Sincere and deceptive statements in Italian criminal proceedings

Tommsso Fornaciari[1] and Massimo Poesio[2]
University of Trento, Italy
[1]*tommaso.fornaciari@unitn.it*
[2]*massimo.poesio@unitn.it*

Abstract

Identifying false or deceptive statements in testimonies is a difficult challenge in criminal proceedings because it is not a task humans find easy. Text classification techniques have shown promise at this task—but so far, they have mainly been tested with laboratory produced data rather than authentic, real life data. We collected what is the first Italian corpus of hearings from criminal proceedings in which the defendant was found guilty of false testimony. In such cases, the transcriptions of each hearing report the words exactly as told by the subjects, and the judgment points out the statements found by the Court to be false. This characteristic makes it possible to annotate sincerity and deception of statements in such data on the basis of unusually solid objective information. We used these data to train models to classify statements as sincere or deceptive, showing that in spite of the difficulty humans have at this classification task, it is possible to obtain a performance well above chance level from automatic classifiers using very simple surface linguistic features.

Keywords: FORENSIC LINGUISTICS; DECEPTION DETECTION; TESTIMONY IN COURT; TEXT CLASSIFICATION

1. Introduction

1.1. Detecting deception

Identifying deceptive statements in testimonies could provide very useful support to investigative work, particularly when other kinds of evidence are scarce or absent. In spite of this, modern studies demonstrate that human performance in recognizing deception is not much better than chance (Bond and De Paulo, 2006). Furthermore, in some studies human skills seem to be not particularly improved even after specific training (Levine, Feeley, McCornack, Hughes, and Harms, 2005). Other studies instead try to demonstrate that the ability of humans as lie-detectors is underestimated (Frank and Feeley, 2003). In any case, even in papers in which positive effects of training are found, the difficulty of the task is openly recognized (Porter, Woodworth, and Birt, 2000).

Probably the difficulty in recognizing deceptive statements has led to the development of a wide variety of approaches to discover deceptive statements. They can be very different from each other, but all of them typically involve two steps:

- To identify some clues of deceptiveness in the communicative act;
- To verify if the statements held as false are actually false.

The choice of clues to be considered in the analysis determines the strategy in trying to detect deception. Several authors try to put together different analysis techniques, hoping to optimize the accuracy in detecting falsehoods. This is the case with De Paulo, Lindsay, Malone, Muhlenbruck, Charlton, and Cooper (2003), who consider more than 150 cues, verbal and non-verbal, directly observed through experimental subjects. Also Jensen, Meservy, Burgoon, and Nunamaker (2010) recently focused on cues coming from audio, video and textual data, with the aim of building a paradigm useful to identify deceptiveness.

1.2. Stylometry

With the contribution of modern linguistics and psychology, the analysis of language has become increasingly effective and has been applied to specific aspects of the discourse. In recent years stylometric methods which typically analyse linguistic style in text through statistical techniques, have in fact been demonstrated to be effective in several forensic tasks, such as author profiling (Coulthard, 2004; Solan and Tiersma, 2004), including deducing the age and sex of authors of written texts (Koppel, Schler, Argamon, and Pennebaker, 2006), author attribution (Luyckx and Daelemans, 2008; Mosteller and Wallace, 1964) and plagiarism analysis (Stein, Koppel, and Stamatatos, 2007). Stylometry is also becoming more and more important in Deceptive Language Analysis.

Stylistic features have emerged as useful markers to evaluate the truthfulness of the speakers (or writers). A lot of studies have been carried out following this path (for example Porter and Yuille, 1996), in a variety of contexts. For example Adams (1996), working in the context of Police Forces, asserted the necessity to take into account the personal style of communication together with the content of the testimonies. In Italy Anolli, Balconi, and Ciceri, (1999), working on Italian linguistic data, tried to identify styles of communication which are specific to deceptive language.

1.3. Deceptive language analysis

1.3.1. Field and laboratory studies

Regarding deceptive language, the existing papers can be roughly divided in two main families: field studies and laboratory studies. Field studies, such as those using Criteria Based Content Analysis (Vrij, 2005), one of the foremost techniques for the evaluation of children's statements in cases of suspected sexual abuse, are interesting for forensic practice but, as noticed by Vrij himself (2005), it is often difficult to verify the sincerity of the statements. Typically, in practical cases the content of the testimonies themselves and non-verbal cues play an important role in the assessment of sincerity. Such methods of research are quite different from those employed in our paper, which relies instead on stylometric analyses.

