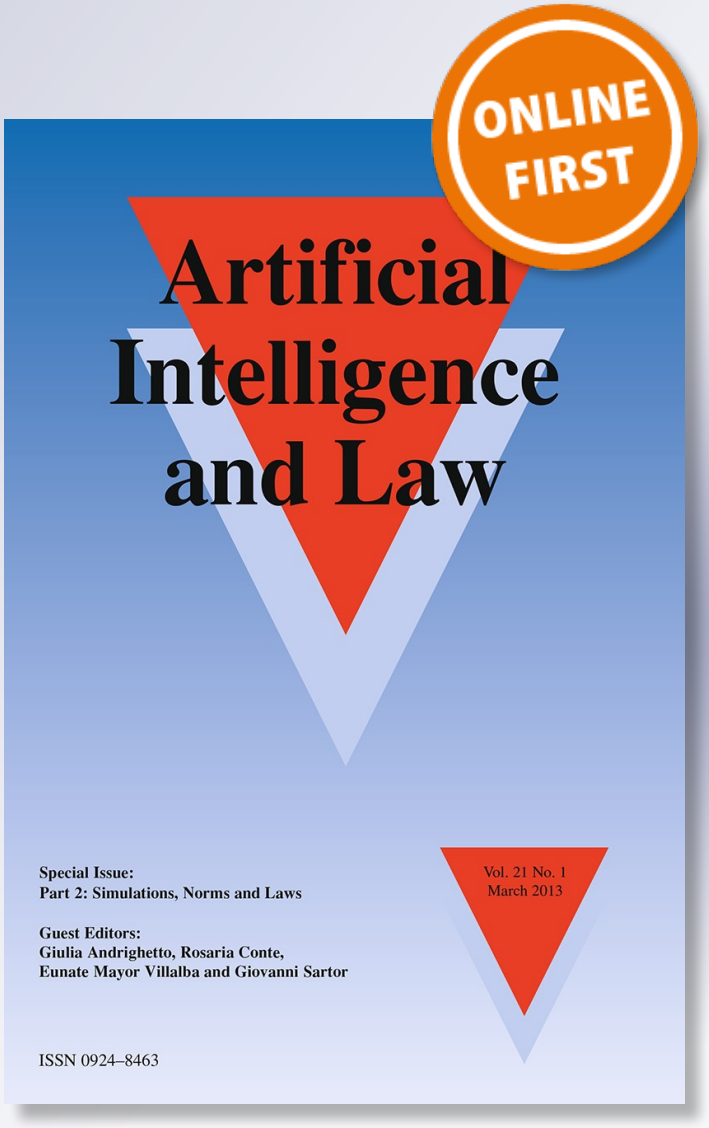
Automatic deception detection in Italian court cases

Tommaso Fornaciari & Massimo Poesio

Artificial Intelligence and Law

ISSN 0924-8463

Artif Intell Law
DOI 10.1007/s10506-013-9140-4



Springer

Artif Intell Law
DOI 10.1007/s10506-013-9140-4

Automatic deception detection in Italian court cases

Tommaso Fornaciari · Massimo Poesio

Abstract Effective methods for evaluating the reliability of statements issued by witnesses and defendants in hearings would be an extremely valuable support to decision-making in court and other legal settings. In recent years, methods relying on stylometric techniques have proven most successful for this task; but few such methods have been tested with language collected in real-life situations of high-stakes deception, and therefore their usefulness outside lab conditions still has to be properly assessed. In this study we report the results obtained by using stylometric techniques to identify deceptive statements in a corpus of hearings collected in Italian courts. The defendants at these hearings were condemned for calumny or false testimony, so the falsity of (some of) their statements is fairly certain. In our experiments we replicated the methods used in previous studies but never before applied to high-stakes data, and tested new methods. We also considered the effect of a number of variables including in particular the homogeneity of the dataset. Our results suggest that accuracy at deception detection clearly above chance level can be obtained with real-life data as well.

Keywords Deception detection · Stylometry · Criminal proceedings

1 Introduction

Effective methods for tagging potential deception on the basis of verbal or non-verbal cues (by hand or automatically) would have a number of applications in court

T. Fornaciari (✉)
Center for Mind/Brain Sciences, University of Trento, Trento, Italy
e-mail: tommaso.fornaciari@unitn.it

M. Poesio
School for Computer Science and Electronic Engineering, University of Essex, Colchester, UK
e-mail: poesio@essex.ac.uk

Published online: 21 February 2013

and other legal settings. The focus of the research presented in this paper is tagging potential deception in court testimonies to support criminal investigations in cases in which external evidence of the truthfulness of these testimonies is not (yet) available, but deception detection methods could also be applied in other legal, policing and security applications, for example to identify fake reviews of books or hotels, and in human resources evaluation. There has therefore been a great deal of research in the topic—see, e.g., De Paulo et al. (2003), Ekman (2001), Fitzpatrick and Bachenko (2009), Hancock et al. (2008), Newman et al. (2003), Strapparava and Mihalcea (2009), Vrij (2008), and many many others. Among other results, this line of research showed that, regarding behavioral clues to deception, "there is no clue or clue pattern that is specific to deception, although there are clues specific to emotion and cognition" (Frank et al. 2008). Meta-studies such as De Paulo et al. (2003) and Hauch et al. (2012), on the other end, identified a number of verbal cues systematically correlated with lying and truth telling: e.g., liars tend to use more negative emotion words, more motion verbs, and more negation words, whereas truth-tellers tend to use more self-references (*I, me, mine*) and more 'exclusive' words (i.e., exception connectives: *except*, *without*, etc.). [See also Newman et al. (2003)]. As a result, automatic methods focusing on verbal cues have been developed able to detect deception with reasonable accuracy (Newman et al. 2003; Strapparava and Mihalcea 2009).

This field of research suffers, however, from a serious problem: the difficulty of collecting data suitable to study the problem, or to develop automatic methods to identify deception. It is often difficult or impossible to verify the truthfulness of statements contained in data collected in natural environments (Vrji 2005). As a result, many if not most studies in the area, and in particular the just mentioned papers proposing computational techniques for deception detection, rely on data collected in laboratory conditions (Newman et al. 2003; Strapparava and Mihalcea 2009). But as the authors themselves point out (Newman et al. 2003), lying imposes a cognitive and emotional load on individuals which is not easy to reproduce artificially, and anyway achieving true 'high-stakes' deception would have serious ethical implications (Fitzpatrick and Bachenko 2009). (In the context of police investigations, the awareness of the legal consequences of a testimony and the emotional impact of speaking about criminal events can turn out to be very stressful for the subjects who issue statements.) Therefore it is by no means obvious that the results obtained with data collected in the lab will generalize to real life scenarios. For example, Undeutsch (1984) claimed that, due to the lack of ecological validity, laboratory studies are not very useful in testing the accuracy of tools for the evaluation of witnesses' reliability, such as the analyses based on Statement Validity Analysis—SVA (Vrji 2005). [Gokhmann et al. (2012) provide a useful review of the types of data used in deception detection research.]

As a result, Newman et al. (2003) identify the fact that "… external motivation to lie successfully was practically nonexistent…" among their participants as one of the main limitations of their work, the first and best known attempt to develop a computational method for deception detection relying entirely on verbal cues. A second limitation they identify is the fact that their model is limited to the English language; and given that differences in rates of self-reference is one of the main

cues for identifying truth-tellers, they see Romance languages such as Italian or Spanish as particularly interesting languages to test the cross-linguistic validity of their claims. In the research discussed in this paper we addressed these two limitations of the earlier study. Specifically, we set ourselves two objectives:

1. to collect a dataset in the context of criminal proceedings that would not suffer from the shortcomings of the datasets employed to develop earlier computational models of deception detection;
2. to compare the results obtained with this dataset with those obtained in earlier studies both from an accuracy point of view and from the point of view of the verbal cues employed.

In order to accomplish the first objective, we created a corpus of hearings in Italian courts for cases of **calumny** and **false testimony**, in which the defendant is accused to have issued deceptive statements during a previous hearing. When the defendants are found guilty, the trials end with a judgment which reconstructs the investigated facts and specifies quasi-verbatim the lies told in the courtroom. This information allowed us to annotate the utterances produced by the defendants as **true**, **false** or **uncertain** with great accuracy. The resulting corpus, called DeCour (for DEception in COURt) is the first resource for studying Italian true and false statements in a real life scenario. [And because the data are in a Romance language, the second limitation pointed out by Newman et al. (2003) can be addressed as well.]

DeCour was used to train text classification models classifying utterances as false or not-false purely on the basis of verbal information. Besides replicating the methods used by Newman et al. (2003), we also applied to the task a number of ideas from the field of Stylometry (see following Section).

The structure of the paper is as follows. Section 2 is a summary to the field of deception detection and to the application of stylometric techniques in this area. In Sect. 3 our dataset is described in more detail. In Sect. 4 we discuss the machine learning and experimental methods we used to identify deceptive statements in DeCour. Finally, the results are presented in Sect. 5 and discussed in Sect. 6.

2 Background

2.1 Detecting deception

Detecting deception in communication is a challenge for humans. Human performance at recognizing deception was found to be not much better than chance in a number of studies (Bond and De Paulo 2006). Other studies claim that even specific training is not particularly effective to improve human skills (Levine et al. 2005). On the other end, there are studies suggesting that the ability of humans as lie-detectors is underestimated (Frank and Feeley 2003). In any case, even in papers which reveal positive effects of training, the difficulty of the task is out of the question (Porter et al. 2000).

2.2 Approaches to deception detection

In part no doubt because of the very difficulty of the task, a wide variety of approaches to discover deceptive statements have been tried. The literature about deceptive communication can be divided in three main branches, depending on the cues investigated:

- Studies focused on non-verbal behaviour;
- Studies focused on verbal behaviour;
- Recent studies based on neuro-physiological, and in particular neuro-imaging techniques.

All of these approaches are however based on the same theoretical assumption, whether explicitly or implicitly: this is the idea, historically formalized by Undeutsch as the hypothesis which takes his name (Undeutsch 1967), that the cognitive elaboration of an untruthful narrative differs from the elaboration of a truthful one, therefore this difference should be traceable in the features of the narrative itself. Undeutsch was interested in verbal behavior, but his theoretical framework is also suitable to study non-verbal communication, and is consistent with recent findings using neuro-imaging techniques (Davatzikos et al. 2005; Ganis et al. 2003; Merikangas 2008; Simpson 2008).

2.3 Non-verbal approaches

The best known method for detecting deception, the polygraph, relies on non-verbal cues, but the literature contains a great number of papers studying the relation between deception and various aspects of non-verbal behaviour. One of the best known authors in this area is Ekman (2001), who studied in particular facial expressions. Other cues are the time taken to respond (response latency), or pupil dilatation (Wang et al. 2010). Many authors use combinations of cues in their attempt to improve accuracy at detecting falsehoods. This is the case of De Paulo et al. (2003), who consider more than 150 cues, verbal and non-verbal, observed through subjects mostly in lab conditions. Jensen et al. (2010) focused on cues coming from audio, video and textual data, with the aim of building a paradigm useful to identify deceptiveness.

