L E S L I

Linguistic Evidence in Security, Law and Intelligence

Volume 3 (2019) | ISSN 2327-5596 (online) | DOI 10.5195/lesli.2019.19 | http://lesli-journal.org

Developing and Analyzing a Spanish Corpus for Forensic Purposes

Ángela Almela[a], Gema Alcaraz-Mármol[b], Arancha García-Pinar[c] and Clara Pallejá[c]

a Universidad de Murcia, Spain
b Universidad de Castilla-La Mancha, Spain
d University Center for Defense-UPCT, Spain

Abstract

In this paper, the methods for developing a database of Spanish writing that can be used for forensic linguistic research are presented, including our data collection procedures. Specifically, the main instrument used for data collection has been translated into Spanish and adapted from Chaski (2001). It consists of ten tasks, by means of which the subjects are asked to write formal and informal texts about different topics. To date, 93 undergraduates from Spanish universities have already participated in the study and prisoners convicted of gender-based abuse have participated. A twofold analysis has been performed, since the data collected have been approached from a semantic and a morphosyntactic perspective. Regarding the semantic analysis, psycholinguistic categories have been used, many of them taken from the LIWC dictionary (Pennebaker et al., 2001). In order to obtain a more comprehensive depiction of the linguistic data, some other ad-hoc categories have been created, based on the corpus itself, using a double-check method for their validation so as to ensure inter-rater reliability. Furthermore, as regards morphosyntactic analysis, the natural language processing tool ALIAS TATTLER is being developed for Spanish. Results shows that is it possible to differentiate non-abusers from abusers with strong accuracy based on linguistic features.

Keywords: forensic linguistics, linguistic corpus, morphosyntactic analysis, semantics.

Introduction

This study is part of a project, which is the first study of the Spanish language associated with a specific type of criminals in Spain, by means of a natural language processing tool. In this study, the focus is on the gender-based abusers, i.e. people who have been convicted of domestic or gender-based violence. Our ultimate aim is to identify the abuser's linguistic profile. The text analysis tools used in this study are LIWC: Linguistic Inquiry and Word Count (Pennebaker, Francis and Booth 2001) and ALIAS: Automated Linguistic Identification & Assessment System (Chaski 1997, 2001, 2005).

This journal is published by the University Library System, University of Pittsburgh as part of its D-Scribe Digital Publishing Program and is cosponsored by the University of Pittsburgh Press.

In this paper, foundational information including a review of corpus linguistics and forensic linguistics is presented in Section 1, informational about our data collection, experimental subjects, and ethical issues in Section 2, and data analysis in Section 3, discussion of results in Section 4, with conclusions and future work presented in Section 5.

1. Corpus linguistics

Corpus linguistics is defined by its focus on a corpus. A corpus is a collection of texts that are used for language analysis. As Renouf noted, "[t]he term 'corpus' will be used to refer to a collection of texts, of the written or spoken word, which is stored and processed on computer for the purposes of linguistic research" (Renouf 1987:1). Among the types of corpora, we have to mention two types, those which are general (containing texts of different nature) such as the Brown Corpus or the British National corpus, and those which are specific (containing texts that belong to a particular genre). Additionally, there are other features of corpora which influence the research in corpus linguistics, e.g. multilingual (texts in different languages), parallel (same texts in two languages), learner (use of a language by non-native speakers), diachronic (comprehending a specific time span), or monitor (corpora which under continuous and ongoing construction).

Despite the fact that corpus linguistics (CL) is a well-established discipline nowadays, there are still discrepancies as regards its definition. Some scholars consider corpus linguistics as a method (Parodi 2008, Saldanha 2009). Thus, Parodi (2008) states that corpus linguistics constitutes a set of methodological principles to study any linguistic domain, and it supports research of the language in use. Sinclair (1998) affirms that with CL almost any linguistic pattern can be investigated, whether lexical, grammatical, phonological or morphological. Sinclair supports the idea that CL is of a methodological nature. Yet, some others (Teubert 2005, Tognini-Bonelli 2001, Halliday 1993) argue that the role of CL goes far beyond methodology. In this line Tognini-Bonelli (2001) argues that CL "has become a new research enterprise and a new philosophical approach to linguistic enquiry" (p.1). Leaving these controversies aside, all seem to agree that CL provides an objective view of language that overcomes the merely anecdotal or a single scholar's introspection, by the systematic examination of a group of texts (corpus) used to find the answer to a research question related to patterns of language in use.

