UNIVERSITY OF TRENTO

CiMeC
Center for Mind/Brain Sciences

DOCTORAL SCHOOL

Cognitive and Brain Sciences – XXV cycle

PhD Dissertation

Deception Detection in Italian Court testimonies

Tommaso Fornaciari

Advisor
Prof. Massimo Poesio

Head of the School
Prof. David Melcher

December 2012

Acknowledgements

The research activity described in this dissertation was complex and would not have been possible without the helpfulness of several people at each step of its realization.

Prof. Roberto Cubelli of the University of Trento gave me the first advices about the Doctoral School and about the cultural path I could have followed.

During the phase of feasibility study, I received remarkable advices from Dr. Biagio Mazzeo, Public Prosecutor at the Public Prosecutor's Office of Genova, from Dr. Michela Guidi, Public Prosecutor at the Public Prosecutor's Office of Forlì, from Dr. Piero Tony, Chief Prosecutor of the Public Prosecutor's Office of Prato, and from Dr. Rosario Minna, Chief Prosecutor of the Public Prosecutor's Office of Ferrara. Without their ideas and their knowledge of Italian legal system, I would not have found the way to realize this project.

I also found full cooperation at the Courts where the data were collected. Dr. Francesco Scutellari, President of the Court of Bologna, Dr. Heinrich Zanon, President of the Court of Bolzano, Dr. Francesco Antonio Genovese, President of the Court of Prato and Dr. Sabino Giarrusso, President of the Court of Trento readily authorized the exam of the Court files and the data collection, allowing the creation of DeCour.

Into the same Courts, I had the fortune to meet very kind people - I would say guardian angels -, who helped me to solve several bureaucratic and practical problems. They are Carmelo Di Pietro of the Court of Bologna, Dr. Manfred Kelder of the Court of Bolzano, Rita Fava of the Public Prosecutor's Office of Prato and Dr. Sandro Pettinato of the Court of Trento.

During my learning activities I was strongly supported by Prof. Marco Baroni of the University of Trento, Prof. Carlo Strapparava of the Fondazione Bruno Kessler (FBK) and Prof. Alessandro Moschitti of the University of Trento, who gave me the possibility of benefiting from their great expertise in computational linguistics.

I was kindly hosted for one month by Prof. Walter Daelemans at his CLiPS - Computational Linguistics and Psycholinguistics research center of the University of Antwerp. I enjoyed his warm hospitality and I received from him and his working group a lot of precious feed-backs, which helped me to improve the data analysis.

During a conference in U.S., I had the fortune to meet Prof. Eileen Fitzpatrick, who believed in my work and gave me the possibility to join her and Joan Bachenko in the organization of the Workshop on Computational Approaches to Deception Detection, held at the 13th Conference of the European Chapter of the Association for Computational Linguistics. This was one of the most rewarding experiences of the PhD path.

Someone claims it is not elegant to thank one's own advisor, as to follow the student's activities is his duty. This is true, but it is also true that the same duty can be fulfilled in several ways. Not only Prof. Massimo Poesio was an extremely well skilled and helpful advisor, but also I had the possibility to appreciate his moral integrity and his disinterested commitment for the development of the students' career. Moreover, I think we were surprisingly on the same wavelength. I felt we had no limbering up: we started working together easily as if we were doing it since long time.

Of course a PhD does not end with the dissertation, but it should be the first step of a longer way. Massimo Poesio, Dr. Carlo Bui of Italian National Police and my uncle Daniele Fornaciari are giving me a great support for the ongoing work, and their help is really precious.

Lastly, I am aware that sometimes it is not easy to be daily close to me, especially when I am fully absorbed by my work. Even so, I got constant comprehension and fondness from my family, in particular from my mother, Maura Donini, and from my girlfriend Heidi Pedersen.

I am deeply grateful to the people I mentioned. All of them did for me something more than what I asked for, or what I would have reasonably expected. In this way, not only they made possible the realization of an idea which without them would not have succeeded. But also, looking backwards to these three years, I realize that their true gift was beyond what they did for me: it was their human warmth, and to make me feel rich in their friendship. I honestly think I cannot to repay this debt but, as my uncle Daniele is used to say, "It is beautiful to be in debt with people we love".

Abstract

Effective methods for evaluating the reliability of statements issued by witnesses and defendants in hearings would be extremely valuable 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 laboratory conditions still has to be properly assessed.

DeCour - DEception in COURt corpus - has been built with the aim of training models suitable to discriminate, from a stylometric point of view, between sincere and deceptive statements. DeCour is a collection of hearings held in four Italian Courts, in which the speakers lie in front of the judge. These hearings become the object of a specific criminal proceeding for calumny or false testimony, in which the deceptiveness of the statements of the defendant is ascertained. Thanks to the final Court judgment, that points out which lies are told, each utterance of the corpus has been annotated as true, uncertain or false, according to its degree of truthfulness. Since the judgment of deceptiveness follows a judicial inquiry, the annotation has been realized with a greater degree of confidence than ever before. In Italy this is the first corpus of deceptive texts not relying on 'mock' lies created in laboratory conditions, but which has been collected in a natural environment.

In this dissertation we replicated the methods used in previous studies but never before applied to high-stakes data, and tested new methods. Among the best known proposals in this direction are methods proposed by Pennebaker and colleagues, who employed their lexicon - the Linguistic Inquiry and Word Count (liwc) - to analyze different texts or transcriptions of spoken language, in which deception could have been used, but collected in an artificial way. In our experiments, we trained machine learning models relying both on lexical features belonging to liwc and on surface features. The surface features were selected calculating their Information Gain, or simply according to the frequency they appear in the texts. We also considered the effect of a number of variables including the degree of certainty the utterances were annotated as truthful or not and the homogeneity of the dataset. In particular, the classification task of false utterances was carried out against the only utterances annotated as true, or against the utterances annotated as true and as uncertain together. Moreover subsets of DeCour were analysed, in which the statements were issued by homogeneous categories of subject, e.g. speakers of the same gender, age or native language. Our results suggest that accuracy at deception detection clearly above chance level can be obtained with real-life data as well.

Publications

Parts of this thesis appeared in the following previous publications:

- Fornaciari, T. and Poesio, M. (2011a). Lexical vs. surface features in deceptive language analysis. In *Proceedings of the ICAIL 2011 Workshop Applying Human Language Technology to the Law*, AHLTL 2011, pages 2–8, Pittsburgh, USA

- Fornaciari, T. and Poesio, M. (2011b). Sincere and deceptive statements in italian criminal proceedings. In *Proceedings of the International Association of Forensic Linguists Tenth Biennial Conference*, IAFL 2011, Cardiff, Wales, UK

- Fornaciari, T. and Poesio, M. (2012b). On the use of homogenous sets of subjects in deceptive language analysis. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 39–47, Avignon, France. Association for Computational Linguistics

- Fornaciari, T. and Poesio, M. (2012a). DeCour: a corpus of DEceptive statements in Italian COURts. In Calzolari, N. C. C., Choukri, K., Declerck, T., Uäÿur Döäÿn, M., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA)

Contents

List of Tables

List of Figures

Testis falsus non erit impunitus,
et qui mendacia loquitur non effugiet.