However, coherently with the cited study of Frank et al. (2008), a common finding in this research is that it is difficult to identify non-verbal cues specific for deception, and also De Paulo et al. (2003) argue that "behaviors that are indicative of deception can be indicative of other states and processes as well". With regard to this, Walczyk et al. (2003) mention the case of Aldrich Ames, the spy who, from 1985 to 1994, provided the former Soviet Union with classified material he obtained as high-level agent of the CIA. During these 9 years, he successfully passed two polygraph tests.

2.4 Hermeneutic approaches

Undeutsch developed a framework called Statement Analysis (Undeutsch 1967, 1982, 1984), inspired by the notion of truth in interpretation as expressed in the field of Hermeneutics developed by Heidegger, Gadamer, and others. In such approaches

the truth of statements is assessed on the basis of principles called 'reality criteria' and designed to ensure that the statement is factual. Statement Analysis and its successors such as Statements Validity Analysis (SVA)—in turn divided in three stages, a semistructured interview, the Criteria-Based Content Analysis (CBCA), and an evaluation of the CBCA outcomes—are commonly used in forensic practice and in the literature. However, according to Vrji (2005) "SVA evaluations are not accurate enough to be admitted as expert scientific evidence in criminal courts but might be useful in police investigations". Thus Adams (1996), among others, asserted the necessity to take into account the personal style of communication together with the content of the testimonies.

2.5 Stylometry

The approach to the analysis of verbal cues for deception identification that is becoming more and more dominant in recent years is **stylometry**. Stylometry studies text on the basis of its stylistic features only. This can be done for a variety of purposes, e.g., in order to attribute the text to an author (**authorship attribution**) or to get information about the author, e.g., her/his gender or personality (**author profiling**). Stylometry actually goes back a very long way—the arguments used by Lorenzo Valla in the Fifteenth century to demonstrate the falsehood of the Donation of Constantine are essentially stylistic ones (Pepe 1996)—but it is only in the Nineteenth century that the field took place with the introduction by De Morgan of quantitative measures in stylistic studies (Lord 1958). (Quantitative) stylometric methodology was subsequently formalized by Lutoslawski (1898). Modern stylometry, which relies mainly on computational methods for automatically extracting low-level verbal cues from large amounts of text and on machine learning techniques, has proven effective in several tasks, including author profiling (Coulthard 2004; Solan and Tiersma 2004) [for example, deducing age and sex of authors of written texts (Koppel et al. 2006; Peersman et al. 2011)], author attribution (Luyckx and Daelemans 2008; Mosteller and Wallace 1964), emotion detection (Vaassen and Daelemans 2011) and plagiarism analysis (Stein et al. 2007).

2.6 Stylometric methods for deception detection

As Koppel et al. (2006) point out, the features used in stylometric analysis belong to two main families: surface-related and content-related features. The second kind of features, in turn, could be divided in two categories: features extracted from lexicons, and features coming from the linguistic analysis of texts themselves.

Surface-related features This type of features includes the frequency and use of function words or of certain n-grams of words or part-of-speech (POS tag), without taking into consideration their meaning.
Content-related features These features attempt to capture the meaning of texts. Such information may come from:

Lexicons lexicons associate each word to a variety of categories of different kinds: grammatical, lexical, psychological and so on. This results in a profile of texts with respect to those categories.
Linguistic analyses more complex analyses such as syntactic analyses, extraction of argument structure or coreference are also possible. Some of these analyses can be carried out automatically, but others, such as those carried out by Bachenko et al. (2008), can only be done by hand.

Newman et al. (2003) was arguably the first study showing that stylometric techniques could be effectively applied to detect deception. In that study, Newman et al. collected in the lab a corpus of sincere and deceptive texts from five different topics and contexts: videotaped, typed and handwritten discussions about attitudes to abortion, feelings about friends, and mock crime. These data were then analysed using a lexical resource: specifically, the Linguistic Inquiry and Word Count (LIWC), a lexicon created by Pennebaker's group (Pennebaker et al. 2001) and categorizing words under a number of categories such as their emotional content, self reference, etc.. The authors reached an accuracy of about 60 % (with a peak of 67 %) in three of the five studies, against a chance performance of 50 %. In the remaining two studies, the performance was not better than chance.

Strapparava and Mihalcea (2009) used surface features only. Strapparava and Mihalcea obtained good results at classifying into "sincere" or "deceptive" texts collected with the Amazon Mechanical Turk service.

Finally, an example of (semi-automatic) approach to deception detection using linguistic analysis is the work presented in Bachenko et al. (2008) and Fitzpatrick and Bachenko (2009). Fitzpatrick and Bachenko are in the process of collecting a high-stakes corpus including criminal statements, police interrogations, and civil testimony (Fitzpatrick and Bachenko 2012). Several linguistic indicators of deception were identified, such as **linguistic hedges** (e.g., *to the best of my knowledge...*), **overzealous expression** (*I swear to God*), **negative emotions** (*I was a nervous wreck*), and a variety of inconsistencies with respect to verb and noun form. The texts were then manually annotated with these indicators. This information was used as features to classify deceptive statements, with very high accuracy (close to 75 %).

3 Data set

In this section we briefly discuss how we collected and annotated a dataset containing examples of 'high stakes' deceptive language produced by subjects for whom the deception had real-life implications: the DeCour corpus.

3.1 Calumny and false testimony in the Italian Criminal Code

DeCour is a collection of hearings for "calumny" and "false testimony" (articles 368 and 372 of the Italian Criminal Code, respectively). While the concept of "false

testimony" is fairly intuitive,[1] in the Italian Criminal Code "calumny" is a particular kind of false testimony, consisting in the attempt to charge on someone else the responsibility of some crime which has been committed.[2] The distinction makes sense because in the Italian legal system nobody can be forced to make statements unfavorable to oneself; thus to lie about a committed crime is not considered a crime, but it is a crime to try to blame someone else. Therefore the hearings in DeCour come from two main situations:

- the defendant in a criminal proceeding tries to calumny someone;
- a witness in a criminal proceeding lies for some reason.

In both cases, a new criminal proceeding is initiated, in which the subjects can issue new statements or not, and having as body of evidence the transcript of the hearing held in the previous proceeding.

DeCour only contains hearings in which at the end the defendant is found guilty of "calumny" or "false testimony". Hence the proceeding ends with a judgment of the Court which summarizes the facts, pointing out precisely the lies told by the speaker in order to establish his punishment. Thanks to the transcription of the hearing in one hand, and to the final judgment of the Court in the other hand, it is possible to annotate the statements of the speakers on the basis of their truthfulness or untruthfulness.

3.2 Validity of the judgments of truth and falsity

Normally in corpus annotation one is only worried about *replicability*—i.e., whether different coders will assign the same code to an item. In this type of task however we are also concerned with *validity*: how confident can we be that the statements marked as false are actually false?

Of course, it is possible that Court judgments are wrong: some evidence coming form the inquiry could be in some way mistaken or misinterpreted by the judge. Since the annotation of DeCour relies on the information provided by the judgment, this would bring about an erroneous evaluation of the statements' truthfulness and would result in some noise in the data. This kind of risk is unavoidable.

[1] To be precise, Art. 372 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.

I.e., this article punishes who, in front of the Judicial Authority, says the false or denies the truth, or does not reveal what he knows about the investigated facts.

[2] Specifically, 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.

I.e., this article is violated whenever an individual tries to shift the blame for some crime on someone who he knows being innocent.

Our analysis of the data we collected suggests that any bias in Court is to the advantage of the defendant. In accordance with the principle of *in dubio, pro reo*,[3] when the least doubt exists about their guilt, defendants are not convicted. While collecting data we ran across several proceedings where the defendant was probably lying, and the judge most likely thought so as well, but in which the defendant was ultimately acquitted for lack of evidence of deception. These proceedings were not included in DeCour. On the other end, when the defendant is convicted, his guilt is always well demonstrated.

Therefore, even though it is not possible to estimate the rate of errors in these judgments, we expect it to be fairly low.

3.3 The hearings

Among the various kind of reports which are produced in a criminal proceeding, the minutes of the hearings held in Court seemed to be most appropriate and useful for our purposes, because they are transcripts which are required to reproduce *verbatim* what the subject said in courtroom.[4] DeCour is composed by the minutes of 35 hearings held in four Italian Courts: Bologna, Bolzano, Prato and Trento. These minutes report verbatim the statements produced by 31 different individuals (four of whom heard twice).

3.4 Preprocessing

3.4.1 Tokenization

The whole corpus was tokenized. The tokens include the words of the texts as well as punctuation. Punctuation marks are considered in blocks: this means that, for example, a single dot or a single question mark constitute a token, but an ellipsis that is three consecutive dots "..." also constitutes a single token. Our analysis units are the **utterances**, defined as strings of text delimited by punctuation marks, such as periods, question marks and ellipses. Taking punctuation marks in blocks prevents the creation of analysis units made uniquely by single punctuation marks. By contrast, apostrophes—which in Italian indicate the lack of the last vowel in the previous word—were not treated as separate tokens, but are kept together with the previous word. This helped the performance of the following lemmatization. Acronyms, such as "S.p.A.", "P.M." and so on, were considered as single tokens too. Otherwise, the dots would separate the letters constituting the acronym, with a proliferation of meaningless tokens and utterances. Lastly, hours expressed in numbers, such as "9:10", were also considered single tokens; in this case, the aim was to keep separated the numbers from the specific case of telling an hour.

[3] When in doubt, side with the accused.

[4] In particular, until 2005 the hearings were mainly recorded on tapes, which were used to be re-employed several times once the transcription was carried out. Therefore the audio tracks of the earliest hearings are definitively lost. Since 2006, instead, the audio tracks are recorded on CD-rom, and an attempt to get them is in process.

3.4.2 Anonymisation

Sensitive data were anonymised, as agreed with the Courts. Proper names of persons and things, such as streets, cars and so on, were substituted with five "x". Therefore, each proper name was counted as the same token "xxxxx", leaving a specific trace in the frequency lists of tokens of the cases in which the subject tells a proper name.

3.4.3 Lemmatization and POS-tagging

The whole corpus was put in lower-case, and then lemmatized and POS-tagged using a version of TreeTagger[5] (Schmid 1994) trained for Italian.

3.5 Annotation

The hearings are dialogs in which at least four roles are always present and have precise duties dictated by rules of the Criminal Proceeding Code. The **judge** is an impartial figure who has to judge the facts. The **prosecutor** brings about the accusations, whereas the **lawyer** is in charge of the defense. All of these individuals ask questions to the **defendant**, who has to answer them. These answers are the object of investigation of this study.