Laboratory studies (Newman, Pennebaker, Berry, and Richards, 2003), on the other hand, focus on mock lies, produced by experimental subjects under laboratory conditions. These studies result in the creation of balanced data sets that typically allow stylometric analyses through machine learning algorithms. Nevertheless, the artificiality of the conditions means that the findings of such studies may not be applicable to real life cases.

As Koppel et al. (2006) point out, the features used in stylometric analysis belong to two main families:

- surface-related features; and
- content-related features.

The first type of feature includes the frequency and use of function words or of certain *n*-grams of words or part-of-speech (POS). The second kind of feature specifies information about the semantic content of words, accessed from dictionaries and lexical resources.

1.3.2. The Linguistic Inquiry and Word Count (LIWC)

Perhaps the best-known lexical resource for deception detection is the Linguistic Inquiry and Word Count (LIWC), created by Pennebaker, Francis, and Booth (2001). This was applied, among other things, to the evaluation of deceptive language. For example Newman et al. (2003) reached an overall accuracy of 60% in classifying deceptive vs. truthful texts. In addition LIWC has been employed in studies on deceptive language carried out by other groups, such as the work by Strapparava and Mihalcea (2009), who obtained results similar to Newman et al. (2003) at classifying into ‘sincere’ or ‘deceptive’ texts collected with the Amazon Mechanical Turk service. Strapparava and Mihalcea actually used surface features in order to classify their texts, but also used the LIWC, even if for post-hoc analysis only, to measure several language dimensions, such as positive or negative emotions, cognitive processes, and so on. In this way, they were able to identify some distinctive characteristics of deceptive texts.

Moreover, the opportunity to work with data in electronic format, and the increasing relevance of Computer Mediated Communication, has contributed to an increase of studies in which deceptiveness is produced through the use of computers: Hancock, Curry, Goorha, and Woodworth (2008), for example, employed LIWC for research about dyadic communication in a synchronous text-based setting. Making use of different variables, Zhou carried out an analogous study of both synchronous (2005), and asynchronous (Zhou, Burgoon, Nunamaker, and Twitchell, 2004) Computer Mediated Communication.

1.4. Our research

Our paper also aims to develop machine learning models of deception detection. However, we aim also to fill a research gap identified most recently by Zhou, Shi, and Zhang (2008), who highlighted the lack of ‘data sets for evaluating deception detection models’ (p. 1078). Our goal is to contribute to research knowledge by analyzing transcriptions of false and true testimonies presented during Court hearings and to distinguish true testimony from false on the basis of stylometric differences.

The theoretical assumption on which this paper is based, historically known as the Undeutsch hypothesis (1967), is that the cognitive elaboration of untruthful statements differs from the elaboration of truthful ones, so that differences should be traceable in the features of the statements themselves. In order to study this hypothesis it is necessary, on one hand, to collect testimonies containing real life linguistic data; and on the other, to know with certainty if statements are sincere or deceptive.

There currently exists a lack of research in which both of these prerequisites are satisfactorily met. With respect to the kind of data collected, the two studies of Fitzpatrick's group (Bachenko, Fitzpatrick, and Schonwetter, 2008; Fitzpatrick and Bachenko, 2009) are the most similar to our research activities. They collected a corpus of criminal statements, police interrogations, and civil testimony. On the other hand, as deception cues to analyze their data they choose several ‘linguistic phenomena’ such as preference for negative expressions in word choice, inconsistencies between verb and noun forms and so on, and their texts were annotated manually on the basis of these ‘phenomena’. They obtained accuracy close to 75% in the classification task. We believe we are the first to apply a stylometric approach to Italian language to detect deception.

The structure of the paper is as follows: In Section 2 we discuss the method used to collect the data for our study. In Section 3 we discuss the methods used to build the models and in Section 4 we present our results.