[Proverbia, 19, 5 Nova Vulgata]

Non nobis, Domine, non nobis,
sed nomini tuo da gloriam.

[Psalmi, 113, 9 Nova Vulgata]

Part I

Introduction and background

Chapter 1

Introduction

In the last twenty years, forensic sciences have grown exponentially, in Italy as in other countries. Police investigations avail themselves more and more of - and also depend on - a wide variety of branches of the science. Modern forensic biology, which relies on DNA analyses and is one of the most revolutionary disciplines for the practice of crime scene investigations, was introduced in Italy in 1989, when the first exams using the technique of Polymerase Chain Reaction (PCR) were carried out at Forensic Science Police Service of the Italian National Police (Polizia di Stato, 2003). Chemistry and physics support the inquiries with a number of methodologies aimed to accomplish several tasks, such as revealing latent fingerprints, or identifying materials, tool-marks and so on (Paceri and Montanaro, 1995). The introduction of data-bases such as Automatic Fingerprint Identification System (AFIS) and Integrated Ballistic Identification System (IBIS), or of expert systems such as the Italian Crime Scene Analysis System (Italian acronym, SASC) and the use of technologies for 3D crime scene reconstruction also transformed radically police investigations, in comparison with a relatively recent past.

Currently it is not far from the truth to assert that any kind of evidence can be the object of technical exams, in order to draw out of it as more information as possible. In this scenario, one could expect that scientific methods are also applied to analyze testimonies issued by people variously involved in criminal proceedings. In the United States a controversial debate regarding the admissibility in Court of physiological measures in order to assess the truthfulness of testimonies began officially on 1923, in the famous case of Frye vs. United States (Saxe and Ben-Shakhar, 1999). This debate did not take place in Italy, where the use of this kind of technologies, based on the polygraph - better known as 'lie detector' -, has never been allowed for reasons of principle (Maffei, 2007). The present formulation of this principle is contained into the art. 188 of Criminal Proceedings Code,[1]

[1]The Criminal Proceedings Code reads:

> Art.188 - Libertà morale della persona nell'assunzione della prova.
>
> Non possono essere utilizzati, neppure con il consenso della persona interessata, metodi o tecniche idonei a influire sulla libertà di autodeterminazione o ad alterare la capacità di ricordare e di valutare i fatti.

which defends the so-called 'moral freedom' of the subjects. This concept, coming from the philosophy of Enlightenment, refers to the faculty of citizens of determining freely (every kind of) their answers in front of the law, and this possibility would be denied employing tools which detect involuntary reactions to the questions. Nevertheless, the Implicit association test (Iat), which relies on reaction times, was employed in a criminal proceeding of 2009 for the first time in Italy, to evaluate the reliability of the narrative of a woman victim of violence (Agosta et al., 2011). However, in spite of this recent case, in the context of police investigations the use of technologies specifically intended to extract from testimonies information useful for the inquiry and which otherwise could not be obtained, is still surprisingly rare.

More precisely, in several kinds of police interviews the collection of information can be carried out through qualified procedures, such as the Cognitive Interview (Fisher and Geiselman, 1992), which was specifically conceived to be employed directly by police officers and to enhance the recovery of the memories of the witnesses. But the Cognitive Interview is a framework for the interviews rather than a tool for data analysis, and its effectiveness depends on the cooperation of the subject interviewed. Mostly in the context of expert witness surveys, and in particular in the evaluation of minors' testimonies, tools such as the Criteria-Based Content-Analysis (CBCA) are usually employed (Vrji, 2005). CBCA is aimed to evaluate the reliability of a testimony, but it is usually applied by psychologists and its result ultimately depends on human evaluation. As a result, for a set of legal, practical and historical reasons, in Italy police investigations lack qualified and easily employable tools which can give a support in testimony evaluation.

The present research project was born from the necessity of filling this gap, implementing a tool aimed to evaluate the truthfulness of witnesses' statements, taking into account the practical and procedural constraints of the Italian legal system. The idea of dealing with polygraph or neuro-imaging techniques was discarded for practical reasons. In spite of the recent evolution of the jurisprudence regarding these approaches, their applicability in Court is limited and in any case their use depends on the availability both of the instruments and of the agreement of subjects to be examined in this way. This means that, even though they are very interesting from a scientific point of view, at least in a mid-term perspective they are realistically doomed to not affect the daily practice of police investigations. By contrast, one of the main ideas of this project is to focus on data which can be easily collected by police forces. On the one hand, this can be considered the approach of the drunk man who looked for the keys of his house under a lamp, because this was the

only place where he could see something. On the other hand, it could be worth to explore the possibility of finding clues of deception in data which are available, especially because this is done carrying out analyses never tried in Italy. Furthermore, as discussed in the next Chapter, the methodological approach which characterizes this study turned out to be not less - and possibly more - effective than other approaches. Hence the decision of analyzing testimonies relying on what these are constituted by: the words issued by the subjects.

1.1 Forensic linguistics

Forensic Linguistics is the branch of linguistics which deals with forensic issues. It is still considered as a rather new field: the expression 'forensic linguistics' itself was used for the first time in 1968 by Jan Svartvik concerning a classic study regarding statements in the Evans case (Svartvik, 1968). In this study, Svartvik analysed four different testimonies of Timothy Evans, finding that two of them could not be a spontaneous production of Evans, but they were altered. Svartvik was able to draw this conclusion from stylistic differences in the use of the grammar in different statements.[2] This was the first case of linguistic analyses applied to a forensic task, and it is considered the birth of this new discipline (Coulthard, 2004). The principal intervention areas of forensic linguistics are typically:

1. Authorship attribution or comparison, regarding one or more texts of uncertain authorship, of which the possible author or authors are known;
2. Authorship profiling, regarding one or more anonymous texts, of which the author is unknown, with the goal of finding out personal information e.g. age, sex, culture and geographical origin;
3. Plagiarism analysis, regarding texts with a certain authorship, but which could be the result of plagiarism of another text;
4. Analysis of veracity or deceptiveness of statements;
5. Phonetic analysis with the goal of identifying the speaker in lawful interception;
6. Analysis of legal text, such as Police reports or Court judgments, with the aim to deal with several topics, such as the meaning of words in a legal context.

[2] Timothy Evans was almost illiterate, and he gave two genuine testimonies, and two other testimonies whose the content was strongly influenced by John Christie, author of several crimes.

Forensic linguistics is a discipline strongly oriented to solve practical problems in a judiciary context, and includes tasks which are also very different from each other. As a consequence, forensic linguistics draws on different means to evaluate the texts object of analysis, from hermeneutical skills to statistical resources, and although it also has at its disposal powerful tools, especially from a computational point of view, it is not always easy to ground a conclusion in solid theoretical foundations (Solan and Tiersma, 2004). For example, the possibility of identifying the author of a text, as asserted by Coulthard, is based on the assumption that everyone has a specific and absolutely individual way of using the language, so called idiolect, which allows for distinguishing every subject from another one (Coulthard, 2004). But this assumption is put in doubt by Olsson, who thinks that it is not correct to assume the existence of an individual idiolect just because it is possible to differentiate the authorship of different writers, for example because in the course of time intra-individual stylistic differences could be found, and he claims: 'it may be better to focus on *distinctive* rather than *unique* style' (Olsson, 2008).

