Article

A Corpus-Based Study of Linguistic Deception in Spanish

Ángela Almela

School of Arts, Universidad de Murcia, 30001 Murcia, Spain; angelalm@um.es

Featured Application: Statistical text classification.

Abstract: In the last decade, fields such as psychology and natural language processing have devoted considerable attention to the automatization of the process of deception detection, developing and employing a wide array of automated and computer-assisted methods for this purpose. Similarly, another emerging research area is focusing on computer-assisted deception detection using linguistics, with promising results. Accordingly, in the present article, the reader is firstly provided with an overall review of the state of the art of corpus-based research exploring linguistic cues to deception as well as an overview on several approaches to the study of deception and on previous research into its linguistic detection. In an effort to promote corpus-based research in this context, this study explores linguistic cues to deception in the Spanish written language with the aid of an automatic text classification tool, by means of an ad hoc corpus containing ground truth data. Interestingly, the key findings reveal that, although there is a set of linguistic cues which contributes to the global statistical classification model, there are some discursive differences across the subcorpora, yielding better classification results on the analysis conducted on the subcorpus containing emotionally loaded language.

Citation: Almela, Á. A. Corpus-Based Study of Linguistic Deception in Spanish. *Appl. Sci.* **2021**, *11*, 8817. <https://doi.org/10.3390/app11198817>

Academic Editors: Arturo Montejo-Ráez and Salud María Jiménez-Zafra

Received: 15 August 2021
Accepted: 17 September 2021
Published: 23 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: text classification; linguistic corpus; deception; linguistic cues; statistical analysis; discriminant function analysis

1. Introduction

The distinction between truth and deception has garnered considerable attention from domains such as formal logic and psychological research. In the field of human kinetics, non-verbal communication has been claimed to play a key role in the detection of deception. More recently, verbal cues to deception have been also explored, as the investigation of linguistic cues to deception in written language has proved to be of utmost importance not only in the forensic context with statements written by witnesses and people involved in crimes, but also because in the increase seen by computer-mediated communication, where written texts constitute a fundamental element.

In the last decade, the field of natural language processing (NLP) has devoted considerable attention to the automatization of the process of deception detection, developing and employing a wide array of automated and computer-assisted methods for this purpose, (see, for example, Ott et al. [1] and Quijano-Sanchez et al. [2]). Researchers in [3] provide a thorough review of this activity. Similarly, another emerging research area is focusing on computer-assisted deception detection using linguistics, [4,5], with promising results. Thus, some computational approaches supervised by experts in the field are considered an efficient way to supplement and support criminal investigators, being of special interest to linguists, jurists, criminologists, and professionals in the field of communications.

Accordingly, in the present study, an overall review of the state of the art regarding linguistic cues to deception is provided, as well as an overview on several approaches to the study of deception and on previous research into its linguistic detection, describing the main controversies in the area (Section 2). Furthermore, the present author draws a distinction between software packages specifically developed for linguistic deception

detection and other verbal assessment tools that are widely used for this and many other purposes (Section 3). Section 4 provides the materials and methods used in the experiment reported, whose results are presented and discussed in Section 5. Lastly, in light of the results obtained, some conclusions are drawn in Section 6 as well as some suggestions for further research.

All in all, this study makes a substantial contribution to the study of computational linguistic tools as an aid to deception detection and deepens the readers' understanding of the linguistic mechanisms underlying deceit. Interestingly, it offers a description of the linguistic cues to deception and promotes a contextualized study of deception, rather than dealing with broader dimensions of analysis.

2. Automated Deception Detection

This section presents the essentials of automated deception detection and advances some prime considerations that, from the present author's viewpoint, should be taken into account when conducting research in this area. For a whole account of theories and controversies in the area of deception detection in general, the reader may resort to [6], which reports past and current research on all aspects of lying and deception, as it is a comprehensive exploration of the state of the art from the combined perspectives of linguistics, philosophy, and psychology.

2.1. Essentials of Linguistic Deception Detection

As stated in [7], context has proved to be an important aspect in research and affects the relation between lying and language. These authors have developed a model called the contextual organization of language and deception (CoLD), which provides a framework including some crucial aspects of context for any deceptive communication. Thus, the nature of the linguistic data in the corpora is worth commenting on. Much has been discussed about the importance of deception in spontaneously produced language. Laboratory-produced lies have been criticized in forensic literature for not being very reliable; for instance, the authors in [8,9] suggest that further research should involve retrospective studies in law enforcement settings to study realistic responses with known outcomes. However, the strength of laboratory-produced data is the possibility for controlling variables and attributes so that the conclusions drawn are experimentally valid. What remains constant during such an experiment are the participants and the topics on which they write, which allows the researcher to avoid confounding intervening variables and to focus on deception in opinions and memories as the only plausible causal factor. Put another way, providing that some variation is observed regarding the dependent variables analyzed, this scientific control will allow the author to assure that the participants' situations are identical until they are asked to lie, and so the potentially new outcome may be attributed to the independent variable. The usefulness of this kind of corpus has indeed been proved in the forensic context, as shown in such studies as [10].

In this respect, it is also worth noting that there are two types of data: low-stakes deception, in which no harm can be done (it is well known that people lie in social situations without intending harm); and high-stakes deception, where real-life damages are possible and likely. This distinction must be considered when drawing conclusions in automated and computer-assisted deception detection research.

Furthermore, a closely related issue in forensic computational linguistics is the importance of working on ground truth data that are forensically feasible. 'Ground truth' data means data for which we know what the correct answers are; thus, for the particular field of deception detection, we need data where we know which texts are true or false. When a method is tested on ground truth data, we can conduct validation testing and accurately report its error rate. In empirical research, validation testing is a technique that determines how well a procedure works, under specific conditions, on a corpus containing texts of known origin [11]. Thus, on a database of ground truth data, the researcher is to apply a replicable analytical method to every text as well as a cross-validation scheme, most

typically by building a statistical or a machine learning (ML) model. Last, the error rate is to be computed from the misclassifications in the analysis.

Within the research paradigm of forensic computational linguistics, in the present article, a corpus-based study is presented, attempting to answer the question ‘Is this truthful or false?’ It is worth noting that automated and computer-assisted methods in other corners, such as author identification, are much more consolidated worldwide and generally admitted in court, such as Chaski’s SynAID [12,13], as compared to computer-assisted deception detection, which is not often used for veracity assessment in the legal setting. In other words, in many, if not all, jurisdictions, experts are not allowed to testify that a person is lying, as only the jury or the judge can do it. Thus, deception detection is only an investigative tool, that is to say, its use is restricted to investigation, not trial. However, some expert witnesses, such as the present author, are currently refining specific computational tools, which have proved reliable in research contexts, in order to promote the implementation of empirical investigative methods in real-life forensic settings.

2.2. *The Role of Linguistic Variables in the Computational Analysis of Deception*

As has been seen, deception detection can play a role in the investigation of different security issues, civil cases, and even some types of crimes, and, according to the Institute for Linguistic Evidence (ILE) (<https://linguisticevidence.org/>, accessed on 4 July 2021) paradigm, standards for forensic computational linguistic methodology include that forensic linguistics provides an empirical analysis grounded in linguistic theory [11]. Furthermore, the adoption of totally automated deception detection methods and mixed machine–human methods entail some basic stages: choosing an appropriate linguistic level, properly codifying the variables of analysis, engaging in statistical analysis, and conducting validation testing.

