lup.lub.lu.se/student-papers/search/publication/1530015

Sacipa, S., Ballesteros, B. P., Cardozo, J., Novoa, M. M., & Tovar, C. (2006). Understanding peace through the lens of Colombian youth and adults. *Peace and Conflict: Journal of Peace Psychology*, *12*(2), 157.

Sánchez, A., & Cantos, P. (1997). El ritmo incremental de palabras nuevas en los repertorios de textos. Estudio experimental y comparativo basado en dos corpus lingüísticos equivalentes de cuatro millones de palabras, de las lenguas inglesa y española y en cinco autores de ambas lenguas. *Atlantis*, 205–223.

Serrano, Y. (2015). The strategic issues of journalistic lexicon when reporting on victims of the Colombia armed conflict. *Journal of Latin American Communication Research*, *5*(1), 64–86.

Shultz, J. M., Ceballos, Á. M. G., Espinel, Z., Oliveros, S. R., Fonseca, M. F., & Florez, L. J. H. (2014). Internal displacement in Colombia: Fifteen distinguishing features. *Disaster Health*, *2*(1), 13–24. http://dx.doi.org/10.4161/dish.27885

Ugarriza, J. E., & Craig, M. J. (2013). The Relevance of ideology to contemporary armed conflicts: A quantitative analysis of former combatants in Colombia. *Journal of Conflict Resolution*, *57*(3), 445–477. http://dx.doi.org/10.1177/0022002712446131

Villegas de Posada, C. (2009). Motives for the enlistment and demobilization of illegal armed combatants in Colombia. *Peace and Conflict: Journal of Peace Psychology*, *15*(3), 263–280. http://dx.doi.org/10.1080/10781910903032609

Wagner, L., & Druckman, D. (2016). Drivers of durable peace: The role of justice in negotiating civil war termination. *Group Decision and Negotiation*, 1–23. http://dx.doi.org/10.1007/s10726-016-9511-9

Wimmer, A., Cederman, L.-E., & Min, B. (2009). Ethnic politics and armed conflict: A configurational analysis of a new global data set. *American Sociological Review*, *74*(2), 316–337. http://dx.doi.org/10.1177/000312240907400208