Each answer—i.e., all the text between the end of the previous intervention by another individual and the beginning of the following intervention—is considered a **turn**. Each turn is constituted by one or more **utterances** which, as said above, are delimited by terminal punctuation marks (period, triple-dots, question and exclamation mark). The individual utterance is the analysis unit of DeCour corpus and has been annotated according to the following annotation scheme:

True The utterance is held as true if coherent with the reconstruction of the facts reported in the final judgment.
False The utterances in contrast with that reconstruction are held as false. The judgment often lists precisely the lies told by the speaker: in this case the false utterances are easily identifiable.
Uncertain Even though the judgments give a complete description of the facts, they cannot account for every statement of the witness/defendant. The utterances whose truthfulness is not clear are classified as "uncertain". This category also includes the utterances lacking in propositional value, which from a logical point of view cannot be true or false, such as questions, meta-communicative acts and so on (for example "May you repeat, please?", "If you think so..."...).

In order to verify agreement in the judgments about truthfulness or untruthfulness of the utterances, three annotators annotated separately about 600 utterances. Reducing the agreement to a binary task—false utterances in one side and not-false utterances in the other side, that is true and uncertain utterances—we obtained a κ (Artstein and Poesio 2008) value of $\kappa = .64$.

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

3.6 Some statistics

DeCour is made of 3,015 utterances, which come from 2,094 turns. 945 utterances have been annotated as false, 1,202 as true and 868 as uncertain. The size of DeCour is 41,819 tokens, including punctuation blocks, distributed as follows:

Label	Utterances	Tokens
True	1,202	15,456
Uncertain	868	10,439
False	945	15,924
Total	3,015	41,819

4 Methods

In the next section we will present several experiments concerned with the development of computational models for deception detection based on machine learning techniques. In this section we discuss the methods used to train those models.

4.1 Features

In the experiments of Newman et al. (2003), lexical features from the LIWC were used. Much work in stylometry however suggests that comparable and occasionally better performance can be achieved using surface features such as n-grams of words and/or POS tags. We tested both types of features in our experiments.

4.1.1 Utterance length

In our experiments the unit of analysis are utterances rather than full documents and therefore (differently from the output of the LIWC) it does not make sense to count the mean number of words for sentence. But we do compute two utterance length features: **with** and **without punctuation**. These two features are used in all experimental conditions.[6]

4.1.2 LIWC features

Our first experiments were devoted to replicate Newman et al.'s study, employing the Italian version of LIWC software (Alparone et al. 2004).[7] The LIWC software outputs a few types of surface information about utterances in addition to the lexical

[6] Because our utterances are transcriptions of spoken language, the punctuation marks were inserted by the transcriber. They seemed nevertheless essential to understand the meaning of many utterances, hence their inclusion.

[7] The LIWC for several languages can be obtained from http://www.liwc.net.

information. Specifically, LIWC outputs **sentence word count**, the **mean number of words per sentence**, the **rate of coverage of the text** by the LIWC dictionary and the **number of words longer than six letters**. In the experiments where LIWC features are employed, we include among the features the utterance's length as said above, the **rate** of words found in the text which are also present in the LIWC dictionary and the **number of words longer than six letters**. The**mean number of words per sentence** is omitted as meaningless for our analysis units.

82 out of the 85 'dimensions' (lexical categories) of the LIWC Italian dictionary are also included among the features in these experiments. The features "Loro", "Passivo" and "Formale"[8] were discarded: "Loro" is used to categorize only one lexical item in the dictionary, whereas "Passivo" and "Formale" are not related to any term.

4.1.3 Lemma and POS n-grams

What we call here surface features are computed from frequency lists of n-grams of lemmas and part-of-speech. Lemma and part-of-speech n-grams of seven items were considered, from unigrams to eptagrams; long n-grams were included to identify conventional expressions. In each experiment, these frequency lists are computed from the subset of DeCour employed as training set in that experiment. More precisely, they come from the utterances classified as true or false in the training set, while utterances classified as uncertain were not considered in order to avoid picking up not discriminating features, coming from utterances whose truthfulness or truthlessness is not decidable or not known. Two different feature selection strategies were tested:

Best Frequencies Separate n-gram frequency lists were computed for true and false utterances in the training set, for both lemma and POS n-grams. The most frequent n-grams for each value of n were then chosen from these lists, in a decreasing number for increasing value of n. This approach was adopted as the higher the n the lower the absolute frequency of each n-gram. The number of the most frequent lemmas and part-of-speech collected for the different n-grams with this method, that we will henceforth call **Best Frequencies**, are shown in Table 1. Concretely, as shown in this Table, the 35 most frequent unigrams of lemmas were collected for true and false utterances, the 14 most frequent unigrams of POS, the 30 most frequent bigrams of lemmas and so on, until a total of 196 features from true and as many from false utterances were obtained. The overall number of surface features and the numbers of features of each type illustrated in Table 1 were arrived at on the basis of extensive empirical experimentation. The figure of 196 features in Table 1 is the number of features separately determined for false and true utterances. These separate lists of features are then merged into a single list, whose size depends on the degree of overlap: if the features chosen for false and true utterances are identical then only 196 features are used in total,

[8] "They", "Passive" and "Formal", respectively.

Table 1 The most frequent n-grams collected

N-grams	Lemmas	POS	Total
Unigrams	35	14	
Bigrams	30	12	
Trigrams	25	10	
Tetragrams	20	8	
Pentagrams	15	6	
Esagrams	10	4	
Eptagrams	5	2	
Total	140	66	196

whereas if n-grams for false and true utterances are completely disjoint then 392 n-grams (196+196) would be collected for each utterance.

Information Gain The second strategy for feature selection we employed is based on the popular Information Gain (IG) metric (Forman 2003; Yang and Pedersen 1997). **Information Gain** "measures the decrease in entropy when the feature is given vs. absent" (Forman 2003) according to the formula:

$$IG = e(pos, neg) - [P_{n-gram}e(tp, fp) + P_{\neg n-gram}e(fn, tn)]$$

in which e is the entropy:

$$e(x, y) = -\frac{x}{x+y}\log_2\frac{x}{x+y} - \frac{y}{x+y}\log_2\frac{y}{x+y}$$

and $P_{n-gram}, P_{\neg n-gram}$ are defined as follows:

$$P_{n-gram} = \frac{tp + fp}{all}$$

$$P_{\neg n-gram} = 1 - P_{n-gram}$$

where:

- tp = true positives: because the scientific focus of this work is to verify if it is possible to identify deceptive statements, true positives are the cases where the utterance is false and the feature is present;
- fp = false positives: the cases where the utterance is true and the feature is present;
- tn = true negatives: the cases where the utterance is true and the feature is absent;
- fn = false negatives: the cases where the utterance is false and the feature is absent;
- pos = positives: number of cases where the utterance is false (and the feature is present or absent: $tp + fn$);
- neg = negatives: number of cases where the utterance is true (and the feature is present or absent: $fp + tn$).

To compute the Information Gain of a feature, we compute the feature frequency lists for *n*-grams of lemmas and POS sequences as above, keeping all the *n*-grams with frequency higher than 5. We then compute the Information Gain of each feature and keep the 250 features with highest Information Gain.

4.2 Evaluation

In this section we discuss how the models were evaluated and the significance of the results assessed.

4.2.1 Evaluation metrics

In order to evaluate the performance of the models, four metrics were used:

Accuracy The overall accuracy is given by the sum of true and false utterances correctly classified, out of all the previsions carried out.
Precision We compute precision with regards to *false* utterances. This is the rate of correctly classified false utterances, out of all the entities classified as false:

$$p_f = \frac{tp}{tp + fp}$$

Recall Recall is the rate of correctly classified false utterances, out of all the false utterances present into the data set:

$$r_f = \frac{tp}{tp + fn}$$

F-measure F-measure is the harmonic mean of precision and recall (Chinchor 1992; Sasaki 2007):

$$f_f = 2 * \frac{p_f * r_f}{p_f + r_f}$$

In the rest of the paper we will omit the $_f$ indices except when required.

4.2.2 Random baseline

The performance of the models was compared to a number of baselines. The first of these baselines is an estimate of random performance computed through a Monte Carlo simulation. The basic idea of this kind of simulation is to perform several times a task over random inputs whose distribution reflects that of real data. Then the overall random performance is assumed as reference point to evaluate the results of tasks not-randomly carried out.

As said above, DeCour consists of 3,015 utterances, labeled as false, true or uncertain. Because our aim is to verify if it is possible to identify deceptive statements, and because many classifiers work best on binary problems, we

considered the 3015 utterances of DeCour as belonging to two subsets only, false and not-false utterances, the second class grouping together true and uncertain utterances. 945 utterances are false (31.34 % of the total) and 2,070 not-false.

In each step of the Monte Carlo simulations, utterances are assigned classes in such a way that the rate of elements classified as false is the same as in the gold standard; then the percentage of correct answers is computed. This procedure is repeated 100,000 times. In less than .01 % of trials the level of 60.03 % of correct predictions was exceeded. Precision at identifying false statements exceeded 37.03 % in less than 0.1 % of all simulations, whereas recall exceeded 35.97 % in less than 0.1 % simulations. These levels were therefore taken as chance level in the data analysis in the following section.

A second Monte Carlo simulation was carried out considering only utterances annotated as true and false, and discarding those classified as uncertain. 2,147 utterances remained, of which 1,202 true and 945 false, as above. Out of the 100,000 simulations, less than .01 % showed an accuracy higher than 54.54 %, while the thresholds for precision and recall were respectively 49.95 and 48.36 %

4.2.3 The majority baseline

Another straightforward kind of baseline is the so-called Majority Baseline: assigning to each utterance the label of the majority class. The accuracy of this baseline is equal to the percentage of items belonging to the majority class. In the case of DeCour, the rate of not-false utterances is 68.66 %; if uncertain utterances are not considered, the rate of true utterances is 55.98 %.

The Majority Baseline can be difficult to beat, but it's not always very helpful: in our application for instance always assigning to utterances the label not-false would give us an accuracy of 68.66 %, but a recall over false utterances (i.e., those we are actually interested in) of 0 %.

4.2.4 A simple heuristic algorithm

Finally, a third baseline was considered, a heuristic algorithm motivated by the observation discussed in previous work (Fornaciari and Poesio 2011) that often in the hearings the prosecutor charges the defendant of facts that are known thanks to the inquiry, and therefore a common form of lie is to deny those facts, or to claim not to know or not to remember them. The heuristic algorithm is as follows:

- The utterances beginning with the words *Sì* (Yes), *Lo so* (I know) and *Mi ricordo* (I remember) are classified as true;
- The utterances beginning with the words *No* (No), *Non lo so* (I don't know) and *Non mi ricordo* (I don't remember) are classified as false;
- All other utterances are randomly classified as true or false, according to the rate of true and false utterances present in DeCour.

After 100,000 trials, the performance of this algorithm was better than that of the Monte Carlo simulation, both regarding overall accuracy and with respect to

precision and recall. Yet with the whole DeCour, less than 0.1 % of the trials reached an accuracy higher than 62.39 %. Also with $p < .001$, the precision threshold was 40.06 % and the recall threshold 41.80 %. Considering only true and false utterances, the levels for the algorithmic baseline were 59.57 % for accuracy, 54.38 % for precision and 52.80 % for recall.