2. The data

2.1. Finding suitable data

In criminal proceedings, investigators interview numerous witnesses, who can produce true or false statements. In many cases the investigators do not know which statement is true or false, and in most cases the transcripts of these testimonies do not reproduce *verbatim* what the subjects said. Instead, they are simply a synthesis of the witnesses' declarations, carried out by the police officer who produces the transcript. Such reports are not a faithful mirror of the linguistic behavior of the subjects; therefore they are not useful from the purposes of the present paper.

However in Italy there is a specific case of testimony that is reported *verbatim*: hearings that take place during a debate in front of the judge. Focusing on this aspect of the criminal procedure is therefore the most promising way of studying deception production. Furthermore, to focus on the debate is a convenient choice from the point of view of the homogeneity of data. It is an event strongly ritualized, in which actors and acts recur in a standard way, and it guarantees a certain regularity of conditions in different hearings.

In addition, there is a type of criminal proceedings in which the truthfulness or deceptiveness of testimonies is easily verifiable. This is the case of criminal proceedings concerning violations of articles 368 and 372 of the Italian Criminal Code[1] that codify the crimes of ‘calumny’ and ‘false testimony’, respectively. They are typically proceedings that originate when statements, issued in hearings related to any crime, are found unreliable, and therefore the statements themselves become the object of a further criminal proceeding for ‘calumny’ or ‘false testimony’. Because these proceedings are related to the lies, necessarily they end with a judgment that points out in a certain, organic and exhaustive way, the lies told by the defendant.

2.2. Data collection

Our first step was to contact the Courts in several Italian towns, in order to receive authorization to examine their dossiers and extract information from them for research purposes. The three Presidents of Court to which the research project has been presented, allowed the collection of the data, with the restriction of publishing them in anonymous form, respecting the privacy of the subjects involved.

This paper is based on the data collected in the Courts of Trento, Bolzano and Prato. Eighteen hearings with false testimonies were identified, issued by a total of seventeen subjects, one of whom was interrogated twice, who appeared in the hearings as defendant, witness or expert witness.

[1] To be precise, art. 368 reads: Chiunque, con denunzia, querela, richiesta o istanza, anche se anonima o sotto falso nome, diretta all'Autorità giudiziaria o ad altra Autorità che a quella abbia obbligo di riferirne, incolpa di un reato taluno che egli sa innocente, ovvero simula a carico di lui le tracce di un reato, è punito con la reclusione da due a sei anni. In brief, it punishes whoever tries to charge the responsibility of some crime on someone who he knows is innocent.
Art. 372 instead reads: Chiunque, deponendo come testimone innanzi all'Autorità giudiziaria, afferma il falso o nega il vero, ovvero tace, in tutto o in parte ciò che sa intorno ai fatti sui quali è interrogato, è punito con la reclusione da due a sei anni. This article punishes someone who, in front of the Judicial Authority, says a falsity or denies the truth, or does not reveal what he knows about the investigated facts.

2.3. Pre-processing

Each transcript was converted to XML format according to a coding scheme where each intervention of the heard subject, in between the interventions of some other individual, is classified as a *turn*. Each turn can be constituted by one or more *utterances*—delimited by terminal punctuation marks—that are the main units of analysis for this paper. This resulted in 1437 utterances available for analysis from the complete corpus of 18 transcripts.

Each utterance of a witness was assigned a label that specifies the truthfulness or truthlessness of the utterance itself. This annotation was carried out by hand, on the basis of information found in the Court's judgment relative to the testimony. Between the white of the truth and the black of the falsity, however, there are wide gradations of gray, and the judgment that describes the facts and points out the lies told, cannot specify the truth of each statement issued in the courtroom. To label the utterances is therefore a complex task, as discussed in the following annotation scheme:

- *'False'*: the utterance is clearly pointed out in the judgment as false, or the falsity is a logic consequence of some ascertained lie.
- *'True'*: the utterances that are coherent with the reconstruction of the facts contained in the judgment are considered true. Also the utterances that explain something not considered in the judgment because they not influential with respect to the investigated facts, are generally considered true.
- *'Not reliable'*: an utterance is considered not reliable if it is related to the investigated facts, but the judgment does not prove its deceptiveness.
- *'True or not reliable'*: like the 'not reliable' utterances, the 'true or not reliable' ones are related to the topic of investigation, and the judgment demonstrates nothing about them. Nevertheless, according to the event and to other statements certainly true or false, and/or on the basis of a weak connection with the interests that the subject tries to defend, it is logical to suppose that they are probably true. In brief, according to common sense, those utterances should be true, but the fact is not demonstrated, and ultimately questionable.
- *'False or not reliable'*: this is the specular situation with respect to the previous point. According to the interests of the subjects, and to the economy of the event and of the testimony, it is reasonable, but not demonstrated, that these utterances are false. In these cases, the final evaluation is not certain, and a note is made about the 'hue' of the statement.
- *'Undecidable'*: the utterances that, from a logical point of view, cannot be either true or false, are considered undecidable. This is the case for many questions (like 'Excuse me, can you repeat?'), but also for several utterances stopped in mid-sentence, that do not have a complete sense. This is also the case for utterances that have a meta-communicative function, and regulate the relations between actors, like 'Now I'll explain.' or 'If you think so...' and so on.

The corpus was tokenized and anonymized in accordance with the agreements with the Courts. Stop words were not removed: on the contrary, in stylometric analysis function words are considered crucial. Blocks of punctuation marks were considered as one token. For example, a single comma was considered a token, and three suspension points were also considered a single token. Finally, the corpus was lemmatized and POS-tagged using a version of TreeTagger[2] (Schmid, 1994) trained for Italian.

While the utterances of other participants in the courtroom are not considered, the 1437 utterances of the heard subjects have been labeled according to this coding scheme. Corpus statistics are provided in Table 1.

Table 1: Corpus statistics

Label	Utterances	Tokens	
		with punct.	without punct.
False	333	5778	4802
True	537	7908	6628
Not reliable	225	3351	2746
True or not reliable	83	1758	1452
False or not rel.reliable	78	1648	1360
Undecidable	181	1146	886
Total	1437	21589	17874

The utterances labeled as ‘True or not reliable’, ‘False or not reliable’, ‘Not reliable’ and ‘Undecidable’ have been discarded, and the analyses concern only the part of the corpus constituted by ‘True’ and ‘False’ utterances: a total of 870.

3. Methods

3.1. Features

Each utterance was described as a feature vector. The features come from a training set made by ten of our eighteen hearings. This subset of hearings provides 623 utterances labeled as ‘true’ or ‘false’ (about 72% of the utterances in our corpus).

These features were selected by looking at the most distinctive features of true and false utterances, derived from the following approach. Frequency lists of all lemmas arising from both true and false utterances were created separately. The 200 most frequent lemmas for each type of utterance were selected and afterwards merged into a single list containing the most frequent lemmas of both classes of utterances. Theoretically, this list could have had a minimum of 200 items, in case of completely identifying the two previous lists, and a maximum of 400 items, in the case of no overlap.

The same procedure was applied to collect the following features, independently for each class, as shown in the Table 2.

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Table 2: Selected features

Features	Selected
Lemmas	first 200
Bigrams of lemmas	first 200
Trigrams of lemmas	first 200
POS	first 25
Bigrams of POS	first 25
Trigrams of POS	first 25
Total	675

The last features in the vector were the length of each utterance, with and without punctuation. Therefore, the theoretical minimum length of the final vector was 677 features (the lengths of the utterances, plus the 675 features stated above) and the theoretical maximum 1352. In the end, the feature vector had 1021 features. This suggests that the features of true and false utterances are quite different, and this is promising for the following analyses.

3.2. Baseliners

Before evaluating the results of the analyses, it was necessary to compute a baseline with which to refer. This was achieved through a simulation using a Monte Carlo technique. First, four hearings were used as a test set, for a total of 148 utterances, about 17% of the total amount of utterances in our corpus. This test set had 81 utterances labeled as ‘true’ and 67 as ‘false’, about 54.73% and 45.27% of ‘true’ and ‘false’ utterances, respectively. Then, 10,000 simulations were carried out, in which a classifier tried to guess the class of each entity of the test set, simply on the basis of the fact that 54.73% of the entities belong to the class ‘true’, and 45.27% belong to the class ‘false’. The result was that more than 99% of simulations did not exceed 60% of correct answers. Therefore 60% of correct classifications was assumed as the threshold for our test set.