In Italy, the activities of voice comparison are usually carried out by forensic experts. Studies related to the analysis of legal texts are also present, but they are not yet widespread and they regard the analysis of the judiciary language, rather than the forensic investigations. An example of this approach can be represented by the group of Prof. Bellucci, of University of Florence.[3] By contrast, the application of modern tools of analysis to spoken or written linguistic productions of forensic interest is not yet the object of in-depth study. Therefore, as far as we know, this research project is the first attempt to study Italian texts employing the methodology of computational linguistics applied to forensic tasks.

1.2 The Morellian method

In spite of the technological novelty for Italian investigations, there is a basic methodological continuity between forensic linguistics and the methods used in police investigations to identify the unknown author of a crime. This is the reliance on the so-called 'Morellian method'. This is the approach applied by Giovanni Morelli to the study of works of art (Morelli, 1880), and identified for the first time as a theoretical framework useful for criminal analysis by Carlo Bui (Bui, 2006). According to Giovanni Morelli, to attribute an anonymous work of art to its author, it is necessary to focus on the more negligible

[3] http://www.patriziabellucci.it/laligi.htm

details, such as ear lobes, nails, the form of the hands and of the fingers, and not on the more striking features, that are more easily imitated. The personality has to be searched 'where the personal effort is weaker', and then where the expression is more spontaneous. Freud agreed with Morelli, saying that the Morellian method is 'strictly related to the approach of medical psychoanalysis. The psychoanalytical method is also used to penetrate secret and hidden things on the basis of unappreciated or unperceived elements, debris or 'rubbish' of our observation' (Freud, 1913). In this perspective, computational linguistics and other sciences applied to forensic tasks are based on the same principle: looking for details neglected by the author of a crime, which can be revealing about his identity. Therefore, the meaning of this research project is to experiment new methods for criminal analysis, focusing on the relatively unexplored field of analysis of testimonies in criminal proceedings, relying on the approach of computational linguistics.

1.3 Object of the research

The focus of this dissertation 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 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 laboratory 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 Assessment (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 dissertation 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.

The structure of the thesis is as follows. In Chapter 2 the previous work in this area is discussed. In Chapter 3 our dataset is described in detail. Chapter 4 presents our machine learning models and experimental methods. Chapter 5 gives some more information about the practical realization of the research project. The results of the experiments are presented in Chapter 6 and discussed in Chapter 7.

Chapter 2

Background

2.1 Detecting deception

Detecting deception in communication is a challenge for humans. Human performance at recognizing deception was studied in several applications of forensic interest (Meissner and Kassin, 2002), even in high-stakes scenarios (Garrido et al., 2002). Human skills were found to be not much better than chance in a number of studies (Bond and De Paulo, 2006). Deception researchers also tried to develop lie detection training procedures to be employed in forensic settings (Kassin and Fong, 1999), but Levine et al. (2005) claim that even specific training is not particularly effective to improve the ability of subjects. On the other end, there are studies suggesting that the ability of humans as lie-detectors is underestimated (Frank and Feeley, 2003). For example, O'Sullivan and Ekman (2004) found that some people - they call 'wizards' - are particularly skilled in detecting deception, but Bond and Uysal (2007) analyzing their results conclude that 'chance can explain results that the authors attribute to wizardry' and 'no truly diagnostic procedure for identifying wizards has ever been reported'. 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; Vrij, 2008).

Because of the existence of many modes of gathering evidence of human behavior, and probably also because of the difficulty of the task itself, a wide variety of approaches to discover deceptive statements have been tried. They can be very different from each other, but all of them involve two steps:

- To identify in the communicative act some clues of deceptiveness;
- To verify their correlation with deceptive and truthful communication.

While the first choice determines the method of study, to test the correlation between cues and deception implies to establish the ground truth which is object of deception. This depends on the context in which the deceptive communication takes place, and hence determines the object of study, or in other word its domain dependence.

As far as the clues of deception are concerned, the most ancient (but not necessarily simplest) approach is based on hermeneutics, that is on the analysis of the content of

the spoken or written language. Present day, the range of clues which can be object of investigation is wider (Vrij, 2008). The literature about deceptive communication can be divided in three main branches:

- Studies focused on Verbal behavior;
- Studies focused on Non-Verbal behavior;
- Recent studies based on physiological variables, and in particular on neuro-imaging techniques.

Regarding the object of study, the literature can be divided in two main families:

- Studies relying on data collected in natural environment;
- Studies relying on data collected in laboratory;

Field studies are usually interesting in forensic applications, because of their realistic nature. The psychological conditions of the subjects are genuine, and this is relevant especially in high-stakes settings. Unfortunately in these studies to establish the ground truth is often difficult (if not impossible). Thence with few exceptions - such as the study of the group of Fitzpatrick (Bachenko et al., 2008; Fitzpatrick and Bachenko, 2009) - in literature it is not easy to find balanced data sets in which deception and truth are comparable (Vrij, 2008; Zhou et al., 2008). By contrast, laboratory studies are characterized by the artificiality of participants' psychological conditions. They focus on mock lies, produced by experimental subjects under laboratory settings. These studies result in the creation of balanced data sets, but their findings may not be generalized to deception encountered in real life.

In the following sections, the three methodological approaches mentioned above are discussed, both applied to some field and laboratory studies.

2.2 Physiologic variables based lie detection

2.2.1 Functional Magnetic Resonance Imaging - fMRI

One of the most innovative approaches to deception deception relies on modern techniques of neuro-imaging. In particular, the more and more widespread use in hospitals and universities of functional Magnetic Resonance Imaging, commonly termed 'fMRI', is deeply affecting cognitive neuroscience (Logothetis, 2008). This is also true regarding the field of deception detection. The technique of fMRI, that measures the changes in blood flow and

oxygen consumption of the brain areas, was employed to visualize the neural activity while the experimental subjects were carrying out tasks in which they had to lie. Langleben et al. (2002) claim that 'there is a neurophysiological difference between deception and truth at the brain activation level that can be detected with fMRI', and Ganis et al. (2003) seem to be confident about the fact that 'at least in part, distinct neural networks support different types of deception'. The opinion of Davatzikos et al. (2005) is the same. It is also remarkable that, in the United States, at least two laboratories[1] already provide a service of deception detection based on fMRI technology.