Corpus linguistics approaches language based on two main principles: empiricism and technology. As for the former, CL is essentially empirical. This means that it explores language in context, or the performance of language, harking back to the Chomskyan distinction between performance --what people do linguistically-- and competence --what people know implicitly of language. Leech (1992) remarks that CL examines a question of performance, of *language in use*. The analysis is therefore based on observation and experience, from which theoretical principles can be formulated. In this sense, it is "bottom-up". It provides an empirical basis for teaching materials, grammars and dictionaries of a general or specific nature, whether written or oral. Regarding technology, at present day, we should not understand CL without computation. Kennedy argues that CL is "inextricably linked to the computer" (1998, p.5). In fact, Baker (2012) highlights the importance of technology as it allows testing without the influence of the researcher's preconceptions. Computation is key to CL as it contributes to a more reliable analysis. Besides, it allows the handling of huge amount of text in a quick and replicable way. Paying attention to all that has been stated above, it seems clear that CL is to be framed within an interdisciplinary and integrated view of language research. Thus, CL involves several disciplines (e.g. linguistics and computer technology) and can be applied to different linguistic aspects (e.g. any level of language analysis and any focus of language use).

1.1. Aspects of analysis and tagging in corpus linguistics

Corpora have been compiled for different purposes, which determine the corpus' nature, size and design. In a broad sense, the two main aspects that are normally analyzed within a corpus are *concordance* and *frequency*. A *concordance* is a list of all the words in a corpus, surrounded by the words which come before and after it, or the word's context. Concordance implies a qualitative approach as it explores the behavior of words in the corpus, i.e. what words a particular word co-occurs with in context. A more quantitative standpoint is found in *frequency*, where the focus is on the number of times a word or a group of words occurs in the corpus itself. Word frequency can be presented as a raw count, but this is not as useful as the "*relative frequency*" or "*normalized frequency*" of the word in relation to the entire word count of the corpus. When the relative frequency is calculated, corpora of different sizes can be compared. Thus, raw counts are hardly ever used in corpus linguistics, since texts are almost always of different sizes.

Corpus annotation --also called *tagging*-- consists of adding linguistic information to a corpus (Leech 2005). Tagging helps the researcher to retrieve information easier and faster. It provides objectivity to the analysis (McEnery 2003). Annotation also allows reusability and multifunctionality of the corpus. Once the information is tagged, it can be used for various purposes. Usually tagging schemes are explained so that every user of a corpus can understand additional information provided by the tags, beyond the particular words in a corpus.

Corpora are usually tagged or "marked up" with different sorts of information. They can be tagged at a phonological, morphosyntactic or semantic level, focusing on boundaries, suprasegmental elements in the case of phonology, part of speech and word categories in a morphosyntactic analysis or in lemmas and semantic fields at a semantic level. Corpora can be tagged in an automatic, semi-automatic or manual way. If we opt for automatic annotation, post-editing human correction is recommended as a standard industry practice (Cantos-Gomez 2013, McEnery and Hastie 2012). In order to improve accuracy of an automated approach, a semi-automatic computer-assisted approach is desirable. Manual tagging is the most accurate but it can be extremely time-consuming. The semi-automatic approach, using human experts to correct any errors of automatic annotation, is especially recommended when there is a justification based on the use for which the corpora is designed, such as forensic uses.

There are many types of taggers that can be used in automatic or semi-automatic annotation. Many of them are designed only for the English language. This is the case of SALTA or AMALGAM, QTAG or CLAWS. Others are multilingual such as Tree tagger, which presents files available for English, German, French, and Italian; TnT for German and English; or HMM-based tagger MBT, which includes Dutch, English, Spanish, Swedish and German. Some of them are free-access like SALTA, AMALGAM or Tree tagger, but CLAWS, for instance, requires a license.

1.2. Corpus Linguistics for forensic purposes

Forensic linguistics studies language in legal issues, that is to say, it deals with linguistic evidence in court. Linguistic evidence is useful in in the analysis of suicide notes, threats, trademark disputes or author identification (Eagleson 1994, Kniffka 2000, Chaski 2005, Juola 2006, Shapero 2011, Solan and Tiersma 2012). The practice of forensic linguistics involves law, language, psychology and other disciplines related to language and computation.

In order for forensic linguistics to serve its purpose of providing linguistic evidence in court, the discipline requires a series of scientific standards which makes it “reliable, replicable and respectable” (Chaski 2012: 4). Some standards for forensic linguistic methodology include that forensic linguistics provides an empirical analysis, is grounded in linguistic theory and can be replicated (Chaski 1997, 2001, 2005, 2012). Coulthard (1994:40) advocated for the use of corpus in forensic linguistics given the possibilities that the empirical exploration of corpora can provide in terms of evidence and investigation. Indeed, Coulthard states that "any improved methodology must depend, to a large extent, on the setting up and analyzing of corpora." In fact, nowadays it is difficult to understand forensic linguistics as a legitimate forensic science without corpus linguistics. Chaski (1997, 2001) developed the first corpus for forensic authorship identification so that methods can be grounded in empirical analysis.