These kinds of analyses can make use of variables from different linguistic levels, namely, the phonemic, morphemic, lexical, syntactic, semantic, and pragmatic. As stated in [11], forensic methods dealing with written data have focused on analytical units at the character, word, sentence, and text levels. Specifically, some studies, such as [14], present automated methods for deception detection operating at the character level, whose analytical units include, among others, single characters, punctuation marks, or character-level n-grams (units of adjacent characters). At the word level, analytical units can be word-level n-grams [15], lexical semantics [16], and vocabulary richness [17]. Sentence-level analytical units can include part-of-speech (POS) tags [18], sentence type [19], average sentence length [20], and average number of clauses per utterance [21]. At the textual level, analytical units can include text length [22] and discourse strategies [23], to name but a few. The easiest patterns to detect by machine are character and word level features. On the contrary, at other linguistic levels, automatic pattern detection is harder, especially with forensic data, as they are often messy. For instance, sentence level features can be extracted automatically, but most parsers require human revision of the output to ensure the accuracy of the analysis.

In their meta-analysis of computational deception detection, [24] explored 44 studies and a set of 79 cues, which seemed reasonably consistent across previous literature. Despite some inconsistencies, the authors reported some common conclusions from the poll of studies reviewed: in broad terms, liars experienced greater cognitive load than truth-tellers; using fewer words related to cognitive processes, they used more negative emotion words, detached themselves from the events narrated, and used fewer sensory–perceptual words. Nonetheless, words expressing uncertainty were found indicative neither of deception nor of truth. All in all, the results varied across the studies according to event type, involvement, intensity of interaction, and motivation, among other variables.

3. Description and Explanation of the Most Significant Methodologies

In this section, the main tools for automated deception detection are presented (a schematic overview is provided in Figure 1). The first group is aimed at the automatic

extraction of lexical features for different purposes, whereas the second group includes software specifically developed for the computational classification of written statements as true or false.

3.1. Automatic Extraction of Linguistic Features Applied to Detecting Deception

One of the earliest attempts at automated content analysis was the General Inquirer [25,26], and some years later, [27] assessed several linguistic cues, using TEXAN, a computer system that analyzed word frequencies by keypunching the words to map them to different lexical categories, with the main purpose of differentiating truths from lies in the written medium.

In the last 20 years, some more modern content analysis approaches were developed in research contexts on similar grounds, outstandingly the linguistic inquiry and word count, or LIWC [28]. 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 was based on the sentence, but both systems relate linguistic text to other categories of cognition. Specifically, the categories used in the original version of LIWC were related to standard linguistic processes, psychological processes, relativity, and personal matters; a detailed description of the individual categories can be found in [29]. It has been also adapted and translated into more than 10 languages, including Spanish [30], as will be seen in the exemplary study presented below. In sum, LIWC provides a tool for studying the emotional, cognitive, and structural components contained in language on a word-by-word basis, working out the percentage of words which fall into those categories. Ref. [16] were the first researchers to use this system for deception detection, yielding above-chance accuracy of classifications for different types of lies. Even if LIWC is not entirely unproblematic as an analytical tool in linguistics [31], over the last few years, it has been widely used in such fields as forensic linguistics [15], sentiment analysis [32], and psycholinguistics [33] with considerable success.

Some other automatic corpus classification tools have been developed beyond word frequency analysis, such as CohMetrix [34,35]. It analyzes cohesion relations, taking into account the meaning and context in which words or phrases occur in texts. Ref. [36] was the first piece of research where it was applied to deception detection.

3.2. Software Developed for the Computational Classification of Written Statements as True or False

The software specifically developed for linguistic deception detection is presented in this section. One of the most famous methods for deception detection is scientific content analysis (SCAN). It was developed in 1987 [37], a polygraph examiner, and methods based on it are generally known as statement analysis. Most of the literature published on this type of analysis is merely descriptive (see, for example, Lesce [38] and McClish [39]), although it was automated with reported accuracy results of 71% in [10]. However, as stated in [40], SCAN and other statement analysis systems have been mainly used and taught by practitioners manually, with several studies having examined SCAN with suggestive but inconsistent results [41,42].

Some other computational tools have been specifically developed for deception detection, such as Agent99Analyzer [43], created to extract linguistic cues to deception from texts and videos, iSkim [44], or CueCal [14]. A somewhat different detection deception software is ADAM, or automated deception analysis machine [45], which focuses on editing processes, such as backspace or spacebar while typing messages as well as measuring response latencies. The main methodological drawback of this approach seems to be that it requires a keystroke analyzer to be on the interviewee's machine, which can be seen as an intrusion of privacy.

Figure 1. Schematic overview of the main computational tools for linguistic deception detection [2,4,10,14,25,26,28,34,35,43,44].

Remarkably, most previous studies in computerized deception detection have relied exclusively on shallow lexico-syntactic patterns. However, [19] were the first researchers to explore syntactic stylometry. Over four different subcorpora including service reviews and essays on different topics, the authors explore features derived from phrase structure grammar (PSG) parse trees, showing that they consistently improve the detection rate over several baselines that are based only on lexical features. Most relevantly, within the four subcorpora examined, they apply their method to the corpus from TripAdvisor collected for [1], improving the classification results obtained by its collectors by reaching over 91% accuracy.

In this line of linguistic sophistication, a valuable contribution to linguistic deception detection has been made by Witness Statement Evaluation Research (WISER), one of the tools provided by ALIAS Technology (<https://aliastechnology.com/>, accessed on 4 July 2021), a company which offers forensic linguistics consulting to attorneys, law enforcement, human resources, and security teams. WISER is a project that makes use of automated text analysis and statistical classifiers to determine the best protocol for the computational classification of true and false statements in the forensic-investigative setting. Ref. [4] tested this text analysis tool, based on ALIAS's module Text Analysis Toolkit Toward Linguistic Evidence Research (TATTLER). It combines linguistic analysis at the phonological, syntactic, and lexico-semantic levels and has been applied to deception detection classification on two types of corpora: low-stakes (laboratory) and high-stakes, actual statements in criminal investigations [46]. The low-stakes, laboratory data comprised two narratives of a traumatic experience, one truthful and the other false, from each participant, while the high-stakes data consisted of actual statements from real criminal investigations with non-linguistic evidence of their veracity or falsehood. The WISER method yielded substantially different results, as 71% of the texts in the laboratory corpus were correctly identified, using leave-one-out cross-validation, while the rate reached 93% for high-stakes deception, which can be considered the most successful rate published to date. Furthermore, this brings to light the contrast between lies told in a low-stakes, laboratory setting and those told in a police investigation. All in all, this study shows how TATTLER linguistic variables work better than text analysis tools used for different purposes, such as LIWC or simplistic NLP models, such as bag of words (BoW). The latter is an approach popular among computer scientists working in text classification. The term bag of words was invented by [47] and

developed by [48], and in this conception of language, each text is seen as a list of words and their frequencies without regard to any morphosyntax or semantics.

As stated above, context has proved to affect the relation between deception and language (see, for example, Almela et al. [22]). Thus, the development of software designed for specific contextual frameworks is especially valuable in deception detection. An outstanding example of contextualized analysis of deception is VeriPol [2], a model for the detection of false robbery reports in Spanish based only on their text. This tool, developed in collaboration with the Spanish National Police and the Ministry of the Interior, combines NLP and ML methods in a decision support system that provides police officers the probability that a given report is false. The impact of this tool was tested by means of an on-the-field pilot study that took place in 10 Spanish police departments in 2017, specifically on a corpus of 588 false robbery reports and 534 truthful robbery reports, which allowed for a robust validation on ground truth data (see Section 2.1). For the analysis, the authors applied feature selection techniques in their approaches, using model variables, such as POS tags, document statistics (e.g., number of tokens, lemmata, and sentences within a document), and unigram lemmata for the performance of ML and statistical classification techniques [2]. They concluded that, in general, the more details are provided in the report, the more likely it is to be truthful. Empirical results show that it is extremely effective in discriminating between false and true reports with a success rate of more than 91%, improving by more than 15% the accuracy of human expert police officers on the same corpus. The pilot study was so successful that nowadays, it is officially used in all the national police offices in Spain. This fact is indeed significant, as, despite the fact that computer-assisted deception detection is not generally accepted in Spanish courts, it is proved that investigative settings may benefit from its assistance. Indeed, the differences between the situation of forensic linguistics in English- and Spanish-speaking countries are worth noting at this point. As explained in [8], there is an ever-growing respect between British police, criminal psychologists and linguists, probably because of the well-established tradition of these disciplines in English-speaking countries. However, in Spain, these areas do not have such a long tradition, hence the difficulty when it comes to securing comprehensive assistance to conduct realistic lie detection studies in languages other than English.