In spite of the enthusiasm of these authors, other researchers are cautious, if not skeptical, about the possibility of mapping cerebral areas involved in the production of deceptive statements (Merikangas, 2008; Simpson, 2008). Moriarty (2009) expresses in a colorful way her theoretical doubts about the connection between neuro-imaging data and deception: 'During the Salem witchcraft trials, Cotton Mather consulted leading treatises on the scientific proof of witchcraft-as science was understood in the Seventeenth Century. In large part, Mather, who fancied himself a man of science, was not impressed with the use of ordeals and torture: "going to the Devil for help against the Devil," as he might have put it. Rather, he was most impressed with a scientific causation argument: If, after a suspected witch curses, there follows death, illness or affliction, there is a presumption of witchcraft. Thus, in the Bridget Bishop trial, evidence was introduced that after Bishop had quarreled with a particular family, the family's pig was taken with strange fits and began foaming at the mouth; these events were believed to be sure evidence that Bishop had bewitched the pig. This supposed relationship, which I have termed elsewhere "Bewitched Pig Syndrome," was considered solid, scientific evidence of witchcraft for more than a century. Today, we might be inclined to note the "*post hoc propter hoc*" fallacy - *"after which, because of which."*'. Moreover Moriarty (2009) points out that fMRI lie detection is a science 'in its infancy' and the recent studies in the field are still lacking of consistency and reproducibility. Spence (2008), who believes that the reliability of fMRI lie detection in 'real world' has still to be proven, notices that 'there is a great deal of variation between the findings described and, crucially, there is an absence of replication by investigators of their *own* findings'. According to Simpson (2008), who carried out a careful review of the recent literature related to this topic, 'the technique does not directly identify the neural signature of a lie. Functional MRI lie detection is based on the identification of patterns of cerebral blood flow that statistically correlate with the act of lying in a controlled experimental situation. The technique does not read minds and determine

[1] See http://www.noliemri.com and http://www.cephoscorp.com

whether a person's memory in fact contains something other than what he or she says it does'. Nevertheless, his opinion is that 'with ongoing research, and likely improvements in accuracy in the laboratory setting, it does not seem unreasonable to predict that fMRI lie detection will gain wider acceptance and, at a minimum, replace the polygraph for certain applications. What seems far less likely is the science-fiction scenario in which a criminal defendant is convicted solely on the basis of a pattern of neuronal activation when under questioning'.

Regarding the replacement of the polygraph with the fMRI, Vrij (2008) does not agree with Simpson, considering that 'fMRI tests are expensive, time consuming, and uncomfortable for examinees' and therefore concluding that fMRI would be 'worthy of introducing as a lie detection tool in real-life situations if they are more accurate than the alternative techniques available to date. So far, research has not yet shown that the fMRI technique does produce more accurate results than traditional polygraph testing'.

2.2.2 The polygraph

It is revealing that both Simpson and Vrij see some analogies between polygraph and fMRI. In fact the same kind of debate regarding the reliability of the results of fMRI concerns the polygraph as well, in spite of the fact that this is a tool for lie detection which has been under evaluation for more than eighty years (Saxe and Ben-Shakhar, 1999). The polygraph, like fMRI, is a device which records some bodily activities: in this case, Electro-Dermal Activity (EDA), blood pressure, and respiration. Similar issues are raised regarding their use. About the polygraph, as well, it can be said that it does not 'read the minds', but simply measures physiological variables, which are assumed to be associated to deception. This association can be of two different kinds, which lead to two different strategies in the use of polygraph: the 'concern approach' and the 'orienting reflex approach' (Vrij, 2008).

The assumption of the concern approach is that the polygraph can be employed to detect signs of stress which are supposed to be related to the production of deceptive statements. Such studies are mainly carried out in criminal investigation settings, which make use of the Comparison Question Test (CQT) (Backster, 1962, 1963; Raskin, 1979, 1982, 1986; Reid, 1947), an interview protocol in five phases aimed to check the bodily reactions of the subjects to crime-related and different kind of control questions. According to relatively old (Brett et al., 1986) and more recent studies (Stern, 2003), in this setting the accuracy of the polygraph in detecting deception can vary from 50% to 95% (Simpson, 2008). Vrij (2008) tried to provide a comprehensive collection of reviews of this kind of studies. As usual in deception detection literature, he found laboratory and field studies. In laboratory studies the stakes are lower than real-life scenarios and their usefulness in

estimating the accuracy of the technique in applied settings is therefore doubtful. Field studies are more realistic, but in these cases confessions themselves are usually accepted as evidence of the ground truth, and this point could be questionable. Also questionable can be to trust in polygraph tests which establish the innocence of the suspect. 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 CIA. During these nine years, he successfully passed two polygraph tests. However, after having selected the studies accomplishing acceptable quality standards, Vrij (2008) found that CQT laboratory examinations reach an accuracy from 74% to 82% in classifying liars, although innocent suspects were correctly classified with an accuracy rate from 60% to 66% and between 12% and 16% of them were believed to lie. The results of field studies showed that between 83% and 89% of the liars were correctly classified. Unfortunately only between 53% and 75% of innocent examinees were correctly identified and - more worrying - by a rate from 12% to 47% of them were incorrectly classified, suggesting that the CQT protocol is vulnerable to false-positive errors (Vrij, 2008).

The orienting reflex-based polygraph tests rely on the assumption that 'an orienting response [*...omissis...*] occurs when someone is confronted with a personally significant stimulus' Vrij (2008). Orienting reflexes can be detected by polygraph through the Guilty Knowledge Test (GKT), developed by Lykken (1959, 1960, 1988, 1991, 1998). The strength and the weakness at the same time of this approach lies in presenting to the subjects stimuli (usually images) they should be familiar with. On the one hand, in many settings to prepare this kind of stimuli is not possible, or it is not possible to be sure that the stimuli can be genuinely significant for the suspects. On the other hand, when it is possible to prepare the right stimuli, it is highly probable that the individuals will present orienting reflexes, although some concerns remain regarding the possibility of lack of memory in the subjects and of getting similar responses from stimuli which are really known or which have only some similarity with something known. However, the reviews considered by Vrij (2008) show that field studies achieve from 76% to 88% of accuracy in identifying liars. Above all, only between 1% and 6% of innocent subjects were misclassified. Unfortunately, only two GKT field studies were found. In one of them, guilty suspects were correctly classified with a percentage of 76%, in the other one only 42% of liars were identified, suggesting a weakness of GKT regarding false-negative errors. By contrast these two studies seem confirm that GKT is hardly prone to false-positive errors, since innocent subjects were misclassified with a percentage between 2% and 6%.

2.2.3 Other technologies

In addition to these best known methods, other techniques are used in deception detection. Making no claim of completeness, we list a few other well-known techniques.

Voice Stress Analysis - VSA

Similarly to the polygraph tests relying on concern approach, the assumption of Voice Stress Analysis (VSA) is that telling lies is more stressful than telling the truth (Gamer et al., 2006). In order to detect signs of stress in the voice, the speech is recorded and its characteristics of intensity, frequency, pitch, harmonics, and so on are considered.[2] This method allows to collect data not invasively and, if needed, covertly. However according to Vrij (2008) voice stress analyses 'may detect truths and lies inaccurately'. Moreover, when the data are gathered covertly, it is not possible to apply the CQT protocol.

Thermal Imaging

This technique, developed by Pavlidis et al. (2002), relies on the same assumptions as the concern approach. In this case, the idea is that subjects, when they lie, present an instantaneous warming of the skin around the eyes, which is a sign of an increased blood flow. Such response is detected by heat detecting high-definition cameras. The method acquired notoriety also because Pavlidis *et al.* published their study in the famous scientific journal *Nature* just few months after the September 11 attacks. As in the case of VSA, the advantage of this technique is that it can be employed covertly. Nevertheless, its achieved accuracy has been questioned, as claimed by Vrij (2008): 'Unfortunately, thermal imaging is not the equivalent of Pinocchio's growing nose'.