3.3. Training

Using the training set mentioned above, models were built using the Naïve Bayes and SVM classifiers in the Weka package[3]. In order to evaluate the models' effectiveness in classification task, the said test set was employed.

4. Results

4.1. Modal performances

The results of the classifiers on the supplied test set are shown in Tables 3 and 4. While SVM[4] performs clearly better than the random classifier, with a remarkable precision detecting deception, Naïve Bayes barely exceeds the baseline of the 60%.

[3] http://www.cs.waikato.ac.nz/ml/weka/

[4] The algorithm for training Support Vector Machines was Sequential Minimal Optimization—SMO.

Table 3: Naïve Bayes performance—supplied test set

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	18	49	0.75	0.269	0.396
True utterances	75	6	0.605	0.926	0.732
Total	93	55			
Total %	**62.84%**	**37.16%**			

Table 4: SVM performance—supplied test set

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	33	34	**0.917**	0.493	0.641
True utterances	78	3	0.696	0.963	0.808
Total	111	37			
Total %	**75.00%**	**25.00%**			

4.2. Deceptive language

Overall, the performances of the SVM models are well above the chance level. Concerning false utterances, which are the target we have to detect, the precision is more than 90%. Instead, the recall needs to be improved, being slightly lower than 50%.

For the next stage of the analysis, it was necessary to determine which kinds of utterances were easier or, conversely, more difficult to classify.

To answer this question, the test set was examined. In general, the statements are very brief. From a total of 148 utterances:

- 95 contain 5 tokens or less (considered without punctuation): that is 64.19% of the test set;
- 27 have from 6 to 10 tokens;
- 26 are longer than 10 tokens.

Figure 1 represents the distribution of the length of the utterances, showing separately the true and the false ones.