4.3 Training the models

In previous work we tested a variety of classification methods, finding that the best performance in general was obtained with Support Vector Machines (SVMs; Cortes and Vapnik 1995), a classification method successfully employed in many applications involving text classification (Yang and Liu 1999). SVMs rely on the identification of optimal hyperplanes in a feature space describing each entity of a data set. In order to do this on data set in which entities are not linearly separable, kernel functions are employed, which re-arrange the entities in a higher dimensional space where linear separation is possible (Zhou et al. 2008).

Therefore, the choice of the most appropriate kernel function is fundamental to obtain good performance in classification task. Linear kernel functions are usually considered useful in text categorization, where often one deals with large sparse data vectors, as in the study of Karatzoglou et al. (2006). Nevertheless in the following experiments radial kernel functions are employed, because on DeCour they gave more uniform results and overall better performance in the various experimental conditions.

Our SVM models were trained and then tested via n-fold cross-validations. In all the experimental conditions, each hearing of DeCour constitutes a fold for the cross-validations, so that the experiments run on the whole corpus have been carried out with a 35-fold cross-validation. Other experiments were also carried out, where only some subsets of DeCour have been taken into consideration; in these cases, some hearings were discarded and thence a n-fold cross-validation corresponding to the number of the employed hearings was carried out.

5 Experiments and results

Thirteen experiments were carried out, divided in three groups. The first group of five experiments were concerned with replicating the methodology of Newman et al. (2003) in a high-stakes deception scenario and with comparing the performance of the lexical features used in that work with that of surface features, which have often been shown to achieve similar or better performance. The goal of the second group of experiments was to compare the performance of the classifier on the entire corpus with the performance on the subset of utterances classified as true or false only—arguably a more realistic application of the methodology we used, which would only be used for utterances that according to the investigators or the judges could be held as relevant to be classified as true or false. Finally, in the last group of experiments we studied whether better results could be obtained by

focusing on more cohesive sets of subjects—only male speakers, only Italian native speakers, and only speakers above 30 years of age.

5.1 Comparing lexical and surface features

5.1.1 Preliminary discussion

The results of these first experiments suggest that the methods employed by Newman et al. do achieve results above chance even with real-life data. These results are lower than those obtained with the majority baseline, but this could not result in usable data. Also, results above the majority baseline can be obtained using surface features only.

5.1.2 Using the LIWC

In the first experiment, LIWC was used to classify deceptive texts in a near-replication of Newman et al. (2003). The most significant differences were that our texts were in Italian and therefore the Italian LIWC was used instead of the English LIWC; that utterances were classified instead of whole texts; and that SVMs were used instead of logistic regression. A 35-fold cross validation was carried out over the whole DeCour corpus. 86 features were used to categorize utterances: the 2 utterance length features from Sect. 4.1.1 and the 84 LIWC features from Sect. 4.1.2.

The results of this experiment are summarized in Table 2.[9] The mean accuracy in detecting false utterances reached in this experiment was 68.28 %, with standard deviation $\sigma = 8.86$. This rate of accuracy is almost 6 points percent higher than that of the heuristic algorithm, but does not exceed the majority baseline, which achieves the highest results.

5.1.3 Surface features

In the second and third experiments, only surface features were used in addition to the utterance length features. As discussed above, two approaches to choosing surface features were tried: simple frequency and Information Gain. As in the first experiment, a 35-fold cross validation was carried out (notice that because the surface features are selected from the training set, this means that different features could potentially be chosen in each of the 35 repetitions).

Best frequencies The results obtained with Best Frequencies are summarized in Table 3. The mean accuracy of the models was 68.29 %, with standard deviation $\sigma = 11.13$. As in the previous experiment, the performance is higher than that of the heuristic baseline and random choice, but not than that of the majority baseline. The average number of features employed in each fold of the experiment using Best Frequencies was 296.54, with standard deviation $\sigma = 2.20$; the best surface features are shown in Table 1.

[9] Here and in the rest of the paper we indicate the highest accuracy achieved in bold.

Table 2 Results with LIWC lexical features on the whole corpus

	Correctly classified entities	Incorrectly classified entities	Precision (%)	Recall (%)	F-measure (%)
False utterances	344	601	51.57	36.40	42.68
True utterances	1,747	323	74.40	84.40	79.09
Total	2,091	924			
Total accuracy	69.35 %	30.65 %			
Mean accuracy	68.28 %				
Monte Carlo baseline	60.03 %				
Majority baseline	**68.66 %**				
Heuristic baseline	62.39 %				

Table 3 Surface features: best frequencies

	Correctly classified entities	Incorrectly classified entities	Precision (%)	Recall (%)	F-measure (%)
False utterances	305	640	53.42	32.28	40.24
True utterances	1,804	266	73.81	87.15	79.93
Total	2,109	906			
Total accuracy	69.95 %	30.05 %			
Mean accuracy	68.29 %				
Monte Carlo baseline	60.03 %				
Majority baseline	**68.66 %**				
Heuristic baseline	62.39 %				

Information gain In a second experiment, the surface features were selected according the Information Gain strategy. The results are summarized in Table 4. The mean accuracy for this experiment was 69.89 %, with standard deviation $\sigma = 9.73$. This is the best result among the first group of experiments; both the majority and the heuristic baseline are improved upon (by 1 and 7 % points, respectively). The feature vectors in this case consisted of 252 features: 250 surface features and the two utterance length features.

5.1.4 Combining lexical and surface features

Finally, we tried combining both the lexical features from the LIWC and the surface features chosen either through Best Frequencies or through Information Gain.

Table 4 Choosing surface features using information gain

	Correctly classified entities	Incorrectly classified entities	Precision (%)	Recall (%)	F-measure (%)
False utterances	393	552	53.11	41.59	46.65
True utterances	1,723	347	75.74	83.24	79.31
Total	2,116	899			
Total accuracy	70.18 %	29.82 %			
Mean accuracy	69.89 %				
Monte Carlo baseline	60.03 %				
Majority baseline	68.66 %				
Heuristic baseline	62.39 %				

LIWC + best frequencies In the first case, the 84 LIWC-related features and the surface features of the second experiment were used; for an average number of features in the 35-fold of 380.54, with standard deviation $\sigma = 2.20$. In this experiment the mean accuracy was 68.96 %, with standard deviation $\sigma = 9.94$: this result is higher than the heuristic baseline (by more than 6 % points) and the majority baseline (although only by a few tenths of point). The overall performance of the 35-fold cross-validation is presented in Table 5.

LIWC + information gain Alternatively, the 84 LIWC features were combined with surface features collected with Information Gain. In this case, 336 features were used in total. The mean accuracy was 68.59 %, with standard deviation $\sigma = 10.03$. This is about 6 % points higher than the heuristic baseline, but it is slightly lower than the majority baseline. Table 6 summarizes the results.

Table 5 LIWC + best frequencies features

	Correctly classified entities	Incorrectly classified entities	Precision (%)	Recall (%)	F-measure (%)
False utterances	327	618	54.77	34.60	42.41
True utterances	1,800	270	74.44	86.96	80.21
Total	2,127	888			
Total accuracy	70.55 %	29.45 %			
Mean accuracy	68.96 %				
Monte Carlo baseline	60.03 %				
Majority baseline	68.66 %				
Heuristic baseline	62.39 %				

Table 6 LIWC + information gain features

	Correctly classified entities	Incorrectly classified entities	Precision (%)	Recall (%)	F-measure (%)
False utterances	382	563	52.54	40.42	45.69
True utterances	1,725	345	75.39	83.33	79.16
Total	2,107	908			
Total accuracy	69.88 %	30.12 %			
Mean accuracy	68.59 %				
Monte Carlo baseline	60.03 %				
Majority baseline	**68.66 %**				
Heuristic baseline	62.39 %				

5.2 Discriminating between clearly false and clearly true utterances

5.2.1 Preliminary discussion

The results discussed in this section suggest that when applying the models to the arguably more realistic data obtained by removing irrelevant utterances, we obtain results well above any baseline as well as well above chance.

In particular, in this second series of experiments the utterances annotated as 'uncertain' were discarded, and only 'true' and 'false' utterances considered. Although this selection might at first seem just a way of improving performance, we believe in fact it reflects more accurately how methods such as those discussed in this paper could actually be used to support investigative and court practice. Investigators and judges are unlikely to be interested in testing every single utterance of the accused. When a witness/defendant issues statements, they often mention facts which are universally known as true (for example introducing more relevant topics: "That evening we were at the disco..."), or not particularly relevant for the purposes of the investigation ("I have my lawyer..."). Furthermore, several utterances have just a meta-communicative value, such as "If you were me...", "I do not understand", "Now let me explain," and so on. Even when these declarations have propositional value, their classification is not useful with respect to the facts that the inquiry has to ascertain. Along with the assertions whose truthfulness is unknown, the category of 'uncertain' utterances contains just this last kind of statements, of which the value true/false is not clear or by definition not appropriate. Thus to remove them from the dataset reduces the noise in the data, by excluding utterances which in any case would not need to be classified. Other than the restriction to a subset of the data, the exact same methods are used in the experiments of this second group than were used in the experiments of the first group.

5.2.2 Using the LIWC

Table 7 shows the results obtained by using the LIWC only, as in the first experiment of the first group, but discarding uncertain utterances. The mean accuracy of the 35-folds is 66.48 %, with standard deviation $\sigma = 9.78$. This is almost 7 % points above the most demanding baseline, which for this set of experiments is the heuristic one (removing the uncertain utterances greatly lowers the majority baseline).

5.2.3 Surface features

Best frequencies Table 8 shows the results obtained in this task by using surface features selected using the Best Frequencies technique. The mean accuracy is 68.62, with standard deviation $\sigma = 10.32$—i.e., 9 % points higher than the heuristic baseline.

Information gain This experiment replicates the third experiment of the first group, but without uncertain utterances. In this case, the performance is not the best of the set of experiments: the mean accuracy is 68.25 % (with standard deviation $\sigma = 9.65$): almost 9 points above the heuristic baseline. All the results are summarized in Table 9.

5.2.4 Combining features

LIWC + best frequencies While in the fourth experiment of the first group, mixing lexical and surface features (collected with the Best Frequencies method) did not lead to good results, using this combination with false / frue utterances only results in the best performance in this second group of experiments. The results are shown in Table 10: the mean accuracy is 69.84 %, with standard deviation $\sigma = 10.29$. The distance between the performance and the heuristic baseline is more than 10 % points.