On another front, present day corpus analysis is bound to computational linguistics, as stated above. Computational methods allow the researcher to deal with huge amounts of linguistic data in an objective way, with the help of statistics (Guillén et al., 2008). Analysis may involve different types of variables that can be explored in an accurate way. Authors such as Koppel et al. (2009:9) advocate for the use of machine learning in the study of authorship attribution: "these modern techniques, together with recent advances in natural language processing, have enabled the development of a plethora of potential markers of authorial style". Chaski (2005, 2007) has applied statistical and machine learning approaches that focus on syntactic variables since these can be related directly to linguistic theory.

LIWC: Linguistic Inquiry and Word Count (Pennebaker, Francis and Booth 2001) provides text analysis algorithms of words related to conceptual categories. The linguistic approach is very similar to content analysis, in line with the General Inquirer, the first computer system for content analysis (Stone, Bales, Namenwirth, and Ogilvie 1962, and Stone, Dunphy, Smith and Ogilvie 1966). One important difference between LIWC and the General Inquirer is that LIWC focuses on the word as the unit of analysis, while the General Inquirer is based on the sentence as the unit of analysis. But both LIWC and the General Inquirer relate linguistic text to other categories of cognition.

ALIAS: Automated Linguistic Identification and Assessment System (Chaski 1997, 2005) provides a database platform for corpora within forensic linguistics. ALIAS combines a database for document storage with text analysis algorithms at phonological-orthographic, morphosyntactic, semantic and discourse levels. Within ALIAS, specific modules provide algorithms for specific forensic purposes. SynAID, UniAIDE, and LexiAIDE, for example, provide algorithms related to authorship identification, while SNARE and S-Qual focus on suicide note assessment, ThreatAssess and T-Qual provide threat assessment, InterTexter for measuring similarity among documents, among others, and TATTLER provides text analysis fundamentals that are used in the algorithms for specific forensic purposes.

In sum, corpus linguistics offers a reliable and valid framework to study language patterns and offer accurate results based on objective techniques. Accordingly, what corpus linguistics in forensic linguistics provides is a reliable foundation for a methodology far from speculation, subjectivity and mere intuition.

2. Data collection: compilation of our corpus

The instrument used to elicit data from students was Chaski's (2001). The questionnaire consisted of ten separate questions, described in Table 1, and, for each of them, subjects were requested to write a different type of text. The questions evoke different text types and communicative purposes, different registers of formality, and different levels of emotionality.

Task ID	Topic
1	Describe a traumatic or terrifying event in your life and how you handled it
2	Describe someone or some people who have influenced you
3	What are your career goals and why?
4	What makes you really angry?
5	A letter of apology to your best friend
6	A letter to your sweetheart expressing your feelings
7	A letter to your insurance company
8	A letter of complaint about a product or service
9	A threatening letter to someone you know who has hurt you
10	A threatening letter to a public official or celebrity whom you do not know

Table 1. Writing topics for writing sample database

2.1. Experimental Group of Gender-based abusers

The experimental group comprises convicted participants, all accused of gender violence against their partners in couples or ex-couples. They are all serving their sentence in the penitentiary institution of Sangonera (Murcia). Thirteen prisoners agreed to answer the questionnaire. The range of age is wide: from 19 to 63 years old.

2.2. Control group

One part of the control group was made up of 60 Spanish undergraduates of Engineering at the Polytechnic University of Cartagena (UPCT). Participants were both males and females, and the age range of the sample was between 17 and 20, with the overwhelming majority of the respondents aged 18. A total of sixty students took part in the study. To encourage participation, each of the students was given half a point to be added to their final course grade.

Another part of the control group consisted of 13 students from Universidad de Castilla-La Mancha. They were all students in the fourth year of the Primary Education Teaching degree. They were asked if they wanted to participate in a study about forensic linguistics, yet no details were given so as not to bias their answers. In order to carry out the data collection, the Dean of the faculty was informed and he had to formally accept the process. Initially, we had up to 30 students but only 14 finished the questionnaire. This decrease in the number was due to two main reasons. First, some of them decided not to participate because they did not want to talk about their personal experiences. Second, some others started the questionnaire but did not finish it. Therefore, only around 50% (13 of approximately 30) of the potential sample could be analyzed.