All in all, computational detection deception in both the WISER and VeriPol studies demonstrate that detection is possible with over 90% accuracy, with high-stakes ground truth data.

4. Materials and Methods

This section will provide the reader with a corpus study of deception in Spanish, an empirical study whose aim is to explore the linguistic cues to deception in written language with the aid of an automatic text classification tool, adopting a forensic computational linguistic approach and testing it on an ad hoc corpus containing ground truth data.

4.1. Contextualizing the Study

Ref. [22] predates the experiment reported here. As stated above, in that study, Almela et al. (2013) conducted a classification experiment, testing the Spanish version of LIWC2001 [30] to classify a corpus similar to that of [15], trained and tested with a support vector machine (SVM) classifier, using the four dimensions of LIWC (standard linguistic dimensions, psychological constructs, general descriptors, and personal concerns) separately and then with the possible combinations of the four dimensions. The authors showed the relatively high performance of the automatic classifier in Spanish written texts through the experiments, conducted on three subcorpora, checking the discriminant power of the variables as to their truth condition, the two first dimensions, linguistic and psychological processes, being the most relevant ones. Specifically, the best performing combinations across all LIWC tests and topics was an F-measure of 84.5%, using the combination of all four categories on the good friend topic. For comparison with the other

LIWC studies that use F-measure, the highest F-measure reported in [1] was 76.9%, using the LIWC features alone on the more lexically constrained hotel reviews, and in [18], it was 79.6%. In [22], the authors state that the higher performance on the good friend topic shows the strong dependence of the task on the topic and attribute the better performance on this topic to the greater emotional involvement that narrators have in describing their best friend.

Building on this previous work, the study presented here is a subsequent experiment conducted on the same corpus, considering some of the authors' suggestions for further research in [22]. Of interest, the novelty of this experiment is twofold:

- (1) Regarding the variables for analysis, a fifth dimension is added to the original LIWC set, comprising some stylometric variables which have proved useful in other NLP tasks [49] (described in depth in Section 4.3.2).
- (2) Statistical tests are applied to the individual categories instead of the ML algorithms usually employed for automatic deception detection. Specifically, a discriminant function analysis and several logistic regressions is performed so as to assess the discriminant power of the independent variables individually, instead of testing the dimensions as a whole (described in detail in Section 4.4). This rule-based feature extraction is chosen to make the classifier more describable.

4.2. Research Question

The present study addresses the following research question:

How successful are LIWC individual categories and the further stylometric variables analyzed for deception classification on a Spanish ad hoc corpus containing written opinions and emotionally loaded language?

4.3. Methodology

This section outlines the different stages of the present study. It comprises three main issues: an introduction to the nature of the study, an account of the analysis variables, and a full description of the corpus.

4.3.1. Nature of the Study

The present study may be classified as quasi-experimental. Quasi-experiments resemble quantitative and qualitative experiments, but they lack random assignment of groups or proper controls [50]. This feature is sometimes seen as an inherent weakness, especially from the viewpoint of experimental purists in the natural sciences. However, this is a very useful design for measuring social variables since it is not always possible to accomplish a purely random allocation of groups when dealing with human subjects. Thus, the present research takes advantage of the possibilities of this experimental design by comparing two groups of participants under similar circumstances. As explained below, an inter-group comparison is drawn, delving into the similarities and differences of the linguistic profiling of deception in written communication across languages. In addition, an intra-group assessment was undertaken in order to explore differences across topics, using the truthful statements as the control subcorpus against which the untruthful dataset is compared. Due to the quasi-experimental nature of the study, the intention is not to generalize the inferences drawn from the data analysis, but to treat them cautiously.

4.3.2. Variables

Most of the core psychologically meaningful categories contained in LIWC [28] and described above were used. It is worth noting that all the variables selected from LIWC reflect the percentage of total words, with three exceptions: raw word count, words per sentence, and percentage of interrogative sentences.

Interestingly, 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 [16] and [18], explored categories from different levels in the hierarchy, using the same experiment, which can be considered as a methodological flaw. In ML classification and statistical techniques, this would result in redundancy, which may yield misleading results. As suggested by such authors as [5], in this case, the results might be skewed by counting those variables twice. 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 the present study, the first option was selected so as to keep the most specific information. Appendix A shows the LIWC categories removed and their correspondences. The first column contains the highest categories, the second one the subcategories, and the third one the subcategories of the previous subcategories—it is worth noticing that the categories which involve no complexity were not included. Categories in capital letters are the most general ones, which were altogether removed. These categories may comprise either categories in bold, which in turn comprise other lower categories, or just in italics, which are the terminal part of the sequence. Only terminal or most specific categories were kept and counted.

Furthermore, a group of punctuation marks measured by LIWC was also explored in the present study, namely period, comma, colon, semicolon, sentences ending with ‘?’, exclamation, dash, quote, apostrophe, parenthesis, and other punctuation. These variables were not previously explored in [22] because, despite being considered part of Dimension I (Linguistic processes), they were not included as LIWC default predictors.

Last, there are some linguistic features not included in LIWC which were deemed relevant for the present study too, gathered in the fifth dimension of variables. To the best of the author’s knowledge, despite having proved useful in areas, such as automated document readability [49], they have not been explored for deception detection yet. They were extracted from the statistics worked out by WordSmith Tools 5.0 (<https://www.lexically.net/wordsmith/index.html>, accessed on 12 March 2021). The first of these variables is a standardized type/token ratio; it is worth noting that the non-standardized version of this ratio was included in the LIWC standard linguistic dimensions, but it proved to be too size-dependent as an index of lexical richness [51]. Thus, the discriminant power of the original version of the ratio may be greater, due to the disparities among the values for the different texts, so it is not as reliable a measure as the standardized version. On the other hand, word length was considered as well. Despite the fact that a category similar to ‘complex words’ was already included in LIWC, namely ‘Sixltr’, all words longer than 6 letters were included. Since the general agreement in corpus linguistics is that complex words should include any word consisting of 8 or more letters [49], their frequency is used for the calculation of one of the independent variables: the ratio of complex words to the number of tokens. Similarly, the ratios of the total amount of 1-, 2-, 3-, 4-, 5-, 6-, and 7-letter words to the number of tokens were worked out. Furthermore, the average word length (in characters) and average text length (in sentences) were considered in this section too. A summary of all the variables is provided in Appendix B, with the variables not previously explored in [22] marked in bold.

4.3.3. Corpus Description

The design of the questionnaire for the compilation of the corpus was focused on three different topics: opinions on homosexual adoption, opinions on bullfighting, and feelings about a good friend. Specifically, the participants received instructions to imagine that they had 10–15 min to express their opinion about the topics. First, they were asked to prepare a text expressing their true opinions on the topics; then, they were asked to prepare a second text expressing the opposite of their opinions, thus lying about their true beliefs. For instance, in the case of the good friend topic, it implied giving positive account on a good friend, and then a false positive account on a bad friend, according to the respondent’s personal experience. The guidelines asked for at least 4–5 sentences in as much detail as possible. Regarding the motivation behind the choice of topics, it

paralleled that in [15]: the three tasks proposed to participants included two controversial topics (homosexual adoption and bullfighting), sensitive subjects, which caused people to entertain a personal opinion on them. As for the third topic, good friend, it was selected so as to offer a counterpart to the previous topics since it entailed less emotional involvement. Interestingly, the controversial topics dealt with in the present study are likely to generate guilt, preoccupation or remorse, despite not being a high-stakes situation.