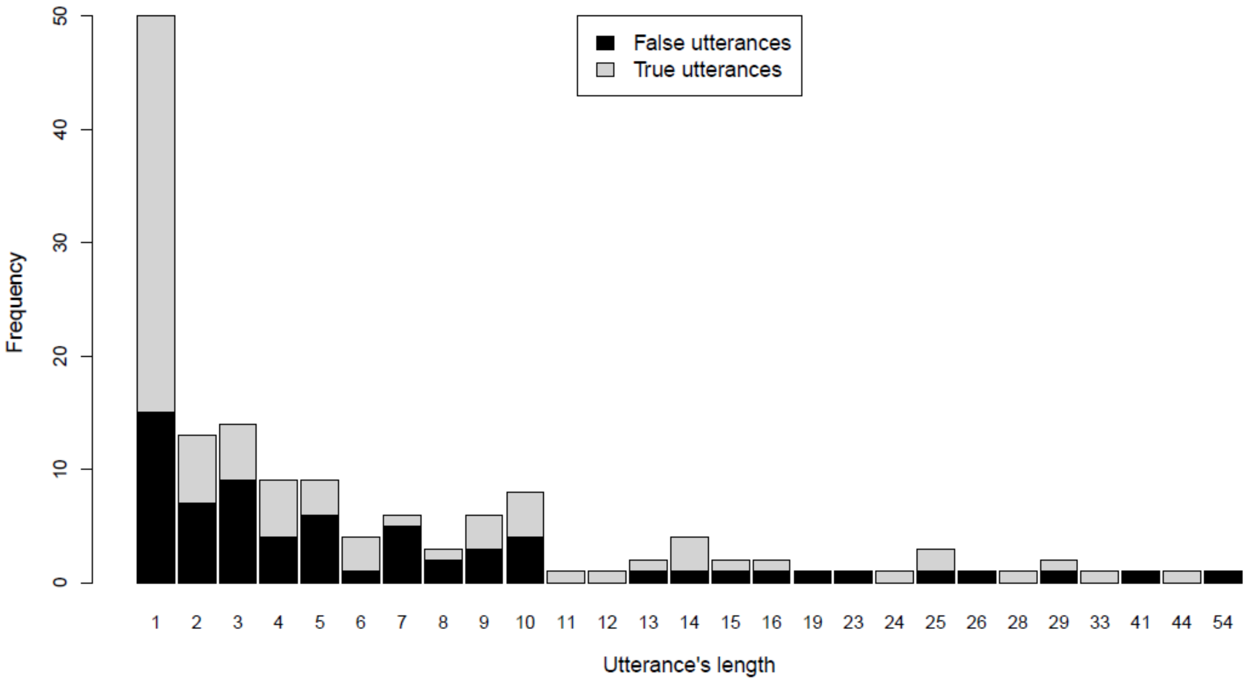

Figure 1: Distribution of length of utterances

Given this distribution, the accuracy of our models was examined on the basis of the length of the utterances. The results show that there is an improvement in performance with shorter statements compared to longer statements, as shown in Table 5.

Table 5: Accuracy of SVM models according to utterance lengths

Length	1-5 tokens		6-10 tokens		11 tokens	
Accuracy	**83.20%**		**59.20%**		**61.50%**	
	Precision	Recall	Precision	Recall	Precision	Recall
False utterances	1	0.61	1	0.267	0.571	0.364
True utterances	0.771	1	0.522	1	0.632	0.8

In particular, maybe contrary to what could be thought, the utterances equal to, or shorter than 5 tokens, are classified with accuracy higher than 80%, while the accuracy for the longer utterances corresponds more or less to the chance level.

Short statements are typically conventional, that is made by stereotyped linguistic formulas, which could be relevant in order to classify statements as true or false. To explore that idea, correspondence analysis has been carried out on the entire corpus (Baayen, 2008). The results are shown in Figure 2.