Last but not least, students in the fourth year of English Studies at the University of Alicante were invited to contribute to our study as well. A student leader contributed greatly to the project by creating a student working group in Facebook to the effect that her fellow students could discuss issues concerning their participation in the Spanish forensic corpus project. In general, students were thrilled to be able to contribute to a real forensic linguistics project, a scientific field new to them, and appreciated that their contributions were considered valuable, creative and productive. Thanks to the student leader's interest, enthusiasm and support, a total number of twenty students volunteered to contribute to the making of the Spanish forensic corpus by completing the ten-item questionnaire. Participants were given a two-month deadline to submit the texts and granted with a certificate of contribution from the Institute for Linguistic Evidence in Delaware (USA). When the deadline was over, each contributor submitted a separate electronic folder comprising ten anonymous files -one per answer to each stimulus question- and an additional file with personal data to the student leader and the latter handed them over to the researcher. Finally, the researcher shared via Dropbox the two hundred texts collected at the University of Alicante for the Spanish forensic corpus project with the lead researchers.

2.3. Challenges

The key challenge in data collection was the access to the participants who were imprisoned. It is worth noting that Trial Court for Gender Violence is a specialized criminal courtroom associated to the Inquiry Courts, established by the Organic Law 1/2004 of Comprehensive Protection Measures against Violence against Women. Extraordinarily these courts have also powers in the civil jurisdiction acting as Courts of First Instance and Inquiry. They are associated to the Judicial District, even though one court can be created to cover the area of two or more districts. Due to the different drawbacks that we were finding, the compilation of data was delayed and the present study is still at a preliminary stage, as we expect to collect additional data from the prison population.

2.4. Compliance with Ethical Principles for Human Subjects Research

The Chaski Writing Sample Database stimulus questions (Chaski 2001) had previously been approved by the Institutional Review Board of the Institute for Linguistic Evidence comprised of a psychiatrist, psychologist, attorney and linguist. In the Spanish setting we did not have to obtain approval from an IRB, but we still followed the ethical standards required by an IRB. The present authors complied with all the ethical requirements for forensic research, namely anonymity, confidentiality, preservation of the data, and restricted access to the data.

3. Data analysis method

The data analysis used both a fully-automated approach and a linguistic team approach. The *automated analysis* was conducted by means of the proprietary software described in 3.1, Linguistic Inquiry and Word Count (LIWC, a comprehensive description can be found in Pennebaker et al. 2001). The *linguistic team approach* involved manual evaluation of the texts by two linguists who worked independently. Theses independent analyses were subsequently compared and compiled by a third staff linguist, since these steps avoid bias in the analysis and ensures inter-rater reliability; in other words, the use of more than one linguist working independently of each other enables blind work and avoids the impressionistic judgments formed by one single analyst. This procedure has contributed substantially to the robustness and reliability of the present scientific method.

3.1. Semantic focus

An accepted set of psycholinguistic categories for the computational analysis of language is provided by Linguistic Inquiry and Word Count (LIWC). LIWC has been used for providing empirical evidence on the relationship between language and the state of mind of subjects (Pennebaker et al. 2001), their mental health (Hancock et al. 2011), and, more recently, the truth value of their speech (Newman et al. 2003; Almela et al. 2013). The program maps each word against a dictionary containing a series of words and the psychologically meaningful categories to which each word is assigned, working out the percentage of words which fall into the four broad dimensions, namely linguistic processes, psychological processes, relativity and personal concerns.

The LIWC dictionary generally arranges categories hierarchically. Thus, some of the categories are the sum of others. For example, the category “Total pronouns” comprises “1st person singular”, “1st person plural”, “Total 1st person”, “Total 2nd person”, and “Total 3rd person”. The categories “1st person singular” and “1st person plural”, in turn, are both subsumed under “Total 1st person”. Some previous studies such as Newman et al. (2003) and Fornaciari and Poesio (2012) explore categories at different levels in the hierarchy using the same automatic classifier, which can be considered a methodological flaw. In machine-learning classification and statistical techniques, this would result in redundancy, which may yield skewed results. In order to avoid this, there are two options: either removing the hierarchically superior categories, or keeping them and leaving the inferior categories out. In this case, the first option has been selected so as to keep as much specific information as possible.

In the team approach, the linguists assessed the texts using categories which were not available in LIWC, but discovered during the manual linguistic analysis. The linguists were able to document semantic categories beyond the word level, such as phrase, sentence and discourse levels. For example, self-contradiction in the text occurs at the sentence and discourse level; self-contradiction is not exclusively word-based and thus not identified by LIWC. The linguists' analysis offers a more semantically rich analysis than a focus on words-only affords.

3.2. Spelling variation and software

It is worth noting that the data from the gender abusers was fraught with spelling errors. Since text analysis tools such as LIWC rely on word-matching, spelling errors cause problems. Therefore, we used a spell-checked version of the gender abusers but not for the college students (whose writings did not contain spelling errors). Due to the prevalence of spelling errors in the data from the experimental group, ALIAS TATTLER was modified to allow the storage of both a spell-checked version as well as the original one for different tasks including semantic analysis.