Event-Related Potentials - ERP

In line with the orienting reflex approach, event-related brain waves can be recorded through electroencephalograms (EEGs) (Rosenfeld, 2002). Among these waves, P300s are a response to personally significant stimuli, which take the name from the fact that their peak appears typically between 300 and 1000 milliseconds after the stimulus. As such, they can be used as clue of deception. The only difference between ERP and GKT polygraph tests consists in the physiological variables taken into consideration: P300 waves versus EDA, blood pressure and respiration. Vrij (2008) reported the results of several studies applying this technique, showing that the performance of ERP tests is similar to that of GKT polygraph tests: an average of 82.29% of liars correctly classified and of 8.75% of truth

[2] http://www.polygraph.org/voicestress.htm

tellers misclassified. However the difficulties in finding the opportune stimuli addressed regarding GKT polygraph tests are present in ERP tests as well.

Transcranial Direct Current Stimulation - tDCS

Lastly, in Italy a study of Priori et al. (2008) made use of transcranial Direct Current Stimulation (tDCS) in the dorsolateral prefrontal cortex (DLPFC), demonstrating that the stimulation affected the reaction times in tasks involving the production of truthful and deceptive responses, with significant differences between the two conditions. As far as we know, this is the only approach which studies deception through the direct manipulation of brain functions.

2.3 Lie detection based on Non-Verbal Behavior

2.3.1 The study of Zuckerman, De Paulo and Rosenthal

Zuckerman et al. (1981) formalized the main theoretical perspectives followed in deception detection through non-verbal behavior analyses. According to these authors, deception should affect:

Emotional reactions. According to Ekman (1989, 2001), three emotions are usually associated to deception: guilt, fear, and delight. All of them can be different in different subjects and can affect the liars' behavior.

Cognitive effort. Liars have to accomplish several cognitively demanding tasks. First, they have to formulate narratives different from the truth they know. Then liars have to monitor that their statements are plausible, and they have to pay attention to not contradict themselves (Vrij, 2008).

Attempted behavioral control. Liars have to monitor their verbal and non-verbal behavior, in order to result convincing. This task could be difficult, since some bodily reactions, such as the tone of voice (Ekman, 1981), are almost beyond the voluntary control (Ekman, 2001).

Arousal. From a practical point of view, arousal and emotional reactions are the same collection of physical phenomena, thence they are not distinguished in research activities (Vrij, 2008).

In high-stakes settings, subjects are more motivated than in low-stakes situations, and their stronger motivation should result in an increased cognitive and emotional load (Vrij,

2008). Thence evidence of this greater involvement should be found in the non-verbal behavior. This is the main reason why the reliability of laboratory studies - where the stakes are necessarily low - is doubtful.

2.3.2 The studies of Ekman

The research activity of Ekman relies on the idea that strong emotions can activate facial muscles almost automatically, and thence to observe micro-expressions could be revealing about deception (Ekman, 2001). For example, if a subject tries to deny to be angry, he will have to suppress typical signs of anger, such as narrowing of the lips, lowering of the eyebrows and so on. But this task is difficult, since emotions can arise suddenly. According to Ekman (2001), subjects can suppress their expressions within 1/25 of second, but this lapse of time is enough for a trained observer to detect such expressions. Moreover several authors (Ekman et al., 1985; Ekman and O'Sullivan, 2006; Hess and Kleck, 1990; Hill and Craig, 2002) found that spontaneous and deliberate expressions are different in latency time, overall duration, duration of peak intensity and onset and offset time (the time from the start of the expression to its peak and from the peak to its disappearance, respectively).

2.3.3 Interpersonal Deception Theory - IDT

Another interesting perspective on deception was proposed by Buller and Burgoon (1996). Interpersonal Deception Theory substantially follows the line of Zuckerman et al. (1981), adding to this formalization its core idea: deception is a form of interaction. As such, both the liar and the other participant(s) to the dialogue influence each other (Burgoon et al., 1996). The influence can be direct or indirect (Burgoon et al., 1999). Direct influence regards phenomena such as matching and synchrony, which may take place during the interaction. This effect, also known as 'chamelon effect' (Chartrand and Bargh, 1999), can be observed after just few minutes of interaction, even between strangers. Indirect effects are instead related to feedback that the liar receives, mainly about his credibility, from the interlocutor and they represent the liar's attempt to modulate his behavior in order to be persuasive.

Although the Interpersonal Deception Theory was formulated with reference to non-verbal behavior, this turns out to be interesting also for verbal behavior based studies. In fact the dialogical interaction between subjects can affect also their verbal behavior (Ireland et al., 2011; Niederhoffer and Pennebaker, 2002).

2.3.4 Non-Verbal clues of deception in literature

Studies focused on non-verbal clues of deception usually rely on the activity of trained raters who watch videos in which liars and true tellers interact, with the aim of analyzing some form of non-verbal behavior. Coding system are adopted in order to detect frequency, duration and intensity of several non-verbal cues and to compare the results for liars and true tellers. As previously discussed, also in this branch of deception detection the difficult choice between laboratory and field studies is present. Field studies are appealing as they are realistic and yet to find good quality videos, containing truthful and deceptive comparable data, and to establish the ground truth is often difficult or impossible. Laboratory studies are not affected by these problems, but the psychological conditions of the subjects are very different from those in natural environment.

Vrij (2008) summarizes a set of 132 studies focused on non-verbal cues to deception. Among the cues taken into consideration in these studies, he distinguishes vocal from visual ones. Table 2.1 reports the list of cues, as formulated by Vrij (2008). In order to summarize the results, Vrij also considered the findings of the quantitative meta-analysis carried out by De Paulo et al. (2003), regarding the same cues of deception. The effect sizes which were found, were evaluated according to the criteria suggested by Cohen (1977). Out of the seventeen cues considered, only three were found significant:

- **Pitch**: liars use a higher pitch of voice than truth tellers, but the effect is small (furthermore, the difference between liars and true tellers usually is only few Hertz, and needs professional devices to be detected);
- **Illustrators**: liars show fewer illustrators than true tellers, with a 'small' effect size;
- **Hand and finger movements**: liars move hands and fingers less than true tellers, with a 'small/medium' effect size. However, Vrij et al. (1997) analyzed this variable on 181 subjects, finding that '64% of them showed a decrease in hand/finger movements during deception, whereas 36% showed an increase of these movements during deception'.

The overall findings of the studies, Vrij claims, 'show an erratic pattern and indicate that many conflicting results have been found'. For example, in some studies speech hesitations are more frequent in liars than in true tellers, in others the opposite is found. The pauses in the speech seem to be longer in liars than in true tellers, but not necessarily more frequent. Sporer and Schwandt (2006b), in their meta-analysis of paraverbal cues, found that liars present longer latencies than truth tellers, but also in this case the effect size was small. Gaze behavior does not seem to be related to deception and this is remarkable,

Table 2.1. Vrij's non-verbal cues for deception detection.