Table 7 Classifying false/true utterances with the LIWC

	Correctly classified entities	Incorrectly classified entities	Precision	Recall (%)	F-measure (%)
False utterances	554	391	65.56 %	58.62	61.90
True utterances	911	291	69.97 %	75.79	72.76
Total	1,465	682			
Total (%)	68.23	31.77			
Mean accuracy	66.48 %				
Monte Carlo baseline	54.54 %				
Majority baseline	55.98 %				
Heuristic baseline	59.57 %				

Table 8 False/true utterances classification with surface features: best frequencies

	Correctly classified entities	Incorrectly classified entities	Precision	Recall (%)	F-measure (%)
False utterances	540	405	69.05 %	57.14	62.53
True utterances	960	242	70.33 %	79.87	74.80
Total	1,500	647			
Total (%)	69.86	30.14			
Mean accuracy	68.62 %				
Monte Carlo baseline	54.54 %				
Majority baseline	55.98 %				
Heuristic baseline	59.57 %				

Table 9 False/true utterances classification with surface features: information gain

	Correctly classified entities	Incorrectly classified entities	Precision	Recall (%)	F-measure (%)
False utterances	533	412	68.77 %	56.40	61.97
True utterances	960	242	69.97 %	79.87	74.59
Total	1,493	654			
Total (%)	69.54	30.46			
Mean accuracy	68.25 %				
Monte Carlo baseline	54.54 %				
Majority baseline	55.98 %				
Heuristic baseline	59.57 %				

Table 10 False/true utterances classification: LIWC + best frequencies

	Correctly classified entities	Incorrectly classified entities	Precision	Recall (%)	F-measure (%)
False utterances	538	407	70.60 %	56.93	63.03
True utterances	978	224	70.61 %	81.36	75.60
Total	1,516	631			
Total (%)	70.61	29.39			
Mean accuracy	69.84 %				
Monte Carlo baseline	54.54 %				
Majority baseline	55.98 %				
Heuristic baseline	59.57 %				

LIWC + information gain The last experiment of this set is the twin of the fifth one of the first series: the LIWC features were combined to surface features collected according to the Information Gain method, and employed for a 35-fold cross-

Table 11 False/true utterances classification: LIWC + information gain

	Correctly classified entities	Incorrectly classified entities	Precision	Recall (%)	F-measure (%)
False utterances	512	433	71.31 %	54.18	61.58
True utterances	996	206	69.70 %	82.86	75.71
Total	1,508	639			
Total (%)	70.24	29.76			
Mean accuracy	68.90 %				
Monte Carlo baseline	54.54 %				
Majority baseline	55.98 %				
Heuristic baseline	59.57 %				

validation experiment, where only true and false utterances were considered. The results are shown in Table 11. The mean accuracy is 68.90 %, with standard deviation $\sigma = 11.18$: that is more than 8 points percent higher than heuristic baseline.

5.3 Selecting more homogeneous sets of defendants

5.3.1 Preliminary discussion

Finally, in the last series of experiments, we attempted to determine whether better results could be achieved by training and testing on more homogeneous sets of speakers. DeCour gave us the opportunity to try three ways of making the sets more homogeneous: (1) only considering defendants of the same gender (unfortunately we only have enough data to try this on male defendants); (2) only Italian native speakers; and (3) defendants of a similar age. We consider each of these in turn.

5.3.2 Only male speakers

A possibility that was often mentioned to us was that male and female speakers lie in different ways, and therefore training and testing on defendants of the same gender could yield better results. Unfortunately DeCour only includes 8 hearings in which the defendant is a woman, which we found is not enough data to build reliable models. We could however try this with male defendants. We removed therefore 10 hearings, in which the defendants are either women or transgender. The remaining subset consisted of 2,234 utterances, of which 712 were false (31.87 % of the total). A new Monte Carlo simulation was carried out, obtaining (with $p < .001$) baselines of 60.11 % for accuracy, 38.48 % for precision and 37.25 % for recall. The heuristic baseline achieved an accuracy of 62.58 %, a precision for false utterances of 41.24 % and a recall of 42.84 %. The Majority baseline was 68.13 %.

As in the previous experiments the highest accuracy was achieved by only using surface features collected through Information Gain, we used this model in the present and the following experiments.

A 25-fold cross-validation was carried out, obtaining a mean accuracy of 69.51 %, with standard deviation $\sigma = 8.81$. This means that the performance exceeds the majority and heuristic baselines. Table 12 presents the overall results in this experiment.

5.3.3 Only Italian native speakers

A second possibility is that Italian native speakers use different cues than non-Italians. In this experiment the nine hearings in which the defendant was not born in Italy were discarded. The remaining dataset consisted of 2,177 utterances, of which 625 (28.71 %) were false. Therefore, the Majority Baseline was 71.29 %. Instead, according to the Monte Carlo simulation, with $p < .001$ the accuracy baseline was 62.56 %, whereas the baselines for precision and recall were 35.52 and 34.48 % respectively. Accuracy, precision and recall for the heuristic baseline were respectively 64.22, 37.93 and 40.64 %.

The mean accuracy of the models, trained with a 26-fold cross-validation, was 70.12 %, with standard deviation $\sigma = 7.99$. This accuracy is not higher than the majority baseline, but exceeds the heuristic one for about 6 points percent. Table 13 summarizes the results of each fold.

5.3.4 Only over 30 years old speakers

In the last experiment, only defendants over 30 years old were considered. This age was chosen as a trade-off between the necessities, on one hand, not to remove too much hearings from DeCour, and on the other hand to divide the subjects in meaningful groups. Because the Courts where the data were collected deal with crimes committed by people over 18 years old, to focus on subjects over 30 years of age meant to discard 14 hearings. The remaining dataset consisted of 1,917 utterances, of which 597 (31.14 %) false. The Majority Baseline was therefore 68.86 %. The threshold of accuracy according to a Monte Carlo simulation was 60.93 % with $p < .001$. The precision baseline was 38.36 % and the recall baseline

Table 12 Only male speakers

	Correctly classified entities	Incorrectly classified entities	Precision (%)	Recall (%)	F-measure (%)
False utterances	292	420	52.52	41.01	46.06
True utterances	1,258	264	74.97	82.65	78.62
Total	1,550	684			
Total accuracy	69.38 %	30.62 %			
Mean accuracy	69.51 %				
Monte Carlo baseline	60.11 %				
Majority baseline	68.13 %				
Heuristic baseline	62.58 %				

Table 13 Only Italian native speakers

	Correctly classified entities	Incorrectly classified entities	Precision (%)	Recall (%)	F-measure (%)
False utterances	255	370	50.20	40.80	45.01
True utterances	1,299	253	77.83	83.70	80.66
Total	1,554	623			
Total accuracy	71.38 %	28.62 %			
Mean accuracy	70.12 %				
Monte Carlo baseline	62.56 %				
Majority baseline	**71.29 %**				
Heuristic baseline	64.22 %				

Table 14 Only over 30 years old speakers

	Correctly classified entities	Incorrectly classified entities	Precision (%)	Recall (%)	F-measure (%)
False utterances	252	345	52.07	42.21	46.62
True utterances	1,088	232	75.92	82.42	79.04
Total	1,340	577			
Total accuracy	69.90 %	30.10 %			
Mean accuracy	70.28 %				
Monte Carlo baseline	60.93 %				
Majority baseline	68.86 %				
Heuristic baseline	63.90 %				

was 36.99 %. The accuracy with $p < .001$ of the heuristic baseline was 63.90 %, the precision 41.12 % and the recall 44.39 %.

After the 21-fold cross-validation, the mean accuracy in classification task was 70.28 %, with standard deviation $\sigma = 7.83$. Table 14 shows the overall performance of the model, which is better than both the majority and heuristic thresholds.

6 Discussion

6.1 Predicting deception

Our first result is that all models proposed in Sect. 4 can identify deceptive statements with an accuracy of around 70 %, which is well above chance and much

better than the simple heuristic algorithm. This suggests that the type of methods proposed by Pennebaker et al. (2001) and Strapparava and Mihalcea (2009), relying only on automatically extracted features, can be applied with a certain degree of success to identify deception even with real-life data collected in high-stakes situations. Not all models outperformed the majority baseline, but for all types of tasks at least one of the non-trivial models achieved a performance better than that tougher baseline by at least 1 % point. In the rest of this subsection we discuss more in detail what makes the task so hard and how the performance could be improved.

6.1.1 The effect of size

The simplest way to improve the performance of a model whose learning curve has not yet plateaued is to increase the size of the corpus. Because the size of DeCour is not very large due to the time it takes to collect the relevant data, the first type of analysis to carry out to investigate the possibility of achieving better performance is simply to study the learning curve of our models.

The learning curve we studied is that of the model obtained in our third experiment in which surface features were collected through Information Gain, since this model achieved the highest mean accuracy among those tested in the first group of experiments employing all data. The learning curve was computed by carrying out cross-validations using 1 hearing for testing and respectively 1, 5, 10, 15, 20, 25, 30 and 34 hearings for feature selection and as training set. The last experiment replicated the one taken as reference point. The results are shown in Fig. 1.

In previous deception detection experiments (Strapparava and Mihalcea 2009), a plateau was observed—increasing the training set size, the models' performance does not improve any longer. In our case however no plateau is visible; on the contrary, the learning curve is growing fairly regularly, suggesting that performance could still be improved by adding more data. The curve also shows that the accuracy of the models is higher than baselines such as the heuristic baseline even when just

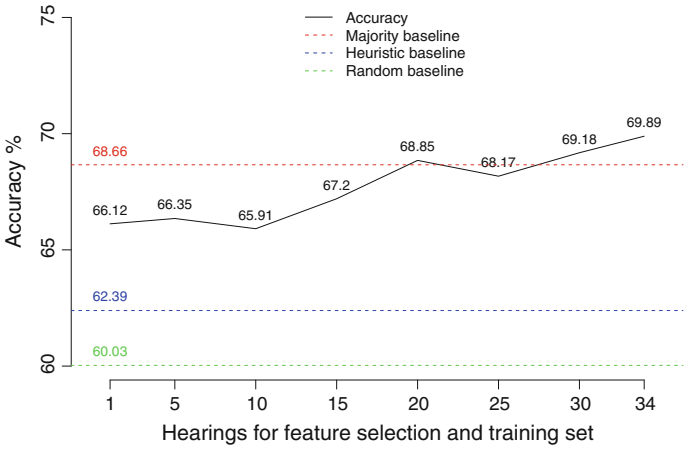


Fig. 1 The learning curve

one hearing is used for training. The features selected from such single hearings are also very similar to those showed and discussed in the next subsections. This suggests that deceptive language is highly stereotyped, and therefore relatively few surface features are sufficient to get results slightly better than chance.

6.1.2 Deception at the utterance level

This is no mean achievement, considering that the task our models have to perform is much more challenging than the one attempted by, e.g., Pennebaker et al. (2001), who only attempted to classify full texts. In DeCour, 496 utterances out of 3015 (16.45 %) are single-word utterances, and 70.44 % of DeCour is constituted by utterances no longer than 15 words. Figure 2 provides the distribution of the lengths of the utterances in DeCour. But as discussed, e.g., in Fitzpatrick and Bachenko (2012), working at the level of the entire narrative identifies the liar, not the lie.

This scenario we are working with may originate two types of criticism. On the one hand, the small amount of information present in the utterances can make them indistinguishable from each other. Some critics may therefore argue that the task is simply impossible; to which the best reply is to show that in fact accuracy above chance can be obtained even with relatively simple methods.