The participants (100) were college students, native speakers of European Spanish. Thus, the task was assigned as an exercise for extra credit in a college course and conducted via email over the course of several days. Personal information, such as age and sex, was not taken into account since it was considered irrelevant to the present analysis. It was deemed of utmost importance to avoid overfitting, which may occur when a sample size is too small in relation to the number of variables used, since this could lead to over-optimistic results. It is generally agreed that, for this kind of analysis, it is necessary that the number of cases be twice the number of variables, expressed as $n = 2k$ [52]. In the present study, a set of 76 independent variables was used; thus, in principle a minimum of 152 contributions would be required. In this case, every subcorpus comprises at least 200 contributions—in the case of the subcorpora organized by topics. In line with [15], 600 contributions were collected—100 true and 100 false statements for each topic—with an average of 94 words per statement and a total of 56,882 words, so statistical overfitting should not be a problem in subsequent analyses. A manual check of the quality of the contributions was made, and each one was entered into a separate text file. Appendix C shows a sample of truthful and untruthful language for each of the three topics, and Figure 2 shows the structure of the sample used for the analysis.

Figure 2. Structure of the dataset.

The dataset was deposited by the present author in a publicly available database, namely <https://github.com/angelalm/DeceptionCorpus>.

4.4. Data Analysis

As regards the statistical methods applied, discriminant function analysis (DFA) and several binary logistic regressions (LR) were calculated with the software package IBM SPSS (<https://www.ibm.com/products/spss-statistics>, accessed on 30 March 2021) so as to assess the discriminant power of the variables individually. On the one hand, DFA had been successfully applied in linguistic analysis for the classification of unknown individuals and the probability of their classification into a certain group [53,54]. In principle, DFA is claimed to make more demanding requirements on the data since it assumes that it shares all the usual assumptions of correlation, requiring linear and homoscedastic relationships—homogeneity of variances—and normal distribution of the interval or continuous data. However, DFA is known to be robust, even when these assumptions are violated, as stated in several modern textbooks about multivariate statistics [55]. At any rate, as LR is well known as an alternative to DFA because it makes less stringent requirements of the data, for the three individual subcorpora, a one sample Kolmogorov–Smirnov test provided evidence against the null hypothesis, implying that the samples were not drawn from a normal

population. As only a few variables met the requirements of normality and only 100 cases are involved, binary logistic regressions were conducted on the individual subcorpora, where the categorical response has only two possible outcomes (untruthful/truthful). Thus, it can be stated that the analyses reported in the present article explore techniques based on statistical approaches instead of methods based on geometrical properties of the data, such as [4,11–13]. It is worth noting that, for each classifier, a leave-one-out cross-validation was run, all sets having an equal distribution between truthful and untruthful statements. This technique, considered exhaustive cross-validation, is used to evaluate how the results of a statistical analysis would generalize to an independent dataset. As explained in [56], the main difference from non-exhaustive cross validation methods, such as *k*-fold cross-validation, is that the latter does not compute all ways of splitting the original sample. Since the aim of this experiment is the prediction of the truth condition of the texts, a cross-validation was applied in order to estimate the accuracy of the predictive models. It involves partitioning a sample of data into complementary subsets, performing an analysis on the training set and validating the analysis on the testing or validation set [57]. For DFA and logistic regression, cross validation shows how reliable the linear function determined by the original group members is when each member is left out of the group.

5. Results and Discussion

First, the DFA shows a successful discrimination between truthful and untruthful accounts in the general corpus (Wilks' $\lambda = 0.699$, $\chi^2 = 210.7$, $p = 00.000$). Specifically, text length proves to be the best single predictor, as shown in Table 1 and Figure 3. Remarkably, the difference between this predictor and the next one in importance is 20 points. Despite this fact, the F-ratio for the next predictor, 1st person singular, is still rather high. There are some other variables identified as predictors shared with studies for English such as [15], namely 2nd person, friendship, insight, exclusive words, and 3rd person. The remaining predictors are words related to certainty, humans, sexuality, number, anger, semicolon, past, assent, future, and tentative words.

Table 1. F-ratios from DFA.

Predictors	LIWC Abbreviation	Examples	F	Sig.
Word count	WC	-	69.812	0.000
1st person singular	I	<i>I, my, me</i>	49.259	0.000
Certainty	Certain	<i>always, never</i>	39.199	0.000
Total second person	You	<i>you, you'll</i>	33.516	0.000
Friends	Friends	<i>pal, buddy, coworker</i>	30.167	0.000
Humans	Humans	<i>boy, woman, group</i>	27.682	0.000
Insight	Insight	<i>think, know, consider</i>	25.708	0.000
Exclusive	Excl	<i>but, except, without</i>	23.601	0.000
Sex and sexuality	Sexual	<i>lust, penis, suck</i>	21.871	0.000
Numbers	Number	<i>one, thirty, million</i>	20.568	0.000
Anger	Anger	<i>hate, kill, pissed</i>	19.397	0.000
Semicolon	SemiC	-	18.329	0.000
Total third person	Other	<i>she, their, them</i>	17.495	0.000
Past tense verb	Past	<i>walked, were, had</i>	16.643	0.000
Assents	Assent	<i>yes, OK, mmhmm</i>	15.909	0.000
Future tense verb	Future	<i>will, might, shall</i>	15.239	0.000
Tentative	Tentat	<i>maybe, perhaps, guess</i>	14.709	0.000

Figure 3. Visual representation of F-ratios from DFA.

As can be seen in Table 2, which gives information about actual group membership vs. predicted group membership, the DFA shows that 76.3% of the original grouped cases were correctly classified, as 77.7% of the truthful statements were correctly classified as truthful (233 out of 300), and 75.0% of the untruthful statements were correctly classified as untruthful (225 out of 300 statements). As regards the leave-one-out classification method, it achieved a success rate of 74%, the percentage of truthful statements correctly classified in the cross-validation being slightly higher than the percentage of untruthful ones (75.7% vs. 72.3%, respectively). Specifically, there is a difference of 10 more statements correctly classified (83 vs. 73 statements).

Table 2. Classification results from DFA (IBM SPSS).

		Deception	Predicted Group Membership		Total
			No	Yes	1
Original ^a	Count	No	233	67	300
		Yes	75	225	300
	%	No	77.7	22.3	100.0
		Yes	25.0	75.0	100.0
Cross-validated ^b	Count	No	227	73	300
		Yes	83	217	300
	%	No	75.7	24.3	100.0
		Yes	27.7	72.3	100.0

^a 76.3% of original grouped cases correctly classified; ^b 74.0% of cross-validated grouped cases correctly classified.

In order to present a comprehensive picture of the effectiveness of the statistical classification methods employed, a summary of the success rates is provided in Figure 4. The experiment conducted on the good friend subcorpus yielded the best results. In this case, the known bundles of truthful and untruthful texts were differentiated with 84.6% cross-validated accuracy, meaning that 84.6% of the time, we can tell truthful and untruthful texts apart from each other and identify them. Specifically, there is a difference of more than 9 points from the previous subcorpus in terms of success, homosexual adoption (84.6% vs.

75.4%), probably due to the fact that when speakers refer to a good friend, they are more likely to be emotionally involved in the experiment; they are not just giving an opinion on a topic which is alien to them, but relating their personal experience with a dear friend and lying about a person that they really dislike. This personal involvement is probably reflected on the linguistic expression of deception, as suggested by [16].