- Vocal cues:
  1. **Speech hesitations**: use of speech fillers e.g., 'ah', 'um', 'er', 'uh' and 'hmmm';
  2. **Speech errors**: grammatical errors, word and/or sentence repetition, false starts, sentence change, sentence incompletions, slips of the tongue, etc.;
  3. **Pitch of voice**: changes in pitch of voice, such as rise in pitch or fall in pitch;
  4. **Speech rate**: number of spoken words in a certain period of time;
  5. **Latency period**: period of silence between question and answer;
  6. **Pause durations**: length of silent periods during speech;
  7. **Frequency of pauses**: frequency of silent periods during speech;
- Visual cues:
  1. **Gaze**: looking into the face of the conversation partner;
  2. **Smile**: smiling and laughing;
  3. **Self-adaptors**: scratching the head, wrists, etc.;
  4. **Illustrators**: hand and arm movements designed to modify and/or supplement what is being said verbally;
  5. **Hand and finger movements**: movements of hands or fingers without moving the arms;
  6. **Leg and foot movements**: movements of legs and feet;
  7. **Trunk movements**: movements of the trunk;
  8. **Head movements**: head nods and head shakes;
  9. **Shifting position**: movements made to change seating position;
  10. **Blinking**: blinking of the eyes.

as popular opinion - even among experts in lie detection - is that liars tend to look away from their interlocutor. However gaze behavior is easy to control and people are aware of its importance for communication, thence it cannot be considered an effective marker for deception (Vrij, 2008).

Together with the discussed cues, De Paulo et al. (2003) considered around one hundred behaviors, of which 21 were significant (including the three already discussed). It turned out that, compared to true tellers, liars tend to have a greater pupil dilation (Wang et al., 2010), and they 'appear tenser, have a more tense voice, have their chin more raised, press

their lips more, and have less pleasant looking faces. They also sound more ambivalent, less certain and less involved, and make more word and sentence repetitions' (Vrij, 2008). However, no cue shows a significance greater than 'small/medium', and again Vrij concludes that 'a cue akin to Pinocchio's growing nose does not exist' (Vrij, 2008).

One of the reasons inducing him to draw this conclusion is that, even though the findings of De Paulo et al. (2003) support the hypothesis of Zuckerman et al. (1981) that subjects experience emotional reactions when deceiving, the emotions felt are not necessarily associated with the act of deceiving in itself. In other word, it is not just liars who feel strong emotions, and even when clues of emotional reactions can be found, the cause for these emotions could be unclear. Thence, coherently with the cited claims of Frank et al. (2008), in the end also De Paulo et al. (2003) state that 'behaviors that are indicative of deception can be indicative of other states and processes as well'.

However it seems that clusters of cues could be effective in deception detection. Vrij (2008) claims that with a combination of four different variables (illustrators, hesitations, latency period, and hand/finger movements) he was able to classify correctly 84.6% of liars and 70.6% of true tellers (Vrij et al., 2000). Similar results were found by Frank and Ekman (1997), who achieved an accuracy up to 80% in detecting deception through micro-expressions observation, but reached a performance even better taking into account micro-facial expressions and tone of voice. In a similar vein, recently Jensen et al. (2010) focused on cues coming from audio, video and textual data, with the aim of building a paradigm for deception detection via a multi-layered model. These authors take into consideration directly objective indicators - they call 'distal cues' - rather than human observations - addressed as 'proximal cues' - that they found to not lead to the best performance in detecting deception. They reached a classification accuracy of 73.3%, and claim: 'Deception indicators are subtle, dynamic, and transitory, and often elude a human's conscious awareness. The increased precision afforded by the distal cues may provide additional information that can be used to classify deception directly. This finding demonstrates the effective use of unobtrusive, automatically extracted features in deception detection' (Jensen et al., 2010).

A last idea comes from the field of non-verbal behavior analysis, which can be particularly interesting for the perspective of this thesis. Although field studies are rare, Vrij (2008) argues that liars' behavior in high-stakes settings, which is characterized by long pauses, word repetitions and decrease in hand/finger movements, suggests an increased cognitive load for the subjects. This confirms the theoretical frame proposed by Zuckerman et al. (1981) and seems promising for the field of research which deals with the cognitively most demanding behavior for liars: verbal behavior.

2.4 Lie detection based on Verbal Behavior

Verbal behavior analysis in deception detection is characterized by two approaches mutually interdependent. The first one refers mostly to semantic analyses, the second one mainly to stylistic analyses.

The focus of semantic analyses is on the content of the communication, and in particular on the internal and external logic of the narrative (Smirnov, 1988), that is on the identification of contradictions or discrepancies between statements, or between statements and objective elements, respectively. Historically this was the first approach to deception detection, and very ancient examples can be found, such as the episode in which the prophet Daniel unmasks the deceptive accusations of the two old judges against Susan, inducing them to issues statements contradicting each other (Daniel 13:1-59 Nova Vulgata). In the Twentieth century, in order to improve the analyses of interviews and testimonies, protocols were developed relying on semantic analyses, but affected by the modern cognitive theories of memory and deception, such as Statement Validity Assessment (SVA), Reality Monitoring (RM) and Scientific Content Analysis (SCA). These methods are discussed in the Subsections 2.4.1, 2.4.1 and 2.4.3.

By contrast, stylistic analyses rely, more or less explicitly, on the same assumption formulated by Zuckerman et al. (1981) for non-verbal behavior analyses. The idea is that emotional reactions, cognitive effort and attempted behavioral control related to deception can affect not only non-verbal behavior, but also linguistic production. Thence clues of their presence should be found in verbal behavior as well. In this perspective, the approach to the analysis of verbal cues for deception identification that is becoming more and more dominant in recent years is stylometry, which will be discussed in Subsection 2.4.4.

2.4.1 Statement Validity Assessment - SVA

In forensic practice, Statement Validity Assessment (SVA) is probably the best known and most employed verbal veracity assessment tool, accepted as evidence in Courts in North American, Austria, Germany, Sweden, Switzerland, and the Netherlands (Vrij, 2008). Initially developed by Trankell (1963), Undeutsch (1967) and Arntzen (1970), this tool was conceived to evaluate the reliability of the testimonies of children in criminal proceedings for sexual abuses. In the end, SVA assumed its current form thanks to the work of Köhnken and Steller (1988). The basic assumption of Statement Validity Assessment is the so-called Undeutsch hypothesis (Steller, 1989), according to which the cognitive elaboration of a memory differs from the elaboration of an imaginative construction, and this difference should be traceable in the features of the issued narrative. Although it led to a

different method of analysis, it can be seen that this assumption is coherent with the latter formulation of Zuckerman et al. (1981).

SVA consists of four phases:

- A preliminary analysis of the case;
- A semi-structured interview aimed to get the statements of the subject;
- The Criteria-Based Content Analysis - CBCA, which is the core of SVA;
- An evaluation of CBCA through the Validity Checklist.

CBCA, in turn, consists of 19 criteria, shown in Table 2.2. They are marked as present or absent in the text by trained evaluators. Then the outcome of CBCA is evaluated through the Validity Checklist. This addresses issues related to possible intervening variables that can affect the validity of the result, such as psychological characteristics and motivation of the subject and characteristics of the interviewer and of the interview itself.

Field and laboratory studies have been carried out in order to evaluate the reliability of SVA. This task turned out to be very difficult. The limit of the field studies is that often convictions and confessions are used to establish the ground truth, but frequently these convictions and confessions are influenced by the results of SVA itself, creating a circular linkage between cause and effect (Vrij, 2008). By contrast, laboratory studies are lacking of ecological validity, so that Undeutsch (1984) claimed that they are not particularly useful in testing the accuracy of such tool. Nevertheless, Vrij (2008) finds that one of the most reliable field studies shows 'several, albeit small, differences between truthful and fabricated statements (Lamb et al., 1997), and all of these differences were predicted by the Undeutsch hypothesis: the criteria were more often present in truthful than in fabricated statements'. By contrast, laboratory studies suggest that CBCA can identify truth and lies with a degree of accuracy of around 70% (Vrij, 2008).