On the other hand, this very shortness of the utterances may be evidence that defendants use language in a way that is easily predictable knowing the ritual of the hearings in Court. Because many of the questions addressed to the defendant are accusations, we may expect he/she to be most likely untruthful while denying them, whereas he/she will be more likely to be sincere when positively asserting known facts. In other words, other critics may argue that in fact the problem of deception detection in this type of context can be solved with fairly simple techniques. To some extent, this is true: the simple algorithm we used as an additional baseline, and based on the heuristic that defendants are most likely untruthful when they deny

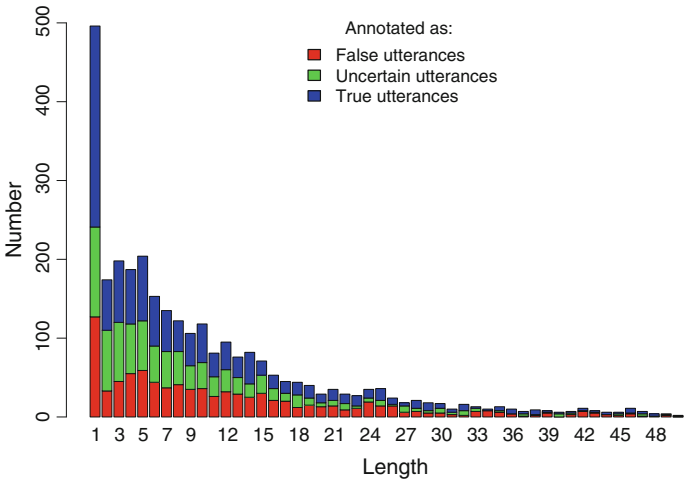


Fig. 2 The distribution of the lengths of the utterances in DeCour

something, is always around 2 % points more accurate than chance. However the fact that this baseline never exceeds an accuracy of 62–63 % suggests that the problem is not so simple.

There also seems to be a correlation between length of the utterance and classification accuracy, as can be seen from Fig. 3, in which utterance length and classification accuracy in the experiment using surface features selected using Information Gain (Table 4) are charted. Clearly, the longer the utterances, the lower the accuracy.

6.1.3 Uncertainty and noise

The models also behave better when applied to cleaner data. In the experiments in which uncertain utterances are excluded the gap between mean classification accuracy using our trained models and the heuristic baseline grows from about 6 to about 9 % points. As explained above, the class of uncertain utterances consists of (1) utterances which cannot have a value of true or false (e.g., questions) or (2) whose truthfulness cannot be decided on the basis of the available evidence. This second group of utterances may therefore contain both false and true statements, which introduces some noise into the dataset; this in turn clearly affects both the training and the testing of the models (even though the uncertain utterances are not employed to identify the features of the models), making the classification task more difficult. This hypothesis that the class of uncertain utterances consists of a blend of false and true ones is supported by looking at Fig. 4. In this Figure we show the distribution of the probabilities assigned by the classifier in the experiment in which we obtained the best results (surface features using Information Gain). If the probability that an utterance is false is >.5, the classifier treats it as false; else, as not-false. We can see that most of the utterances annotated as true in the corpus were given by the classifier a probability of being false of less than .5—in fact, the great majority of those got a probability less than .2. In the case of utterances

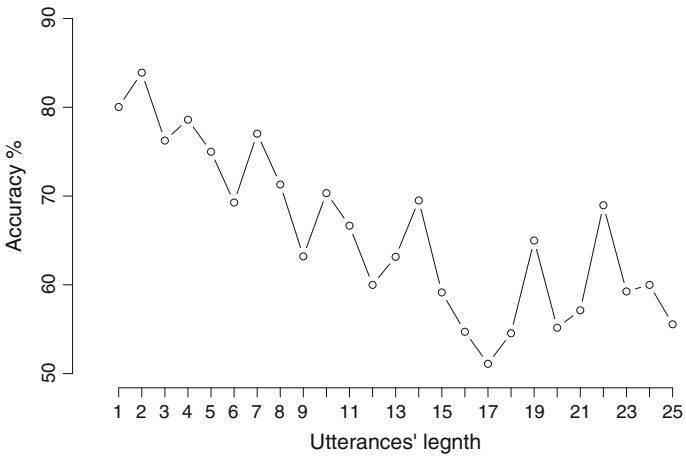


Fig. 3 The relation between utterance length and classification accuracy

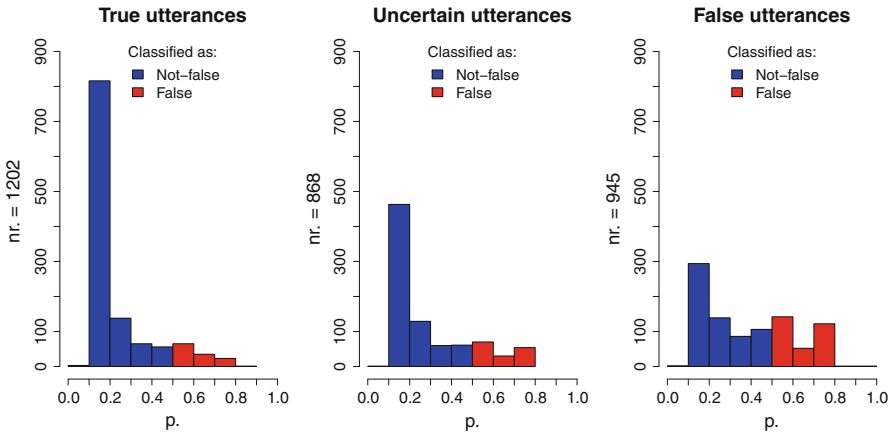


Fig. 4 The probabilities with which the utterances are classified as false or not-false, in each class of utterances

annotated as false, the classifier is less precise, but does assign to many more utterances a probability of being false >.5. The probability distribution of uncertain utterances lies in the middle between these two cases; in particular, the number of utterances whose probability is $.1 < p \leq .2$ is almost exactly halfway between the numbers for true and for false utterances. This suggests that the uncertain class does consist of a blend of true and false utterances, which creates some noise.

As already discussed, attempting to classify all the utterances of a hearing, while useful, does not necessarily reflect how our models would be used in a real life scenario. In the scenarios we envisage, the models would not be used to classify amounts of data so large that cannot be analyzed by humans directly. Every testimony where lies would have to be detected would have been previously examined by human analysts to identify utterances which need not be classified. These include statements such as questions, instructions, or greetings, which do not have propositional value and therefore they cannot be true or false. But these are also statements whose truthfulness is perfectly known, and therefore need not be classified. Therefore we can expect that in a practical situation several statements would be discarded and the dataset would be more similar to the data used in the second set of experiments, rather than the first.

6.1.4 Using more homogeneous data

The last round of experiments, run on subsets of DeCour, were aimed to verify if using more homogeneous data obtained by grouping defendants according to sex, native language and age could lead to better performance in classification task. The results of these studies do not show remarkable improvement in the effectiveness of the models, also because if in one hand the accuracy rises slightly, the baselines too are shifted upwards. Further analyses should be carried out, in order to gain a better

Table 15 Information gain of n-grams of lemmas in DeCour

	N-grams	Translation	IG value
1	non	not	0.0401
2	no	no	0.0212
3	sì.	yes.	0.0179
4	sì	yes	0.0179
5	per	for	0.0159
6	ricordare	to remember	0.0139
7	non ricordare	to not remember	0.0134
8	e	and	0.0126
9	dare	to give	0.0113
10	no.	no.	0.0107
11	o	or	0.0107
12	a	to/at	0.0101
13	ricordare.	to remember.	0.0091
14	da	from/by	0.0077
15	, non	, not	0.0074
16	non ricordare.	to not remember.	0.0072
17	non mi	I do not... (reflexive)	0.0070
18	sapere	to know	0.0066
19	, no	, not	0.0065
20	non avere	to not have	0.0060
21	in	in	0.0058
22	te	you (direct object)	0.0058
23	l'avere	to have... it	0.0057
24	non essere	to not be	0.0056
25	io non	I do not...	0.0055
26	non lo	not... it	0.0052
27	lo	it	0.0052
28	non l'avere	to not have... it	0.0051
29	avere	to have	0.0051
30	niente	nothing	0.0051
31	non lo sapere	to not know it	0.0049
32	e non	and not	0.0049
33	io	I	0.0049
34	non l'	not... it (in front of a wovel)	0.0047
35	lo sapere	to know it	0.0045
36	, ma	, but	0.0042
37	sapere.	to know.	0.0042
38	perché	because	0.0042
39	sì,	yes,	0.0041
40	me	me	0.0041
41	dire	to say	0.0039
42	, io	, I	0.0038

Table 15 continued

	N-grams	Translation	IG value
43	potere	can	0.0038
44	dare,	to give,	0.0038
45	ricordare,	to remember,	0.0036
46	non mi ricordare	I do not remember	0.0036
47	io l'avere	I... to have... it	0.0036
48	mi ricordare	I remember	0.0035
49	proprio	just	0.0035
50	da/di	by/of/with	0.0035

comprehension of the relation between deceptive language and variables such as sex, age and native language.

6.1.5 Linguistically more sophisticated models

Other methods to enhance the models' effectiveness are also possible. One way would be to use more linguistic information. For example, the texts could be parsed to collect syntactical features: in fact, there is some evidence that this kind of features can improve the performance in detecting deception (Feng et al. 2012). This syntactic information could be exploited using tree kernels, already applied to forensic tasks with good results (Giannone et al. 2009) but not yet employed in deception detection.

Finally, according to the Interpersonal Deception Theory—IDT (Buller and Burgoon 1996), speakers in conversations adapt their communication style to that of the interlocutor. Researchers working in other fields (not deception detection) evaluated the degree to which people coordinate their speech in dyadic interactions (Ireland et al. 2011; Niederhoffer and Pennebaker 2002): their approach could possibly be applied for feature selection in deception detection as well. (If the extra cognitive load caused by lying results in more stereotyped linguistic production, it is possible that liars may make more use of the words just heard from the interlocutor, as they are readily available in memory.)

6.2 The language of deception: the case of Italian

A second fruitful way to analyze our results and compare them with Newman et al. (2003) and other studies such as De Paulo et al. (2003) and Hauch et al. (2012) concerns the findings regarding the language used in lies and the difference from that used in truthful statements. Newman, Pennebaker and colleagues concluded that (lab-produced) deceptive language is characterized by fewer first-person singular pronouns, fewer third-person pronouns, more negative emotion words, fewer exclusive words, and more motion verbs. These findings were confirmed by most subsequent research on English. Newman, Pennebaker et al. also wondered about the cross-linguistic validity of these claims—in particular, they observed that

the claims about first-person singular pronouns ought to be tested in Romance languages that do not require a pronoun in many cases of use first-person verbs. The data used in this study allow us, first of all, to revisit these claims in a real, high-stakes setting; and second, to examine the claim about first-person pronouns as Italian is one of the Romance languages with the property mentioned by Newman et al. (2003).