3.3. Can we separate the groups and how are they distinguished?

In order to achieve a balance between the experimental and the control group, the former was taken integrally for analysis, and it was contrasted to a random sample taken from the latter. This sample was equivalent in size to the experimental data, that is to say, 13 contributions from each group.

We decided to keep only thematic categories which at least occurred five times in the corpus as vaiables for statistical analysis.

Regarding the tests carried out, a logistic regression and a discriminant function analysis were conducted with IBM SPSS.

4. Results and discussion

First, it is indeed possible to separate abusers from non-abusers using linguistic analysis and statistical analysis. According to the logistic regression, it is possible to distinguish the gender abuser from the non-gender abusers with 80.8% accuracy (only taking the 5 most relevant categories and 13/13 texts, which is rather positive). Accuracy is slightly lower but still relevant when an alternative statistical classifier, discriminant function analysis, was performed (76.9%).

Second, the thematic categories that serve as variables in the statistical distinction between abusers and non-abusers are striking and complex.

<u>Money</u> and <u>love</u> are the thematic categories that are statistically significant according to both classification tests (see Figure 1). Both <u>money</u> and <u>love</u> categories are more present in the control groups' texts than in the abusers' texts. It

is perhaps expected that money and love, two issues in many marital disputes, differentiate abusers from non-abusers but it is striking that these themes are more present in non-abusers' writing than in abusers' writing.

We can see that there are differences in the other categories, but they are not relevant in the two tests that were carried out.

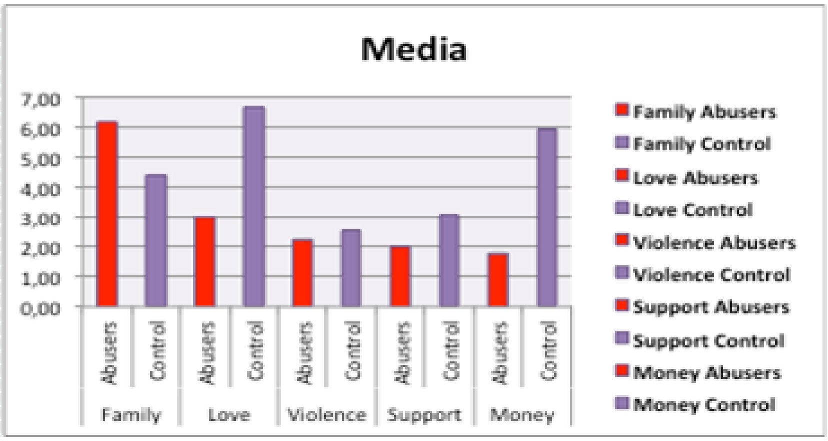

Figure 1. Relevant categories

Regarding other thematic categories for the analysis, family and job were found in all texts. Love and prison are also almost a constant, found in 11 out of the 12 texts analyzed. Categories like violence, threat or anger might have been expected to be more present in the corpus, but violence and anger themes are only found in 5 texts and the threat theme is found in only 2 texts. It is also interesting to mention that the category of ex-couple occurs in more texts (5) than the category of couple which appears in 3 texts. In fact, when the texts talk about love, they usually refer to fraternal and not romantic love. Contradiction is found in 7 out of the 12 texts analyzed.

Regarding the distribution of categories, family appears in all texts, followed by support and suffering. The themes prison and friends are present in at least half of the texts. On the other hand, categories such as self-esteem, abandon, distrust or suspicion are concentrated in one or two texts.

Not all thematic categories occur with the same regularity and strength. In fact, categories related to words feel or improve do not occur at all in the abusers' corpus. Family is the one that is the most common in most texts.

Figure 2 shows that family is in the first position followed by love and violence. On the contrary, health or wickedness were the least relevant for distinguishing the groups. It is also interesting that, contrary to what we might expect, categories such as anger or threat are not among the most relevant for distinguishing the groups. They are rather towards the right tail of the graph.

The shape of Figure 2 is reminiscent of Zipf's law (Zipf 1949). Zipf's law states that, given some corpus of natural language utterances, the frequency of any word or language category is inversely proportional to its rank in the frequency table. Thus, the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc. Thus, the rank-frequency distribution is an inverse relation. We checked that it applied to the semantic categories analyzed as well as to isolated words, as can be seen in Figure 2.

In an attempt to know whether there were any patterns among texts, a cluster analysis was carried out. The dendrogram shown in Figure 3 was obtained. A dendrogram is a graphical summary of the cluster solution. Cases are listed along the left vertical axis. The horizontal axis shows the distance between clusters at the point at which they are joined.