2.4.2 Reality Monitoring - RM

Unlike SVA, Reality Monitoring is not widely employed in forensic practice, maybe because it does not address directly deception. However this tool turned out to be interesting for researchers, due to its basic assumptions. Reality Monitoring, developed by Johnson and Raye (1981, 1998) (and also in Johnson et al. (1993)) relies on an idea very similar to the Undeutsch hypothesis (Undeutsch, 1984): cognitive processes related to perceived and imagined events are different. Therefore the authors expect that perceived events originate memories rich in sensory information, spatial and temporal contextual information and

Table 2.2. CBCA's criteria.

- General characteristics:
  1. Logical structure;
  2. Unstructured production;
  3. Quantity of details;
- Specific contents:
  4. Contextual embedding;
  5. Descriptions of interactions;
  6. Reproduction of conversation;
  7. Unexpected complications during the incident;
  8. Unusual details;
  9. Superfluous details;
  10. Accurately reported details misunderstood;
  11. Related external associations;
  12. Accounts of subjective mental state;
  13. Attribution of perpetrator's mental state;
- Motivation-related contents:
  14. Spontaneous corrections;
  15. Admitting lack of memory;
  16. Raising doubts about one's own testimony;
  17. Self-deprecation;
  18. Pardoning the perpetrator;
- Offence-specific elements:
  19. Details characteristic of the offence.

affective information. By contrast, imagined events should create memories which contain more cognitive operations and less concrete expressions.

Similarly to SVA, the RM protocol dictates that the subjects are interviewed and RM experts check for the presence or absence of RM criteria in the subjects' statements. Table 2.3 lists the Reality Monitoring criteria. However Vrij (2008), who summarizes the findings

Table 2.3. Reality Monitoring criteria.

1. Clarity;
2. Perceptual information;
3. Spatial information;
4. Temporal information;
5. Affect;
6. Reconstructability of the story;
7. Realism;
8. Cognitive operations.

of several studies about Reality Monitoring, reports lack of standardized procedures for the evaluation of such criteria. Therefore these studies could differ in what they actually measure.

An example of this lack of homogeneity is interesting from the point of view of this thesis because it involves a tool employed for the present data anlysis: the Linguistic Inquiry and Word Count - LIWC, the lexicon created by Pennebaker et al. (2001) and described in Subsection 2.4.5. Bond and Lee (2005), in order to verify the presence of the RM criteria, annotated the transcripts of their interviews both manually and using LIWC. The results were unfavourable to LIWC, since in the first case the authors found differences between liars and true tellers, in the second one they did not. The opinion of Vrij (2008) is that 'the problem with using automatic coding is that computer word counting systems ignore context, whereas the RM tool, as well as CBCA, require that the context is taken into account'. Although LIWC is an useful and largely employed resource in detecting deception, this finding is relevant because it reminds that the possibility of improving the performance in deception detection depends on the definition of more and more effective and reliable cues of deception, to be measured possibly in a standardized and automatic way.

As far as the performance of RM tests is concerned, Vrij (2008) found that in literature Reality Monitoring reaches an average accuracy in detecting truth at 71.7% and in detecting lies at 66.1%, with a total average at 68.8%: rates 'higher than could be expected by just flipping a coin'.

2.4.3 Scientific Content Analysis - SCA

A last method addressed by Vrij (2008) is Scientific Content Analysis (SCA), developed by Sapir (2000). To evaluate its effectiveness in detecting deception is difficult, because few studies tested this method. However Vrij (2008) claims that it is used by the Police Forces of several Countries, such as Australia, Canada United States, Belgium, Israel, Mexico, Singapore, South Africa, the Netherlands, and United Kingdom. We mention Scientific Content Analysis because of an interesting analogy with the results of this thesis. In SCA the best criterion as predictor of deceptiveness turned out to be 'Denials of allegations'. As discussed in Subsection 4.2.4, this is also found in DeCour, since in criminal investigations it is frequent that suspects lie denying facts which are charged on their responsibility. This suggests that some characteristics of deceptive language could be highly dependent on the context whereby the communication takes place, and this should be taken into consideration as far as the domain dependence of the scientific findings is concerned.

2.4.4 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 the field became established only in the Nineteenth century with the introduction by De Morgan of quantitative measures in stylistic studies (Lord, 1958). The (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).

Vrij (2008) lists some typical stylometric variables, which are not taken into consideration by the tools previously discussed but are object of other analyses in literature. As in the case of non-verbal behavior, a diagnostic cue like the noose of Pinocchio does not seem to exist. Moreover some cues are considered in some studies as predictors of deception, and in some other as predictors of truthfulness. This is the case, for example, of the length of the statements, although most authors found that liars produce shorter answers than

true tellers, especially in high-stakes scenarios (Sporer and Schwandt, 2006a; Vrij, 2008). This trend is also different from what we found in DeCour, as shown in Section 3.5 and in particular in Table 3.4. Equally ambiguous is the pattern of lexical diversity, that is the ratio between the number of different words in a statement and the total number of words used in the same statement. However Vrij (2008) believes that in high-stakes settings, whereby subjects are strongly motivated to deceive, liars tend to repeat the information they provide and to use a more stereotypical language. This is coherent with the findings in DeCour, as discussed in Section 7.1.

In spite of these unclear patterns, clusters of features could be useful to detect deception (Bond and Lee, 2005; Newman et al., 2003; Zhou et al., 2004). 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.

Several recent studies try to detect deception making use of cluster of features, mostly automatically collected from the analyzed texts. Some of them are discussed in the next Subsections.

2.4.5 The study of Newman, Pennebaker, Berry and Richards

Newman et al. (2003) were arguably the first authors to show that stylometric techniques could be effectively applied to detect deception. They collected a corpus of sincere and deceptive texts through five different laboratory studies. In three of them, the subjects had

both to describe their true opinion about abortion, and also try to support the opposite point of view. The opinions were videotaped, typed and handwritten, respectively. The fourth study was videotaped, and the subjects had to express true and false feelings about people they liked or disliked. Finally, the fifth study, which was also videotaped, consisted of a mock crime, in which the subjects were accused by an experimenter, rightly or not, of a small theft, and they had to reject any responsibility. As a result, Newman *et al.* obtained ten groups of texts, five sincere and five deceptive. These data were then analyzed using a lexical resource: the Linguistic Inquiry and Word Count (LIWC).

Linguistic Inquiry and Word Count - LIWC. This is perhaps the best-known lexical resource for deception detection, developed by Pennebaker et al. (2001). LIWC is a validated tool categorizing words under a number of dimensions. In particular, it is a lexicon, whose English dictionary is constituted of around 4500 words or roots of words, whereby each term is associated with an appropriate set of syntactical, semantical and/or psychological categories, such as emotional words, cognitive words, self references, different kind of pronouns, and so on. When a text is analysed with LIWC, the tokens of the text are compared with the LIWC dictionary. Every time a word present in the dictionary is found, the count of the corresponding dimensions grows. For example, when in one text the LIWC recognizes a word belonging to a category, such as 'I' or 'you' for the category 'pronoun', or 'no', 'neither', 'never' for the category 'negation' and so on, the count of that category grows. The output is a profile of the text which relies on the rate of incidence of the different dimensions in the text itself. LIWC also includes different dictionaries for several languages, amongst which Italian (Alparone et al., 2004). Therefore it was possible to apply LIWC to Italian deceptive texts of DeCour, as discussed in the following Chapters.