Figure 4. Cross-validated classification of truthful and deceptive statements.

Furthermore, Table 3 shows a collection of the predictors identified for truthful, marked with the initial “T”, and untruthful, initial “U”, statements across the examined corpora. It is worth noting that the identification of predictors has proved more successful at pinpointing categories indicative of truthful statements, the most widely shared among subcorpora being text length and 1st person singular.

Qualitative Evaluation

Previous research on deception detection has found that, broadly speaking, deceivers provide shorter responses, compared to truth-tellers (see, for example, DePaulo et al. [58]), as creating and managing misinformation is more cognitively demanding than telling the plain truth. This is also the case with participants in synchronous CMC, where time to plan the responses is limited, almost like in oral communication, which is in line with the present results. Regarding 1st person singular, a previous study conducted in Spanish [59] did not find a significant correlation with this feature. Nonetheless, the authors advanced that the communication topic might make a difference since their participants write about trips, which is unlikely to generate guilt, preoccupation or remorse. On the contrary, the controversial topics dealt with in the present study are more likely to arouse these feelings, despite not being a high-stakes situation.

On the other hand, the strongest predictors for untruthfulness are 2nd and 3rd person. The latter is clearly in line with previous research [14,60]. This cue entails detachment from the self when providing false or imprecise information, indicating the leading role of non-immediacy in deception. Accordingly, there is also a significant 2nd person orientation in untruthful statements, as in [15]. Interestingly enough, it has proved a predictor of deception in the subcorpora of good friend and in the whole corpus, confirming the preference of deceivers for non-immediacy.

Table 3. Predictors identified for truthful and untruthful statements across the subcorpora.

	Bullfighting	Homosexual Adoption	Friend	All
WC	T	T	T	T
1st p. sing.	T	T	T	T
2nd p.			U	U
3rd p.			U	U
Semicolon				T
Number			T	T
Anxiety				T
Insight				T
Sadness			T	
Friends			T	T
Humans		U		U
Posfeel		T		
Certainty			U	U
Achievement			U	
Inhibition			T	
Assent				U
Tentative				T
Future				T
Past				T
Inclusive		U		
Exclusive	T			T
Sexuality				T
Motion		U		

As for the rest of predictors, the results seem to be in line with previous research on the English corpora, with liars experiencing a greater cognitive load than truth-tellers, using fewer words related to cognitive processes and more negative emotion words, as well as fewer sensory-perceptual words [24].

Finally, a novel feature proved significant for the model in Spanish: the semicolon. As mentioned above, it was not previously explored in [22], as neither this one nor the other punctuation marks were included as LIWC default predictors. Although the average sentence length does not appear in any of the discriminant models, both variables are integrally related. As explained above, participants produced a larger number of words when telling the truth, especially the Spanish ones, hence the discriminant power of the semicolon in this language. Significantly, this is one of the novel findings in this study.

Overall, statistical classification methodologies with individual categories have performed better than the ML techniques with whole dimensions reported in [22]. Furthermore, the distribution of the classification results parallels that from the experiment with whole categories.

6. Conclusions and Suggestions for Further Research

All in all, the computational detection of verbal deception has come a long way in a short time, with accuracy scores ranging from 60% on laboratory data [60] to 93% accuracy on high-stakes corpora, as reported in [4]. Remarkably, research on high-stakes, real-life type of data has proved far more successful than results on low-stakes, laboratory data, although some relatively successful experiments using this kind of corpus were reported

in this work, which represents a step forward. Specifically, as regards the percentage of untruthful statements correctly classified in the cross-validation, the classifier yielded 74% accuracy for the whole corpus (DFA), 70.8% for the bullfighting subcorpus (LR), and 75.4% for the homosexual adoption subcorpus (LR). As regards the experiment conducted on the good friend subcorpus, untruthful texts were differentiated with 84.6% cross-validated accuracy (LR). As was stated, the main factor leading to success in these cases seems to be the delimitation of the topic and the communicative context, due to the strong dependence of the task on the topic and on the author's degree of emotional involvement. Thus, the highest degree of accuracy on the last dataset may be attributed to the fact that when referring to a good friend, the participants are more likely to be emotionally involved in the experiment; they are not just voicing an opinion on a topic which is alien to them, but relating their personal experience with a dear friend and lying about a person that they really dislike. This personal involvement is probably reflected on the linguistic expression of deception, as suggested in some previous studies [16,22].

Thus, even if the classification results from the experiments reported in the present article are not as high as those obtained on high-stakes datasets, the relative strength compared to earlier work on low-stakes corpora is worth noting. Furthermore, although the results may seem not good enough to use forensically, basing on the literature review conducted, it can be assumed that a classification method that proves acceptably successful on low-stakes deception will work even better on high-stakes data.

New methods for automated deception detection are continually being developed, especially in the computational paradigm, and in order for the area to move in the right direction, the availability of data tagged for ground truth seems crucial [40,61]. In this sense, collaboration with law enforcement may be of utmost importance. Significantly, within the ILE paradigm, the present author is currently involved in a project for the refining of WISER, given its successful classification performance, as well as its adaptation to Spanish from English.

As a further proposal for future research, a deeper comparison and analysis of other existing methods for deception detection on the same dataset could strengthen the contributions of the newly introduced predictors, as in the outstanding case of semicolon.

All things considered, the use of corpus tools developed out of linguistic theory is of the utmost importance as is the adoption of reliable scientific methods. Researchers should keep on testing methods on real life data, deploying their knowledge of linguistics—theory, corpus linguistics, and computational linguistics—to improve both low-stakes and high-stakes deception detection.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Universidad de Murcia.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: According to suggested Data Availability Statements in section “MDPI Research Data Policies” at <https://www.mdpi.com/ethics>, the dataset has been deposited by the present author in a publicly available database, namely <https://github.com/angelalm/DeceptionCorpus>.

Acknowledgments: I would like to express my gratitude to the anonymous referees for their careful review and insightful comments. Furthermore, I am also grateful to Carole E. Chaski, PhD for critically reading a previous version of this manuscript and stimulating discussions during the preparation of this article.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Selection of redundant LIWC categories for the experiment.

I. Linguistic dimensions		
TOTAL PRONOUNS	Total 1st person	<i>1st person singular</i>
		<i>1st person plural</i>
	<i>Total 2nd person</i>	-
	<i>Total 3rd person</i>	-
II. Psychological processes		
AFFECTIVE OR EMOTIONAL PROCESSES	Positive emotions	<i>Positive feelings</i>
		<i>Optimism and energy</i>
		<i>Anxiety or fear</i>
	Negative emotions	<i>Anger</i>
		<i>Sadness or depression</i>
COGNITIVE PROCESSES	<i>Causation</i>	-
	<i>Insight</i>	-
	<i>Discrepancy</i>	-
	<i>Inhibition</i>	-
	<i>Tentative</i>	-
	<i>Certainty</i>	-
SENSORY AND PERCEPTUAL PROCESSES	<i>Seeing</i>	-
	<i>Hearing</i>	-
	<i>Feeling</i>	-
SOCIAL PROCESSES	<i>Communication</i>	-
		<i>1st person plural</i>
	Other references to people	<i>Total 2nd person</i>
		<i>Total 3rd person</i>
	<i>Friends</i>	-
	<i>Family</i>	-
	<i>Humans</i>	-
III. Relativity		
TIME	<i>Past tense verb</i>	-
	<i>Present tense verb</i>	-
	<i>Future tense verb</i>	-
SPACE	<i>Up</i>	-
	<i>Down</i>	-
	<i>Inclusive</i>	-
	<i>Exclusive</i>	-

Table A1. Cont.