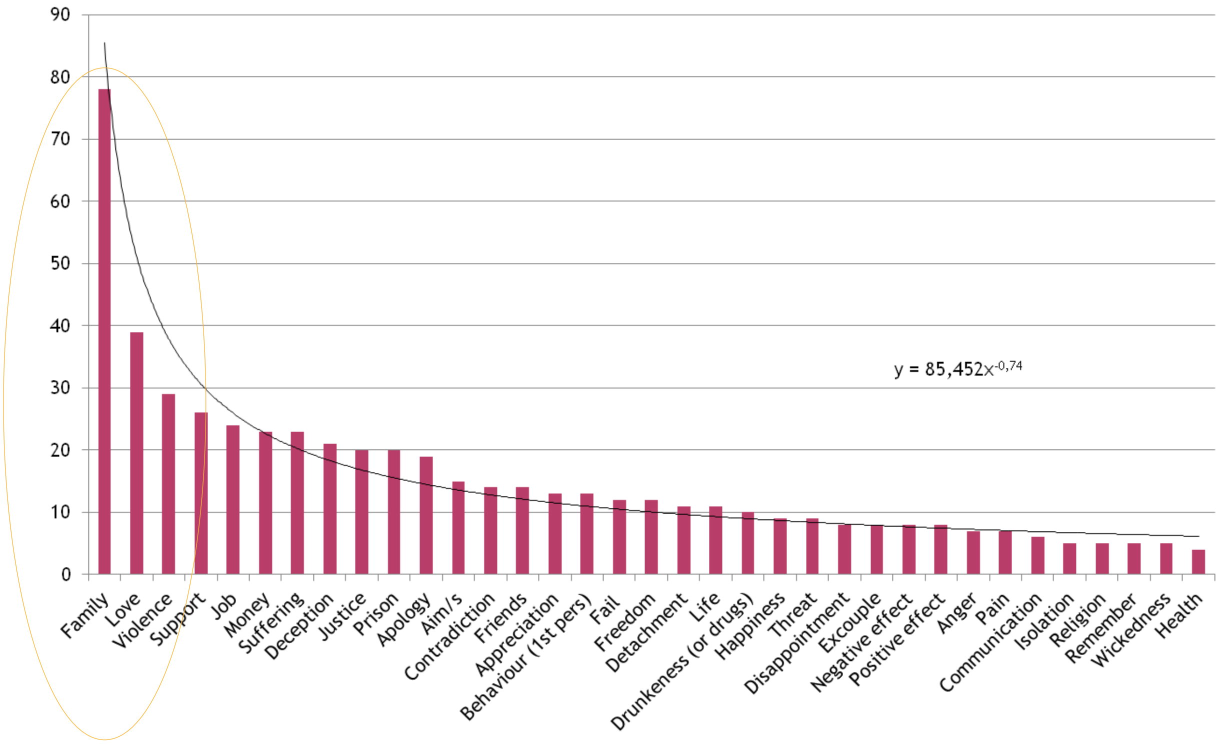

Figure 2. Distribution of texts across variables

Parsing the classification tree to determine the number of clusters is a subjective process. Generally, we begin by looking for "gaps" between joints along the horizontal axis. The cluster analysis reveals that texts 9, 13 and 10 are quite similar in their linguistic profile and their categories, and that texts 2, 11, 6 and 8 are the most dissimilar texts.

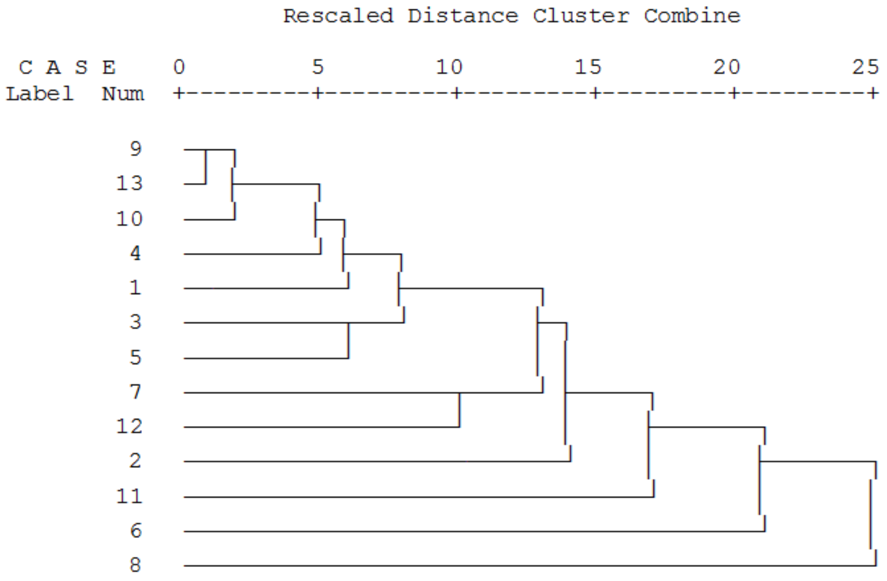

Figure 3. Grouping of texts based on variables

The cluster analysis also reveals that there are some categories that tend to co-occur (see Figure 4). This is the case of anger, ex-couple, happiness, health, isolation, negative effect, pain, positive effect, pregnancy, religion, remember, and wickedness. Some other co-occurrences found are behavior with fail, or job with justice and prison, which seems to suggest a causal relation among them, at least in the authors' cognition.