Most informative n-grams The Information Gain measure of n-grams of lemmas employed in the previously discussed experiments can also be used to get some insight regarding the most typical stylistic traits of deceptive statements. As the goal in this case was to capture the profile of deceptive language rather than training models for the classification task, the whole DeCour was used to compute Information Gain. Only true and false utterances were considered, discarding the more confusing class of uncertain utterances. Table 15 shows the 50 most informative n-grams in DeCour. One obvious consideration is that expressions of negation or assertion, such as "yes" or "not" or statements of remembering or not remembering, of knowing or not knowing, are particularly revealing in deception detection.

However Information Gain does not indicate if a feature is more typical of true or false utterances. Table 16 contains the lists of the twenty most frequent tokens, bigrams, trigrams and tetragrams of true and false utterances.[10] The affirmative answer "yes" is highly frequent in true statements, but it does not appear among the 20 most frequent unigrams in deceptive utterances, as it is only found 111 times.

Conversely, in deceptive statements negative adverbs such as "no" and "not" are more frequent than in true ones, in spite of the fact that DeCour contains only 945 false utterances and 1202 true utterances. Phrases expressing not remembering or not knowing are present in both classes of utterances, but their use is definitely more common in the false ones. This difference becomes even clearer when we take into account the fact that many frequent bigrams are in fact part of frequent trigrams. So for example, out of the 69 bigrams "mi ricordo"/"I remember" found in the false utterances, 49 were actually produced as part of the trigram "non mi ricordo"/"I do not remember". This means that in DeCour the distribution of "mi ricordo" (not included in longer trigrams) and "non mi ricordo" among true and false utterances is as in the following Table:

	True utterances	False utterances
mi ricordo	16	20
non mi ricordo	20	49

The table clearly suggests that these phrases are used differently in true and false utterances although a χ^2 test carried out on this table produces a $p = .1715$, which is statistically not significant (mainly because of the small size of the data). As already

[10] "xxxxx" substitutes an anonymized token, such as proper names or surnames, names of places and so on.

Table 16 N-grams frequency in DeCour

Tokens	Freq.	Bigrams	Freq.	Trigrams	Freq.	Tetragrams	Freq.
True utterances							
sì	431	xxxxx xxxxx	66	**non mi ricordo**	20	mi ha detto che	4
che	389	c'era	53	c'era un	13	**non me lo ricordo**	4
xxxxx	327	mi hanno	40	che c'era	12	**ora non mi ricordo**	4
e	284	**mi ricordo**	36	mi ha detto	10	tant' F vero che	4
di	268	l'ho	32	**mi ricordo che**	9	a fare un giro	3
non	258	mi ha	31	xxxxx e xxxxx	9	altra parte della strada	3
mi	255	non mi	30	xxxxx xxxxx xxxxx	9	anche lui si dimenava	3
a	218	sono stato	30	c'era la	8	c' era la mia	3
la	217	un pò	29	**non lo so**	8	che c' era la	3
è	206	ho detto	28	io gli ho	7	ci hanno portato in	3
io	191	che non	27	mi hanno detto	7	dall' altra parte della	3
ho	185	che era	26	non ho mai	7	e mi ha detto	3
in	180	che mi	25	non è che	7	ho detto anche al	3
era	174	quello che	25	un pò di	7	ho detto che non	3
sono	168	a xxxxx	24	xxxxx xxxxx e	7	ho visto un' auto	3
il	160	io non	24	ce l'ho	6	in entrambi i sensi	3
un	144	io ho	23	ci hanno portato	6	in provincia di xxxxx	3
l'	120	non lo	23	gli ho detto	6	l' ho detto anche	3
perché	116	e mi	21	ho detto che	6	la pattuglia della polizia	3
no	102	di xxxxx	20	mi hanno fatto	6	non ce l' ho	3
False utterances							
non	644	l'ho	85	**non mi ricordo**	49	non l' ho mai	9
che	394	non mi	84	**non lo so**	38	**non me lo ricordo**	9
ho	317	**mi ricordo**	69	non l'ho	28	che a me mi	8
e	302	**non ricordo**	68	non è che	17	a me mi risulta	6
mi	302	io non	61	io l'ho	16	io non ho mai	6
io	291	non lo	60	mi ha detto	16	**io non mi ricordo**	6
è	235	ho detto	53	io non ho	14	**non mi ricordo proprio**	6
no	222	non è	53	non ho mai	14	a me non mi	5
di	220	non ho	51	il mio amico	13	ad un certo punto	5
xxxxx	214	lo so	41	l'ho visto	13	non l' ho visto	5
la	196	mi ha	41	gli ho detto	12	**non mi ricordo non**	5
a	186	xxxxx xxxxx	37	**me lo ricordo**	10	io l' ho allontanato	4
perché	180	non l'	36	non me lo	10	io l' ho detto	4
l'	178	che mi	35	a me mi	9	io non l' ho	4
ricordo	162	a me	34	a me non	9	**io non lo so**	4
il	156	**non so**	33	che a me	9	non lo so perché	4
sono	149	ho visto	30	ho detto che	9	perché non che	4
un	140	c'era	28	l'ho mai	9	a che fare con	3
era	132	che no	27	me l'ha	9	**adesso non mi ricordo**	3
in	123	mi hanno	27	non c'era	9	allora gli ho detto	3

Table 17 LIWC categories most prevalent in true utterances

LIWC dimensions	False utterances' mean values	True utterances' mean values	Difference
Certainty	0.0973	0.2681	−0.1708
Prepositions	0.1472	0.1691	−0.0219
Space	0.0256	0.0348	−0.0093
Time	0.0603	0.0669	−0.0066
Home	0.0028	0.0086	−0.0058
Positive feelings	0.0160	0.0217	−0.0057
Leisure	0.0047	0.0094	−0.0047
Numbers	0.0067	0.0102	−0.0036
Nonfluencies	0.0015	0.0047	−0.0033
Optimism and energy	0.0066	0.0096	−0.0030
Occupation	0.0068	0.0093	−0.0024
We	0.0072	0.0096	−0.0024
Work	0.0026	0.0048	−0.0022
Past tense verb	0.0904	0.0920	−0.0017
They verb	0.0196	0.0209	−0.0014
Money	0.0034	0.0046	−0.0012
Eating, drinking, dieting	0.0021	0.0032	−0.0011
School	0.0002	0.0012	−0.0010
Friends	0.0029	0.0038	−0.0009
Inhibition	0.0040	0.0047	−0.0007

discussed in 4.2.4, this difference is to be expected in a hearing scenario, where a defendant's lies will be most likely in the forms of denials of true accusations.

Association between lies and LIWC categories Newman et al. (2003) summarize their main findings about deceptive language as follows:

> liars tend to tell stories that are less complex, less self-relevant, and more characterized by negativity.

We can verify whether these findings by Newman et al. about deceptive language still hold for our data thanks to the Italian version of LIWC that we used to compute lexical features. The mean values of the LIWC dimensions with the greatest differences in value for true and false utterances are shown in Tables 17 and 18, ordered according to the difference between the values of the two categories (in particular, this difference concerns the means of the normalized frequencies of each LIWC dimension in true and false utterances).

Our conclusions (see previous subsection) about the prevalence of positive statements among true utterances and of negative statements among false ones are confirmed by the fact that the greatest differences among false and true utterances lie in the LIWC dimensions Certainty (with substantially higher value among true utterances) and Negation (viceversa). Confirming the results of Newman et al. (2003), false utterances have higher values for the dimensions Negative Emotions,

Table 18 LIWC categories most prevalent in false utterances

LIWC dimensions	False utterances' mean values	True utterances' mean values	Difference
Negations	0.2682	0.0742	0.1940
Cognitive processes	0.1794	0.0997	0.0797
Present	0.2146	0.1454	0.0692
I verb	0.1580	0.0957	0.0623
Total pronouns	0.1885	0.1473	0.0412
Transitive	0.0527	0.0192	0.0335
I	0.1099	0.0794	0.0305
Introspection	0.0584	0.0353	0.0231
To have	0.0561	0.0336	0.0225
Perceptual processes	0.0537	0.0316	0.0221
If	0.0642	0.0485	0.0157
Discrepancy	0.0309	0.0162	0.0147
Past participle	0.0764	0.0622	0.0142
Causation	0.0382	0.0270	0.0112
Communication	0.0452	0.0354	0.0098
Exclusive	0.1044	0.0946	0.0098
Negative emotion	0.0209	0.0112	0.0097
Articles	0.1735	0.1642	0.0093
Hearing	0.0304	0.0214	0.0091
Seeing	0.0148	0.0067	0.0082

Exclusive and Discrepancy. They also have higher values for content expressing cognitive/perceptual processes (expressed by LIWC dimensions such as Cognitive processes, Perceptual processes, Introspection, Hearing and Seeing). True utterances have greater values for references to time, space, concrete topics (dimensions such as Home, Leisure, Work, School, Friends) and positive feelings.

A particularly interesting finding is the greater presence among false utterances of personal pronouns in general, and in particular of first person pronouns, as showed by the greater use of "Io"/"I" and "me"/"me". This finding is interesting because it goes against the recurrent finding in the literature that people, when they lie, are prone to use other-references rather than self-references (Hancock et al. 2008; Newman et al. 2003).

In Italian, as in other Romance languages, subject pronouns can be omitted. Therefore if it is a general truth that deceptive language tends to contain less self-references than truthful languages, one would expect to find an even lower rate of self-references in Italian than in English. The distribution of pronouns in DeCour would therefore seem to be inconsistent with the previous literature.

In order to investigate in depth this discrepancy, DeCour was parsed making use of the online service Tanl Italian Parser offered by the University of Pisa.[11] Minor errors in the output of the parser were then hand-corrected using simple heuristic

[11] http://paleo.di.unipi.it/it/parse.

Table 19 First person pronouns and verbs in true and false utterances

	False utterances	True utterances
Number	945	1,202
Tokens	15,924	15,456
Pronoun "Io"–"I"	291	191
First person pronouns ("Io"–"I", "me/mi"-"me")	393	257
First person verbs	1,057	756
First person verbs without pronouns	664	499
Pronoun "Io"–"I" without verb	7	12
First person pronouns without verb	26	34
Ratio first person pronouns/number of utterances	0.4158	0.2138
Ratio first person pronouns/number of tokens	0.0246	0.0166
Ratio pronoun "Io"/first person verbs	0.2753	0.2526
Ratio first person pronouns/first person verbs	0.3718	0.3399
Ratio first person verbs without pronouns/first person verbs	0.6282	0.6601
Ratio first person verbs/number of utterances	1.1185	0.6290
Ratio first person verbs/number of tokens	0.0664	0.0489

rules, in particular in order to fix the problems caused to the parser by the ambiguity of "ricordo" (which can be used both as a name—"memory"—or as first person of the verb "I remember") and of "sono" (which without pronoun can be the first singular or the third plural person of the verb "to be"). The statistics about first person pronouns among false and true utterances including also the dropped first person pronouns that we obtained in this way are summarized in Table 19.