IV. Personal concerns		
OCCUPATION	<i>School</i>	-
	<i>Job or work</i>	-
	<i>Achievement</i>	-
LEISURE ACTIVITY	<i>Home</i>	-
	<i>Sports</i>	-
	<i>Television and movies</i>	-
	<i>Music</i>	-
	<i>Money and financial issues</i>	-
METAPHYSICAL ISSUES	<i>Religion</i>	-
	<i>Death and dying</i>	-
PHYSICAL STATES AND FUNCTIONS	<i>Body states, symptoms</i>	-
	<i>Sex and sexuality</i>	-
	<i>Eating, drinking, dieting</i>	-
	<i>Sleeping, dreaming</i>	-
	<i>Grooming</i>	-
	<i>Swearing</i>	-

Appendix B

Table A2. Variables in the experiment.

Variables	Class
Word count	LIWC
Words per sentence	LIWC
Words longer than 6 letters	LIWC
Period	LIWC
Comma	LIWC
Colon	LIWC
Semicolon	LIWC
Sentences ending with ‘?’	LIWC
Exclamation	LIWC
Dash	LIWC
Quote	LIWC
Apostrophe	LIWC
Parenthesis	LIWC
Other punctuation	LIWC
1st person singular	LIWC
1st person plural	LIWC
2nd person	LIWC
3rd person	LIWC
Negations	LIWC
Assents	LIWC
Articles	LIWC
Prepositions	LIWC
Numbers	LIWC
Positive feelings	LIWC
Optimism and energy	LIWC
Anxiety or fear	LIWC
Anger	LIWC
Sadness or depression	LIWC

Table A2. Cont.

Variables	Class
Causation	LIWC
Insight	LIWC
Discrepancy	LIWC
Inhibition	LIWC
Tentative	LIWC
Certainty	LIWC
Seeing	LIWC
Hearing	LIWC
Feeling	LIWC
Communication	LIWC
Friends	LIWC
Family	LIWC
Humans	LIWC
Past tense verb	LIWC
Present tense verb	LIWC
Future tense verb	LIWC
Up	LIWC
Down	LIWC
Inclusive	LIWC
Exclusive	LIWC
Motion	LIWC
School	LIWC
Job or work	LIWC
Achievement	LIWC
Home	LIWC
Sports	LIWC
Television and movies	LIWC
Music	LIWC
Money and financial issues	LIWC
Religion	LIWC
Death and dying	LIWC
Body states, symptoms	LIWC
Sex and sexuality	LIWC
Eating, drinking, dieting	LIWC
Sleeping, dreaming	LIWC
Grooming	LIWC
Swearing	LIWC
Standardized type/token ratio	Styl.
Mean word length	Styl.
Sentences/WC	Styl.
1-letter words/WC	Styl.
2-letter words/WC	Styl.
3-letter words/WC	Styl.
4-letter words/WC	Styl.
5-letter words/WC	Styl.
6-letter words/WC	Styl.
7-letter words/WC	Styl.
Complex words/WC	Styl.

Appendix C

Table A3. Random sample 1 of truthful and untruthful statements in Spanish.

TRUTH	LIE
HOMOSEXUAL ADOPTION	
<i>Para mí no está clara la repercusión que tendría sobre los niños el hecho de que las parejas homosexuales adopten. Sería necesario un estudio previo de las posibles consecuencias o secuelas psicológicas, o de la ausencia de ellas, en el mejor de los casos.</i>	<i>La familia es y ha sido siempre la formada por un hombre y una mujer. No debemos cambiar esto, pues es un claro síntoma de la degeneración de la sociedad. Hemos de defender las tradiciones que llevan funcionando bien durante miles de años.</i>
Translation into English:	Translation into English:
It is not clear to me what the repercussions would be for children if homosexual couples were to adopt. A prior study of the possible psychological consequences or sequelae, or the absence of them at best, would be necessary.	The family is and has always been the one formed by a man and a woman. We must not change this, as it is a clear symptom of the degeneration of society. We must defend the traditions that have been working well for thousands of years.
BULLFIGHTING	
<i>Es una salvajada. Regodearse en el sufrimiento de un animal, disfrutar viendo cómo realiza sus últimos movimientos, agotado y herido. ¿Cómo puede ser un arte esto? Sin duda hay muchas personas que están familiarizadas con las corridas de toros. Es para ellos una situación normal.</i>	<i>Los espectáculos relacionados con los toros son una tradición antiquísima y un arte. Es más, los toros de lidia se pasan la vida al aire libre y son bien mimados por sus criadores, disfrutando así de una vida muchísimo mejor que la que se les ofrece a los animales de granja.</i>
Translation into English:	Translation into English:
It is a savagery. To wallow in the suffering of an animal, to enjoy watching it make its last movements, exhausted and wounded. How can this be art? Undoubtedly, there are many people who are familiar with bullfighting. For them, it is a normal situation.	Bullfighting shows are an ancient tradition and an art. Moreover, fighting bulls spend their lives outdoors and are well pampered by their breeders, enjoying a much better life than that offered to farm animals.
GOOD FRIEND	
<i>Cuando conocí a José María pensé que era uno más, que incluso no nos podríamos llevar bien. Qué equivocación más grande, ¡y qué afortunada! Es hoy uno de mis mejores amigos, que me encontré de casualidad en una de mis muchas andanzas por el mundo.</i>	<i>Sergio es un chaval inteligente, que sabe lo que quiere. Es realmente una buena persona, con la que puedes contar para todo. Su principal cualidad es su simpatía y amabilidad con todos, no importa que no te conozca de nada, siempre te da una oportunidad.</i>
Translation into English:	Translation into English:
When I first met José María I thought he was just another guy, and that we might not even get along. What a big mistake, and how fortunate! Today he is one of my best friends, whom I met by chance in one of my many wanderings around the world.	Sergio is an intelligent guy, who knows what he wants. He is a really good person, you can count on him for everything. His main quality is his sympathy and kindness with everyone, it doesn't matter if he doesn't know you at all, he always gives you a chance.

Table A4. Random sample 2 of truthful and untruthful statements in Spanish.