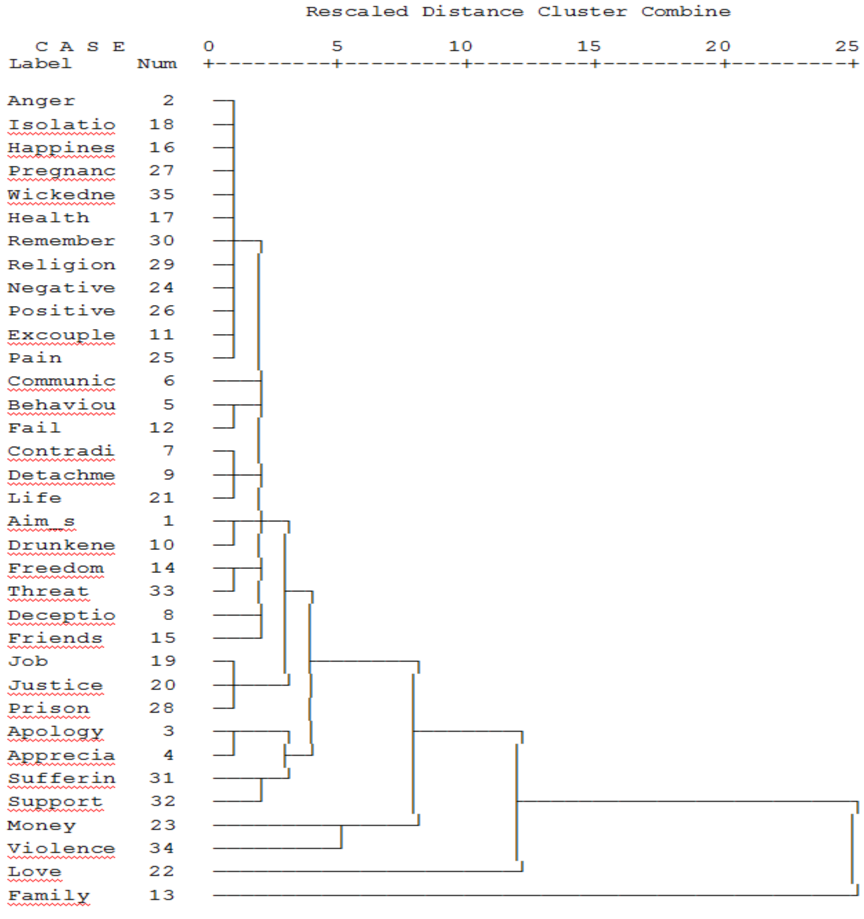

Figure 4. Grouping of co-occurring variables

The statistically significant differences found in the category love between the abusers and the control group may suggest the lack of empathy distinctive of the former (Johnson, 2006). The most unexpected finding was the overwhelming presence of the category money in the control discourse, which might be related to the fact that the abusers, being imprisoned, did not have direct contact with real money at the very moment of the experiment, which might have led to the suppression of this category in their discourse.

5. Conclusions and further research

LIWC has provided an automated text analysis which we supplemented with manual analysis to present higher level semantic analysis. While we could use ALIAS for some semantic, word-based analysis, ALIAS is a powerful piece of software for forensic linguistic evidence which can provide more sophisticated analysis. Therefore, we are working to bring ALIAS syntactic analysis techniques to Spanish. We expect to be able to analyze the data advanced with ALIAS in the near future, performing morphosyntactic analysis.

Gender violence in Spain has become a global social problem, thus it needs to be tackled from different perspectives. Forensic linguistics is one of the disciplines that can contribute to this. However, there is a long path towards the consolidation of this discipline in Spain. We need to use stringent protocols of action and accurate instruments that allow us to study the abusers' language and extract linguistic patterns that may arise from their discourse. As Chaski puts it,

> *It is the linguists' responsibility to create the theoretically sound hypotheses, test these hypotheses and perform the empirical evaluation of our own methods. It is the linguists' responsibility to recognize junk science before it gets to court. These are especially interesting obligations in the case of forensic linguistics, because, as any linguist knows, everyone has something to say about language, and linguistics has many sister-disciplines within academia. (2001: 2).*

Therefore, this project can give us the opportunity to create linguistic profiles that may help in the investigation and prosecution of domestic violence and gender-based abuse.