As shown by the Table, only 37.2 % first-person verbs in Italian have a subject pronoun. But irrespective of whether we count the percentage of first-person pronouns per utterance, or the percentage of first-person verbs, the reduced number of self-references found by Newman et al. (2003) and others in deceptive language is not confirmed for our data.

We found however one construction in which the difference between deceptive and truthful language lies in the greater use of first-person pronouns in true statements. The common statement "I do not remember" can be expressed in Italian either as "[io] non ricordo" or in so-called 'reflexive form' "[io] non *mi* ricordo". In general the reflexive form is of more common use in Italian, and this preference is maintained in true utterance, where the reflexive form "non mi ricordo" is used three times as much as the non-reflexive form "non ricordo," which is only used 6 times. But with false utterances, the preference is reversed: "non ricordo" is used 68 times, as opposed to 49 times for "non mi ricordo". The situation can be summarized as in the following table.

	True utterances	False utterances
non mi ricordo	20	49
non ricordo	6	68

The χ^2 test (equal expected counts) gives a $p = 0.0025$ for this contingency table, highly significant. In other words, the bigram "non ricordo" is an excellent clue of deception.

7 Conclusions

To our knowledge, this is the first study in Italian to report on the use of deceptive language in such a high-stakes setting as a court, and one of the first studies anywhere. For what concerns the perspective of automatic deception detection, the results of our models suggest that stylometric techniques such as those previously used for lab-produced deceptive language can be effective even when the deceptive communication takes place in natural settings and when attempting to classify short text such as utterances as opposed to full hearings. Furthermore, we found that comparable results can be obtained using lexical features and surface features, opening the way to the application of such techniques to languages for which the LIWC is not available. But whereas our models achieve high precision at identifying false statements, recall needs to be improved—i.e., additional markers of deception have to be discovered.

Regarding deceptive language, we could verify many of the findings of previous studies concerning deception markers, which suggests that the cognitive elaboration of deception is basically the same in English and Italian in spite of the different native language of the speakers. We couldn't find however support for one of the recurrent findings in the previous literature, the reduced use of self-referring expressions in deceptive language—in fact, we found the opposite.

Acknowledgments To create DeCour has been very complex, and it would not have been possible without the kind collaboration of a lot of people. Many thanks to Dr. Francesco Scutellari, President of the Court of Bologna, to Dr. Heinrich Zanon, President of the Court of Bolzano, to Dr. Francesco Antonio Genovese, President of the Court of Prato and to Dr. Sabino Giarrusso, President of the Court of Trento.

References

Adams SH (1996) Statement analysis: what do suspects' words really reveal? FBI Law Enforc Bull 65(10):12–20

Alparone F, Caso S, Agosti A, Rellini A (2004) The Italian LIWC2001 dictionary. LIWC.net, Austin

Artstein R, Poesio M (2008) Inter-coder agreement for computational linguistics. Comput Linguist 34(4):555–596

Bachenko J, Fitzpatrick E, 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, pp 41–48, Stroudsburg, PA, USA. Association for Computational Linguistics

Bond CF, De Paulo BM (2006) Accuracy of deception judgments. Pers Soc Psychol Rev 10(3):214–234

Buller D, Burgoon J (1996) Interpersonal deception theory. Commun Theory 6:203–242

Chinchor N (1992) Muc-4 evaluation metrics. In: Proceedings of the 4th conference on message understanding, MUC4 '92, pp 22–29, Stroudsburg, PA, USA. Association for Computational Linguistics

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
Coulthard M (2004) Author identification, idiolect, and linguistic uniqueness. Appl Linguist 25(4):431–447
Davatzikos C, Ruparel K, Fan Y, Shen D, Acharyya M, Loughead J, Gur R, Langleben D (2005) Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. NeuroImage 28(3):663–668
De Paulo BM, Lindsay JJ, Malone BE, Muhlenbruck L, Charlton K, Cooper H (2003) Cues to deception. Psychol Bull 129(1):74–118
Ekman P (2001) Telling lies: clues to deceit in the marketplace, politics, and marriage. W.W. Norton
Feng S, Banerjee R, Choi Y (2012) Syntactic stylometry for deception detection. In: Proceedings of the 50th annual meeting of the association for computational linguistics (volume 2: Short Papers), pp 171–175, Jeju Island, Korea. Association for Computational Linguistics
Fitzpatrick E, Bachenko J (2009) Building a forensic corpus to test language-based indicators of deception. Lang Comput 71(1):183–196
Fitzpatrick E, Bachenko J (2012) Building a data collection for deception research. In: Proceedings of the EACL workshop on computational approaches to deception detection, pp 31–38, Avignon, France. Association for Computational Linguistics
Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3:1289–1305
Fornaciari T, Poesio M (2011) Sincere and deceptive statements in Italian criminal proceedings. In: Proceedings of the international association of forensic linguists 10th biennial conference, pp 126–138, Cardiff, Wales, UK
Frank MG, Feeley TH (2003) To catch a liar: challenges for research in lie detection training. J Appl Commun Res 31(1):58–75
Frank MG, Menasco MA, O'Sullivan M (2008) Human behavior and deception detection. In: Voeller JG (ed) Wiley handbook of science and technology for homeland security. Wiley, New York
Ganis G, Kosslyn S, Stose S, Thompson W, Yurgelun-Todd D (2003) Neural correlates of different types of deception: an fMRI investigation. Cereb Cortex 13(8):830–836
Giannone C, Basili R, Del Vescovo C, Naggar P, Moschitti A (2009) Kernel-based relation extraction from investigative data. In: Proceedings of the third workshop on analytics for noisy unstructured text data, AND '09, pp 93–100, New York, NY, USA. ACM
Gokhmann S, Hancock J, Prabhu P, Ott M, Cardie C (2012) In search of a gold standard in studies of deception. In: Fitzpatrick E, Bachenko J, Fornaciari T (eds) Proceedings of the EACL workshop on computational approaches to deception detection, pp 23–30
Hancock JT, Curry LE, Goorha S, Woodworth M (2008) On lying and being lied to: a linguistic analysis of deception in computer-mediated communication. Discourse Process 45(1):1–23
Hauch V, Blandón-Gitlin I, Masip J, Sporer SL (2012) Linguistic cues to deception assessed by computer programs: a meta-analysis. In: Fitzpatrick E, Bachenko J, Fornaciari T (eds) Proceedings of the workshop on computational approaches to deception detection, pp 1–4, Avignon
Ireland ME, Slatcher RB, Eastwick PW, Scissors LE, Finkel EJ, Pennebaker JW (2011) Language style matching predicts relationship initiation and stability. Psychol Sci 22(1):39–44
Jensen ML, Meservy TO, Burgoon JK, Nunamaker JF (2010) Automatic, multimodal evaluation of human interaction. Group Decis Negot 19(4):367–389
Karatzoglou A, Meyer D, Hornik K (2006) Support vector machines in r. J Stat Softw 15(9):1–28
Koppel M, Schler J, Argamon S, Pennebaker J (2006) Effects of age and gender on blogging. In: AAAI 2006 spring symposium on computational approaches to analysing weblogs
Levine TR, Feeley TH, McCornack SA, Hughes M, Harms CM (2005) Testing the effects of nonverbal behavior training on accuracy in deception detection with the inclusion of a bogus training control group. West J Commun 69(3):203–217
Lord RD (1958) Studies in the history of probability and statistics.: Viii. de morgan and the statistical study of literary style. Biometrika 45(1/2):282–282
Lutoslawski W (1898) Principes de stylomtrie. Revue des tudes grecques 41:61–81
Luyckx K, 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, pp 513–520, Stroudsburg, PA, USA. Association for Computational Linguistics
Merikangas JR (2008) Commentary: functional mri lie detection. J Am Acad Psychiatry Law 36(4):499–501

Mosteller F, Wallace D (1964) Inference and disputed authorship: the federalist. Addison-Wesley, Reading
Newman ML, Pennebaker JW, Berry DS, Richards JM (2003) Lying words: predicting deception from linguistic styles. Pers Soc Psychol Bull 29(5):665–675
Niederhoffer KG, Pennebaker JW (2002) Linguistic style matching in social interaction. J Lang Soc Psychol 21(4):337–360
Peersman C, Daelemans W, Van Vaerenbergh L (2011) Age and gender prediction on netlog data. Presented at the 21st Meeting of Computational Linguistics in the Netherlands (CLIN21), Ghent, Belgium.
Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count (LIWC): LIWC2001. Lawrence Erlbaum Associates, Mahwah
Pepe G (ed) (1996) La falsa donazione di Costantino. Tea storica. TEA
Porter S, Woodworth M, Birt AR (2000) Truth, lies, and videotape: an investigation of the ability of federal parole officers to detect deception. Law Hum Behav 24(6):643–658
Sasaki Y (2007) The truth of the F-measure. Teach Tutor mater, pp 1–5
Schmid H (1994) Probabilistic part-of-speech tagging using decision trees. In: Proceedings of international conference on new methods in language processing
Simpson JR (2008) Functional mri lie detection: too good to be true? J Am Acad Psychiatry Law 36(4):491–498
Solan LM, Tiersma PM (2004) Author identification in american courts. Appl Linguist 25(4):448–465
Stein B, Koppel M, Stamatatos E (2007) Plagiarism analysis, authorship identification, and near-duplicate detection pan'07. SIGIR Forum 41:68–71
Strapparava C, 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
Undeutsch U (1967) Beurteilung der Glaubhaftigkeit von Aussagen [Veracity assessment of statements]. In: Undeutsch U (ed) Handbuch der psychologie: vol 11. Forensische Psychologie. Hogrefe, Gottingen, pp 26–181
Undeutsch U (1982) Statement reality analysis. In: Trankell A (ed) Reconstructing the past: the role of psychologists in criminal trials. Kluwer, Deventer, pp 27–56
Undeutsch U (1984) Courtroom evaluation of eyewitness testimony. Appl Psychol 33(1):51–66
Vaassen F, Daelemans W (2011) Automatic emotion classification for interpersonal communication. In: 2nd workshop on computational approaches to subjectivity and sentiment analysis (WASSA 2.011)
Vrij A (2008) Detecting lies and deceit: pitfalls and opportunities. Wiley series in psychology of crime, policing and law, 2nd edition. Wiley, Chichester
Vrji A (2005) Criteria-based content analysis—a qualitative review of the first 37 studies. Psychol Public Policy Law 11(1):3–41
Walczyk JJ, Roper KS, Seemann E, Humphrey AM (2003) Cognitive mechanisms underlying lying to questions: response time as a cue to deception. Appl Cogn Psychol 17(7):755–774
Wang JT, Spezio M, Camerer CF (2010) Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. Am Econ Rev 100(3):984–1007
Yang Y, Liu X (1999) A re-examination of text categorization methods. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '99. ACM, New York, pp 42–49
Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. CiteSeerX—Scientific Literature Digital Library and Search Engine [http://citeseerx.ist.psu.edu/oai2] (United States)
Zhou L, Shi Y, Zhang D (2008) A statistical language modeling approach to online deception detection. IEEE Trans Knowl Data Eng 20(8):1077–1081