TRUTH	LIE
HOMOSEXUAL ADOPTION	
<i>Yo pienso que es un tema muy delicado y tal vez ahora mismo los hijos de parejas homosexuales podrían ser discriminados en el colegio, tendrá que cambiar la sociedad poco a poco pero aun así pienso que es importante tener un referente masculino y otro femenino en la educación de un niño.</i>	<i>Me gustaría decir estoy cansado de las discriminaciones que sufren las parejas homosexuales en la sociedad hoy en día. Son parejas como cualquier otra y sienten lo mismo que las demás. Por lo tanto pienso que sería correcto que pudieran adoptar ya que querrían a su hijo de la misma manera que las parejas heterosexuales. El respeto a los demás y la tolerancia es uno de los valores centrales de la educación en una familia.</i>
Translation into English:	Translation into English:
I think it is a very delicate issue and maybe right now the homosexual couples' children could be discriminated at school; society will have to change little by little, but I still think it is important to have a male and female reference in the education of a child.	I would like to say that I am tired of the discrimination that homosexual couples suffer in today's society. They are couples like any other and feel the same as others. Therefore, I think it would be right for them to be able to adopt since they would love their child in the same way as heterosexual couples. Respect for others and tolerance is one of the core educational values in a family.
BULLFIGHTING	
<i>El animal agoniza en una sopa de sangre, siente miedo, dolor, angustia, desesperación. No tiene posibilidades reales de defenderse, no tiene noción de lo que sucede a su alrededor, no tiene capacidad de razonar y por ende, de imaginarse cuándo cesarán todas esas desagradables sensaciones. El toro no lucha por su vida. Es sometido a una serie de torturas sistemáticas que lo humillan, lo denigran y lo hacen padecer infinito dolor.</i>	<i>Los toros y las corridas como acto o evento social me parece algo que está hace muchísimos años y da de comer a muchísimas familias, a pesar que dicen que es cruento, piensen que si se quitaran las corridas mucha gente quedaría en paro y lo más señalado es que nos comeríamos los toros igualmente, así que no es interesante el acabar con la famosa fiesta taurina y algo más, ¿toda la carne que comemos todos que pasa? ¿Es sintética?</i>
Translation into English:	Translation into English:
The animal dies in a soup of blood, feels fear, pain, anguish, despair. It has no real possibility of defending itself, it has no notion of what is happening around it, it has no capacity to reason and, therefore, to imagine when all these unpleasant sensations will cease. The bull does not fight for its life. It is subjected to a series of systematic tortures that humiliate it, denigrate it and make it suffer from infinite pain.	Bullfighting as a social act or event seems to me something that has been around for many years and feeds many families, even though they say it is cruel; think that if bullfighting were banned, many people would be unemployed, and the most important issue is that we would eat bulls anyway, so it is not interesting to ban the famous bullfighting tradition, and something else: What happens with the meat that we all eat? Is it synthetic?
GOOD FRIEND	
<i>Mi mejor amigo es la persona con la que paso prácticamente todo mi tiempo libre. Es la persona con la que siempre puedo contar, sea cual sea el problema que tenga. Siempre solemos tener los mismos gustos y aficiones. Nos conocemos desde el colegio y a pesar de los años siempre hemos mantenido una amistad, aunque durante los dos últimos años está siendo mi prioridad. Espero que no se acabe nunca.</i>	<i>Mi amigo X es una de esas personas con las que siempre te lo pasas bien, tiene una gran capacidad para hacerte sentir bien y que eres especial. Es una persona muy sociable y abierta con todo el mundo. Aunque si hay una cualidad que lo distingue es su fidelidad y confianza.</i>
Translation into English:	Translation into English:
My best friend is the person I spend practically all my free time with. He is the person I can always count on, no matter what problem I have. We always tend to have the same tastes and hobbies. We have known each other since school and, despite the years, we have always maintained a friendship, although for the last two years he has been my priority. I hope it never ends.	My friend X is one of those people with whom you always have a good time, he has a great ability to make you feel good and feel that you are special. He is a very sociable and open person with everyone. Although if there is one quality that distinguishes him it is his loyalty and trust.

References

- Ott, M.; Choi, Y.; Cardie, C.; Hancock, J.T. Finding deceptive opinion spam by any stretch of the imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 309–319.
- Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, J.; Camacho-Collados, M. Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police. *Knowl. Based Syst.* **2018**, *149*, 155–168. [\[CrossRef\]](#)
- Vogler, N.; Pearl, L. Using linguistically defined specific details to detect deception across domains. *Nat. Lang. Eng.* **2019**, *26*, 349–373. [\[CrossRef\]](#)
- Chaski, C.E.; Almela, A.; Holness, G.; Barksdale, L. WISER: Automatically Classifying Written Statements as True or False. Oral communication presented. In Proceedings of the American Academy of Forensic Sciences 67th Annual Scientific Meeting, Orlando, FL, USA, 16–21 February 2015; pp. 576–577.
- Picornell, I. Cues to Deception in a Textual Narrative Context: Lying in Written Witness Statements. Ph.D. Dissertation, Aston University, Birmingham, UK, 2012.
- Meibauer, J. (Ed.) *The Oxford Handbook of Lying*; Oxford University Press: Oxford, UK, 2018.
- Markowitz, D.M.; Hancock, J.T. Deception and Language: The Contextual organization of Language and Deception (CoLD) framework. In *The Palgrave Handbook of Deceptive Communication*; Docan-Morgan, T., Ed.; Palgrave Macmillan: New York, NY, USA, 2019; pp. 193–212.
- Bull, R.; Cook, C.; Hatcher, R.; Woodhams, J.; Bilby, C.; Grant, T. *Criminal Psychology: A Beginner's Guide*; Oneworld Publications: Oxford, UK, 2006.
- Sporer, S.L.; Manzanero, A.L.; Masip, J. Optimizing CBCA and RM research: Recommendations for analyzing and reporting data on content cues to deception. *Psychol. Crime Law* **2020**, *27*, 1–39. [\[CrossRef\]](#)
- Fitzpatrick, E.; Bachenko, J. Detecting Deception across Linguistically Diverse Text Types. In Proceedings of the Linguistic Society of America Annual Meeting, Boston, MA, USA, 3–6 January 2013.
- Chaski, C.E. Author Identification in the Forensic Setting. In *The Oxford Handbook of Language and Law*; Solan, L.M., Tiersma, P.M., Eds.; Oxford University Press: Oxford, UK, 2012.
- Chaski, C.E. Empirical Evaluations of Language-based Author Identification Techniques. *Forensic Linguist.* **2001**, *8*, 1–66. [\[CrossRef\]](#)
- Chaski, C.E. Best Practices and Admissibility of Forensic Author Identification. *J. Law Policy* **2013**, *21*, 233.
- Zhou, L.; Burgoon, J.K.; Nunamaker, J.F.; Twitchell, D. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decis. Negot.* **2004**, *13*, 81–106. [\[CrossRef\]](#)
- Mihalcea, R.; Strapparava, C. The lie detector: Explorations in the automatic recognition of deceptive language. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Association for Computational Linguistics, Singapore, 4 August 2009; pp. 309–312.
- Newman, M.; Pennebaker, J.; Berry, D.; Richards, J. Lying words: Predicting deception from linguistic styles. *Personal. Soc. Psychol. Bull.* **2003**, *29*, 665–675. [\[CrossRef\]](#)
- Almela, Á.; Alcaraz-Mármol, G.; Cantos, P.Y. Analysing deception in a psychopath's speech: A quantitative approach. *DELTA Doc. Estud. Lingüíst. Teór. Apl.* **2015**, *31*, 559–572. [\[CrossRef\]](#)
- Fornaciari, T.; Poesio, M. Automatic deception detection in Italian court cases. *Artif. Intell. Law* **2013**, *21*, 303–340. [\[CrossRef\]](#)
- Feng, S.; Banerjee, R.; Choi, Y. Syntactic stylometry for deception detection. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Jeju Island, Korea, 8–14 July 2012; pp. 171–175.
- Pérez-Rosas, V.; Mihalcea, R. Experiments in open domain deception detection. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 17–21 September 2015; pp. 1120–1125.
- Yancheva, M.; Rudzicz, F. Automatic detection of deception in child-produced speech using syntactic complexity features. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 4–9 August 2013; pp. 944–953.
- Almela, A.; Valencia-García, R.; Cantos, P. Seeing through Deception: A Computational Approach to Deceit Detection in Spanish Written Communication. In Proceedings of the Workshop on Computational Approaches to Deception Detection, Association for Computational Linguistics, Avignon, France, 23 April 2012; pp. 15–22.
- Rubin, V.L.; Vashchilko, T. Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. In Proceedings of the Workshop on Computational Approaches to Deception Detection, Association for Computational Linguistics, Avignon, France, 23 April 2012; pp. 97–106.
- Hauch, V.; Blandón-Gitlin, I.; Masip, J.; Sporer, S.L. Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. *Personal. Soc. Psychol. Rev.* **2015**, *19*, 307–342. [\[CrossRef\]](#)
- Stone, P.J.; Bales, R.F.; Namenwirth, J.Z.; Ogilvie, D.M. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behav. Sci.* **1962**, *7*, 484–494. [\[CrossRef\]](#)