Finally, we have demonstrated that corpus development using the protocol in Chaski (1997 and 2001) is possible. Using the collected corpus, we have demonstrated that the Chaski Writing Database stimulus questions evoke data from which it is possible to distinguish abusers from non-abusers through linguistic patterns, with about 80% statistical accuracy.

REFERENCES

Almela, A., Alcaraz-Mármol, G. and Cantos, P. (2015). Analysing deception in a psychopath's speech: A quantitative approach. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 31(2), 559-572.

Almela, A., Valencia-García, R. and Cantos, P. (2013). Seeing through Deception: A Computational Approach to Deceit Detection in Spanish Written Communication. *Linguistic Evidence in Security, Law and Intelligence*, 1(1), 3-12.

Baker, P. (2012). Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies*, 9(3), 247-256.

Cantos Gomez, P. (2013). *Statistical Methods in Language and Linguistic Research*. Sheffield, UK: Equinox Publishing Ltd.

Chaski, C.E. (2001). Empirical Evaluations of Language-based Author Identification Techniques. *International Journal of Speech, Language and Law* (previously *Forensic Linguistics),* 8(1): 1-66.

Chaski, C.E. (2005). Who's at the keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence*, Spring 2005.

Chaski, C.E. (2007). The Keyboard Dilemma and Author Identification, in *Advances in Digital Forensics III,* Sujeet Shinoi and Philip Craiger, eds., New York: Springer.

Chaski, C.E. (2012). Best Practices and Admissibility of Forensic Author Identification. *Journal of Law and Policy,* 21(2). Brooklyn Law School.

Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *International Journal of Speech, Language and Law* (previously *Forensic Linguistics),* 1(1), 27-43.

Eagleson, R. (1994). Forensic analysis of personal written texts: a case study. In J. Gibbons (Ed.), *Language and the Law*. London: Longman.

Fornaciari, T. and Poesio, M. (2012). Sincere and deceptive statements in Italian criminal proceedings. In *Proceedings of the International Association of Forensic Linguists Tenth Biennial Conference* (pp. 126–138).

Guillén, V., Vargas, C., Pardiño, M., Martínez, P. and Suárez, A. (2008). Exploring State-of-the-art Software for Forensic Authorship Identification. *International Journal of English Studies*, 8(1), 1-28.

Hancock, J.T., Woodworth, M.T. and Porter, S. (2011). Hungry like the wolf: A word-pattern analysis of the language of psychopaths. *Legal and Criminological Psychology,* 18(1), 1-13.

Johnson, S.A. (2006). *Physical Abusers and Sexual Offenders: Forensic and Clinical Strategies.* New York: Taylor and Francis.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3), 233-334.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London/New York: Longman.

Kniffka, H., (2000). Anonymous Authorship Analysis without Comparison Data? A Case Study with methodological impact. *Linguistische Berichte,* 182, 179-198.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology,* 60(1), 9-26.

Leech, G. (2005). Adding Linguistic Annotation. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Oxbrow Books.

Leech, G. (1992). Corpora and theories of linguistic performance. In Jan Svartvik (Ed.), *Directions in corpus linguistics.* Berlin: Mouton De Gruyter (pp. 105-122).

McEnery, T. (2003). Corpus Linguistics. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice.* Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Newman, M. L., Pennebaker, J. W., Berry, D. S. and Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29, 665-675.

Parodi, G. (2008). Lingüística de corpus: Una introducción al ámbito. *Revista de Lingüística Teórica y Aplicada*, 46(1), 93-119.

Pennebaker, J. W., Francis, M. E. and Booth, R. J. (2001). *Linguistic Inquiry and Word Count.* Mahwah (NJ): Erlbaum Publishers.

Renouf, A. (1987). Corpus Development, in Sinclair, J. M. (ed.) *Looking Up*. Glasgow/London: Harper Collins Publishers.

Saldanha, G. (2009). Principles of corpus linguistics and their application to translation studies research. *Tradumàtica* 7, 1-7.

Shapero, J. J. (2011). *The Language of Suicide Notes*. Unpublished Thesis. University of Birmingham

Stone, P.J., Bales, R.F., Namenwirth, J.Z., and Ogilvie, D.M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Journal of the Society for General Systems Research,* October 1962.

Stone, P.J., Dunphy, D., Smith, M.S., and Ogilvie, D.M. (1966). *The General Inquirer: a computer approach to content analysis.* Cambridge, MA: MIT Press.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10(1), 1-13.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam and Philadelphia: John Benjamins.

Zipf. G.K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, Massachusetts: Addison-Wesley.