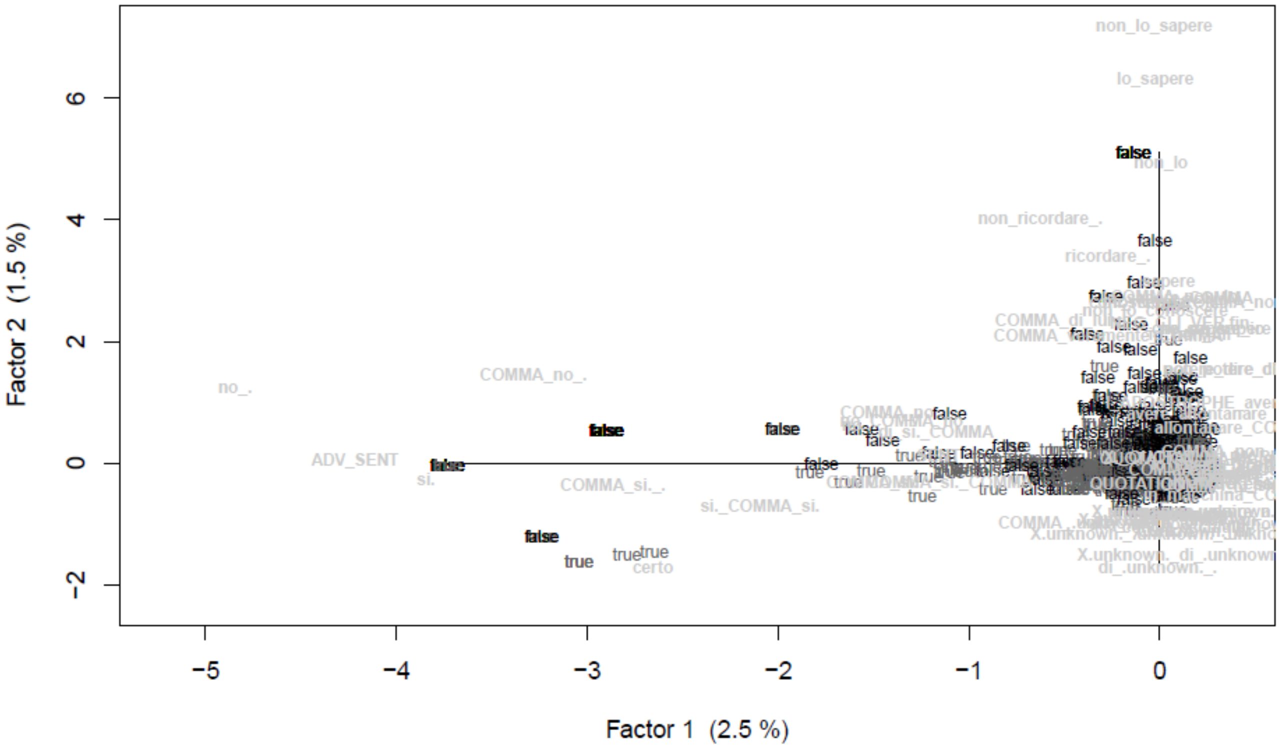

Figure 2: The correspondence analysis

The most ‘extreme’ features, useful to classify the utterances, are just brief and highly conventional expressions including, for example:

- ‘(Do not) know’ - ‘(Non) ricordare’;
- ‘(Do not) remember’ - ‘(Non) sapere’;
- ‘Yes’ - ‘Sì’;
- ‘Not’ - ‘No’;
- ‘Sure’ - ‘Certo’, and so on...

Therefore, it is possible to suppose that, when the language is more conventional, is easier to be recognized as true or false.

5. Conclusions

Our data show that it is possible to train models to classify statements as true or false, with performances clearly above the chance level. However long utterances are more difficult to classify, probably because their complexity represent noise for the models. This suggests that a future research direction may be to employ vectors containing less features, but constituted by longer n-grams, to detect expressions longer than three lemmas.

Deception is generally accepted to create an increase in cognitive load (Vrij, A., Fisher, R., Mann, S., and Leal, S., 2006). This increased cognitive load required to produce a deceptive statement, in culmination with the stress related to being involved in the hearings themselves, could account for the shorter, more conventional utterances identified in Section 4.2 which require less cognitive load to produce. In a not yet published paper by Tomblin et al. (in preparation) ‘Formulaic Language’ is used as a marker of deception and they arrive at analogous conclusions.

According to our data, in any case, it seems that commonly used expressions can be useful in identifying not only deceptive, but also sincere statements. In fact, if on one hand short negative answers, as well as brief denials of knowing or remembering, are typical of deceptive language, on the other hand short affirmative statements are generally truthful. It depends, at least in part, on the dynamic of the event of the hearing itself: prosecutors pose questions about facts which are the object of investigation, with the aim to verify the information collected during the inquiry. Therefore it is possible in the hearings to find several questions to which the interrogated subjects have to answer confirming or denying facts ascertained during the investigation activities. It is obvious that the subjects tend to be sincere when they confirm what the prosecutor already knows, and conversely they often lie through denials of responsibilities which are explicitly charged on them.

Therefore, to have better insights about deceptive language, and to be more precise at recognizing it, it is necessary to carry out more refined analyses. Another possible way to improve analyses could be to employ linguistic tools for the lexical evaluation of the texts. As underlined above, a well known resource of this kind, already employed to detect deception in texts (Newman et al., 2003), is the Linguistic Inquiry and Word Count (LIWC). In fact, exploratory analyses seem to be promising, regarding the possibility to improve the model performances, employing a combination of surface and lexical features. On the other hand, short utterances are already well classified.

Furthermore, the false utterances are recognized with a high degree of precision. It means that, at least for certain kinds of statements, deceptive language is clearly different to truthful language and it can be recognized. In a real life scenario such as the context described in this paper, the ability to confidently detect deception is an important contribution.

Acknowledgements

The data collection is a complex task that could not have been carried out without the help of a lot of people. Many thanks to Dr. Heinrich Zanon, President of the Court of Bolzano; to Dr. Sabino Giarrusso, President of the Court of Trento; and to Dr. Francesco Antonio Genovese, President of the Court of Prato. Many thanks also to Dr. Piero Tony, Chief Prosecutor of the Public Prosecutor's Office of Prato, to Dr. Sandro Pettinato of the Court of Trento, to Dr. Biagio Mazzeo, Prosecutor in the Public Prosecutor's Office of Genova, to Dr. Michela Guidi, Prosecutor in the Public Prosecutor's Office of Prato, and to Rita Fava of the Public Prosecutor's Office of Prato. Many thanks also to Dr. Marco Baroni of the University of Trento and to Dr. Carlo Strapparava of FBK - Fondazione Bruno Kessler.