26. Stone, P.J.; Dunphy, D.; Smith, M.S.; Ogilvie, D.M. *The General Inquirer: A Computer Approach to Content Analysis*; MIT Press: Cambridge, MA, USA, 1966.
27. Knapp, M.L.; Hart, R.P.; Dennis, H.S. An exploration of deception as a communication construct. *Hum. Commun. Res.* **1974**, *1*, 15–29. [\[CrossRef\]](#)
28. Pennebaker, J.W.; Francis, M.E.; Booth, R.J. *Linguistic Inquiry and Word Count*; Erlbaum Publishers: Mahwah, NJ, USA, 2001.
29. Pennebaker, J.W.; Graybeal, A. Patterns of natural language use: Disclosure, personality, and social integration. *Curr. Dir. Psychol. Sci.* **2001**, *10*, 90–93. [\[CrossRef\]](#)
30. Ramírez-Esparza, N.; Pennebaker, J.W.; García, F.A.; Suriá, R. La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Rev. Mex. Psicol.* **2007**, *24*, 85–99.
31. Hunt, D.; Brookes, G. *Corpus, Discourse and Mental Health*; Bloomsbury Publishing: London, UK, 2020.
32. Salas-Zárate, M.P.; López-López, E.; Valencia-García, R.; Aussenac-Gilles, N.; Almela, A.; Alor-Hernández, G. A study on LIWC categories for opinion mining in Spanish reviews. *J. Inf. Sci.* **2014**, *40*, 749–760. [\[CrossRef\]](#)
33. Almela, A.; Alcaraz-Mármol, G.; Garcia, A.; Pallejá-López, C. Developing and Analyzing a Spanish Corpus for Forensic Purposes. *LES LI Linguist. Evid. Secur. Law Intell.* **2019**, *3*, 1–13. [\[CrossRef\]](#)
34. Graesser, A.C.; McNamara, D.S.; Louwerse, M.M.; Cai, Z. Coh-Metrix: Analysis of text on cohesion and language. *Behav. Res. Methods Instrum. Comput.* **2004**, *36*, 193–202. [\[CrossRef\]](#)
35. McNamara, D.S.; Graesser, A.C.; McCarthy, P.M.; Cai, Z. *Automated Evaluation of Text and Discourse with Coh-Metrix*; Cambridge University Press: Cambridge, MA, USA, 2014.
36. Bedwell, J.S.; Gallagher, S.; Whitten, S.N.; Fiore, S.M. Linguistic correlates of self in deceptive oral autobiographical narratives. *Conscious. Cogn.* **2011**, *20*, 547–555. [\[CrossRef\]](#)
37. Sapir, A. *Scientific Content Analysis (SCAN)*; Laboratory of Scientific Investigation: Phoenix, AZ, USA, 1987.
38. Lesce, T. SCAN: Deception Detection by Scientific Content Analysis. *Law Order* **1990**, *38*, 8. Available online: <http://www.lsiscan.com/id37.htm> (accessed on 25 November 2020).
39. McClish, M. *I Know You Are Lying. Detecting Deception through Statement Analysis*; The Marpa Group, Inc.: Winterville, GA, USA, 2001.
40. Fitzpatrick, E.; Bachenko, J.; Fornaciari, T. *Automatic Detection of Verbal Deception*; Morgan and Claypool Publishers: Williston, VT, USA, 2015. [\[CrossRef\]](#)
41. Adams, S.H.; Jarvis, J.P. Indicators of veracity and deception: An analysis of written statements made to police. *Speech Lang. Law* **2006**, *13*, 1–22. [\[CrossRef\]](#)
42. Kang, S.M.; Lee, H. Detecting deception by analyzing written statements in Korean. *Linguist. Evid. Secur. Law Intell.* **2014**, *2*, 1–10. [\[CrossRef\]](#)
43. Fuller, C.M.; Biros, D.P.; Burgoon, J.K.; Adkins, M.; Twitchell, D.P. An analysis of text-based deception detection tools. In Proceedings of the 12th Americas Conference on Information Systems, Acapulco, Mexico, 4–6 August 2006; pp. 3465–3472.
44. Zhou, L.; Booker, Q.E.; Zhang, D. ROD: Towards rapid ontology development for underdeveloped domains. In Proceedings of the 35th Hawaii International Conference on System Sciences, Honolulu, HI, USA, 10 January 2002; pp. 957–965. [\[CrossRef\]](#)
45. Derrick, D.; Meservy, T.; Burgoon, J.; Nunamaker, J. An experimental agent for detecting deceit in chat-based communication. In *Proceedings of the Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium*; Jensen, M., Meservy, T., Burgoon, J., Nunamaker, J., Eds.; Grand Wailea: Maui, HI, USA, 2012; pp. 1–21. [\[CrossRef\]](#)
46. Chaski, C.E.; Barksdale, L.; Reddington, M.M. Collecting Forensic Linguistic Data: Police and Investigative Sources of Data for Deception Detection Research. In Proceedings of the Linguistic Society of America Annual Meeting, Minneapolis, MN, USA, 2–5 January 2014.
47. Harris, Z. Distributional Structure. *Word* **1954**, *10*, 146–162. [\[CrossRef\]](#)
48. Salton, G.; McGill, M. *Introduction to Modern Information Retrieval*; McGraw-Hill: New York, NY, USA, 1983.
49. Cantos, P.; Almela, A. Readability indices for the assessment of textbooks: A feasibility study in the context of EFL. *Vigo Int. J. Appl. Linguist.* **2019**, *16*, 31–52. [\[CrossRef\]](#)
50. Shadish, W.R.; Cook, T.D.; Campbell, D.T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*; Houghton Mifflin Company: Boston, MA, USA, 2002.
51. Chipere, N.; Malvern, D.; Richards, B.J. Using a corpus of children’s writing to test a solution to the sample size problem affecting Type-Token Ratios. In *Corpora and Language Learners*; Aston, G., Bernardini, S., Stewart, D., Eds.; John Benjamins: Amsterdam, The Netherlands, 2004; pp. 139–147.
52. Kline, P. *A Handbook of Test Construction*; Methuen: New York, NY, USA, 1986.
53. Berber-Sardinha, T.; Veirano, M. (Eds.) *Multidimensional Analysis*; Bloomsbury Publishing: London, UK, 2019.
54. Cantos, P. *Statistical Methods in Language and Linguistic Research*; Equinox: London, UK, 2013.
55. Tabachnick, B.G.; Fidell, L.S. *Using Multivariate Statistics, New International Edition*, 6th ed.; Pearson Education Limited: Harlow, UK, 2013.
56. Molinaro, A.M.; Simon, R.; Pfeiffer, R.M. Prediction error estimation: A comparison of resampling methods. *Bioinformatics* **2005**, *21*, 3301–3307. [\[CrossRef\]](#)
57. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*; Morgan Kaufmann: San Mateo, CA, USA, 1995; pp. 1137–1143.

-
58. DePaulo, B.M.; Lindsay, J.J.; Malone, B.E.; Muhlenbruck, L.; Charlton, K.; Cooper, H. Cues to deception. *Psychol. Bull.* **2003**, *129*, 74–118. [[CrossRef](#)]
 59. Masip, J.; Bethencourt, M.; Lucas, G.; Sánchez-San Segundo, M.; Herrero, C. Deception detection from written accounts. *Scand. J. Psychol.* **2012**, *53*, 103–111. [[CrossRef](#)]
 60. Burgoon, J.K.; Blair, J.P.; Qin, T.; Nunamaker, J.F. Detecting deception through linguistic analysis. *Intell. Secur. Inform.* **2003**, *2665*, 91–101. [[CrossRef](#)]
 61. Vivancos-Vicente, P.J.; García-Díaz, J.A.; Almela, A.; Molina, F.; Castejón-Garrido, J.A.; Valencia-García, R. Transcripción, indexación y análisis automático de declaraciones judiciales a partir de representaciones fonéticas y técnicas de lingüística forense. *Proces. Leng. Nat.* **2020**, *65*, 109–112.