References

Adams, S. H. (1996) Statement analysis: what do suspects’ words really reveal? *The FBI Law Enforcement Bulletin*, 65(10): 12—20.

Anolli, L., Balconi, M., and Ciceri, R. (1999) Ulisse o Richelieu? Stili verbali della comunicazione menzognera. *Lingua e stile*, 34(3): 379—402.

Baayen, R. (2008) *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge University Press.

Bachenko, J., Fitzpatrick, E., and Schonwetter, M. (2008) Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 41—48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bond, C. F. and De Paulo, B. M. (2006) Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3): 214—234.

Coulthard, M. (2004) Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25(4): 4310—447.

De Paulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003) Cues to deception. *Psychological Bulletin*, 129(1): 74—118.

Fitzpatrick, E. and Bachenko, J. (2009) Building a forensic corpus to test language-based indicators of deception. *Language and Computers*, 71(1): 183—196.

Frank, M. G. and Feeley, T. H. (2003) To catch a liar: challenges for research in lie detection training. *Journal of Applied Communication Research*, 31(1): 58—75.

Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. (2008) On Lying and being lied to: a linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1): 1—23.

Jensen, M. L., Meservy, T. O., Burgoon, J. K., and Nunamaker, J. F. (2010) Automatic, multimodal evaluation of human interaction. *Group Decision and Negotiation*, 19(4): 367—389.

Koppel, M., Schler, J., Argamon, S., and Pennebaker, J. (2006) Effects of age and gender on blogging. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*.

Levine, T. R., Feeley, T. H., McCornack, S. A., Hughes, M., and Harms, C. M. (2005) Testing the effects of nonverbal behavior training on accuracy in deception detection with the inclusion of a bogus training control group. *Western Journal of Communication*, 69(3): 203—217.

Luyckx, K. and Daelemans, W. (2008) Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, 513—520, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mosteller, F. and Wallace, D. (1964) *Inference and disputed authorship: The Federalist*. Addison-Wesley.

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(5): 665—675.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Lawrence Erlbaum Associates, Mahwah.

Porter, S., Woodworth, M., and Birt, A. R. (2000) Truth, lies, and videotape: an investigation of the ability of federal parole officers to detect deception. *Law and Human Behavior*, 24(6): 643—658.

Porter, S. and Yuille, J. C. (1996). The language of deceit: an investigation of the verbal clues to deception in the interrogation context. *Law And Human Behavior*, 20(4): 443—458.

Schmid, H. (1994) Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Solan, L. M. and Tiersma, P. M. (2004) Author identification in American courts. *Applied Linguistics*, 25(4): 448—465.

Stein, B., Koppel, M., and Stamatatos, E. (2007) Plagiarism analysis, authorship identification, and near-duplicate detection pan'07. *SIGIR Forum*, 41: 68—71.

Strapparava, C. and Mihalcea, R. (2009) The lie detector: explorations in the automatic recognition of deceptive language. In *Proceeding ACLShort '09 - Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*.

Tomblin, S., Taylor, P., Vrij, A., Leal,. S., Mann, S., Nash, R. & Menacere, T. (in preparation) Formulaic language occurs more often in deceptive statements.

Undeutsch, U. (1967) Beurteilung der Glaubhaftigkeit von Aussagen [Veracity assessment of statements]. In Undeutsch, U., editor, *Handbuch der Psychologie: Vol. 11. Forensische Psychologie*, 26—181. Hogrefe, Gottingen, Germany.

Vrij, A. (2005) Criteria-based content analysis—A qualitative review of the first 37 Studies. *Psychology, Public Policy, and Law*, 11(1): 3—41.

Vrij, A., Fisher, R., Mann, S., and Leal, S. (2006) Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, 10(4): 141—142.

Zhou, L. (2005) An empirical investigation of deception behavior in instant messaging. *IEEE Transactions on Professional Communication*, 48(2): 147—160.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., and Twitchell, D. (2004) Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13(1): 81—106.

Zhou, L., Shi, Y., and Zhang, D. (2008) A statistical language modeling approach to online deception detection. *IEEE Transactions on Knowledge and Data Engineering*, 20(8): 1077—1081.