The texts collected by Newman et al. (2003) were preliminarily analyzed using the LIWC. Of the 72 linguistic dimensions considered by the program, the authors selected the 29 variables considered more promising to detect deception. In particular, they excluded the categories that could reflect the content of the texts (such as 'leisure', 'money', 'religion' and so on), those used less frequently in the texts, and those specific of one form of communication (for example the 'nonfluencies', that are specific of spoken language). At the end, they considered the 29 variables listed in Table 2.4.

For the analyses, first, the values of the 29 variables were standardized by conversion of the percentages outputted by the LIWC to z scores. Then, the authors carried out the analyses described as follows: 'We first performed a forward-entry logistic regression, predicting deception based on usage of the 29 LIWC categories in four of the five studies. This

Table 2.4. LIWC dimension employed in the experiments of Newman et al. (2003)

- Standard linguistic dimensions:
  1. Word Count;
  2. % words captured by the dictionary;
  3. % words longer than six letters;
  4. Total pronouns;
  5. First-person singular;
  6. Total first person;
  7. Total third person;
  8. Negations;
  9. Articles;
  10. Prepositions;
- Psychological processes:
  11. Affective or emotional processes;
  12. Positive emotions;
  13. Negative emotions;
  14. Cognitive processes;
  15. Causation;
  16. Insight;
  17. Discrepancy;
  18. Tentative;
  19. Certainty;
  20. Sensory and perceptual processes;
  21. Social processes;
- Relativity:
  22. Space;
  23. Inclusive;
  24. Exclusive;
  25. Motion verbs;
  26. Time;
  27. Past tense verb;
  28. Present tense verb;
  29. Future tense verb.

logistic regression produced a set of beta weights predicting deception. Second, these beta weights were multiplied by the corresponding LIWC categories in the remaining study and added together to create a prediction equation. These equations formed our operational definition of linguistic profiles. Finally, a second logistic regression was performed, using this equation to predict deception in the remaining study. These three steps were repeated for each of the five studies. In all analyses, deception was coded as a dichotomous variable

with truth-telling coded as 1 and lying coded as 0'.

Whereas chance performance was 50% of correct classifications, the authors reached an accuracy of about 60% (with a peak of 67%) in three of the five studies. In the remaining two studies, the performances were not better than chance. To evaluate simultaneously the five studies, from the 29 LIWC categories, the following five were selected:

1. First-person singular pronouns;
2. Third person pronouns;
3. Negative emotions words;
4. Exclusive words;
5. Motion verbs.

They were the variables that were significant predictors in at least two studies, and also in this case the accuracy of the previsions was about 60%.

The results of Newman et al. (2003) suggested that 'deceptive communication was characterized by the use of fewer first-person singular pronouns (e.g., *I*, *me*, *my*), fewer third-person pronouns (e.g., *he*, *she*, *they*), more negative emotion words (e.g., *hate*, *anger*, *enemy*), fewer exclusive words (e.g., *but*, *except*, *without*), and more motion verbs (e.g., *walk*, *move*, *go*)'. The finding about the use of less first-person pronouns is interpreted as an attempt by the liar to 'dissociate' himself from the lie. The presence of negative emotions should reflect the guilt felt by the liar. The use of less 'exclusive' words and more 'motion' verbs should be a sign of lower cognitive complexity, due to the cognitive load of telling lies. The lower rate of third-person pronouns was not expected by Newman et al. (2003) and interpreted as the result of the content of the examined texts: in particular, the topic of abortion would induce to use of third persons.

In the end of the paper, the authors address two limitations of their study. They deserve to be mentioned, because this thesis has been conceived exactly to overcome these limitations. First, their model is limited to English. The authors are aware that in other languages, such as Romance languages, which can omit the pronouns, deceptive communication could show different patterns, especially regarding the use of the pronouns. Second, the emotional involvement of their subjects was low, but motivation can affect significantly deceptive communication. The present dissertation focuses just on Italian language - a Romance language - and on transcripts of hearings in Courts, that is on data produced in an high-stakes setting.

2.4.6 The study of Strapparava and Mihalcea

In a similar vein, Strapparava and Mihalcea (2009) collected a corpus of truthful and deceptive statements making use of the Amazon Mechanical Turk service.[3] The subjects were asked to prepare two brief speeches, the one expressing their true opinion on a topic, and the other one expressing false opinion on the same topic. The selected topic were 'abortion', 'death penalty' and 'best friend'. For each of them, the authors collected 100 truthful and 100 deceptive statements, with an average of 85 words per statement. In this paper the computational approach to deception detection was different from the previously discussed paper of Newman et al. (2003). The authors employed two classifiers: Naïve Bayes - NB and Support Vector Machines - SVM. Furthermore these algorithms were fed with simple surface features. LIWC was employed, but only for an analysis *ex post*, in order to get some insight about deceptive language. Although their models relied only on surface features, the accuracy in detecting truthful and deceptive texts was around 70%: even better than the one of Newman et al. (2003). This suggested that simple surface features can be successfully employed in deception detection.

The present dissertation followed the same path. We used surface features as well, in order to verify their effectiveness on the Italian texts of DeCour. We also made use of Support Vector Machine classifiers, which empirically turned out to be the most effective to classify or data set.

[3] https://www.mturk.com/mturk/welcome

2.4.7 The study of Fitzpatrick and Bachenko

While Newman et al. (2003) and Strapparava and Mihalcea (2009) focused on laboratory data whereby deception clues were automatically detected through LIWC, Fitzpatrick and Bachenko (2009) tried to collect a corpus of deceptive statements coming from real life cases in which deception indicators were manually identified. Like DeCour in Italian, as far as we know this is currently the only corpus of this kind in English and it relies on open source data such as 'Court TV and other web sources as well as published works and local police documents', whose the ground truth was known. Unfortunately, a direct parallelism between this study and the present dissertation was not possible, as the clues of deception employed by Fitzpatrick and Bachenko (2009) were not directly applicable to DeCour. In fact the indicators of Fitzpatrick and Bachenko (2009) were selected according to the English literature about deceptive language, and their presence into the texts was marked by hand. These indicators belonged to three categories:

- Lack of commitment to a statement, that is 'devices used to avoid making a direct statement of fact';
- Use negative expressions;
- Inconsistencies between verb and noun forms.

According to the widespread idea that a single clue of deception is not 'sufficient to determine whether the language is deceptive or truthful', Fitzpatrick and Bachenko (2009) identified 'areas of a narrative that contain a clustering of deception indicators'. Using this method, the authors were able to distinguish truthful from deceptive 'areas of narrative' with an accuracy of 74.9%, suggesting that deception clues which were reliable in laboratory studies, can be effective when applied to real life cases as well. This dissertation is meant to verify the same point regarding Italian field cases.

2.5 Summary

Several approaches have been tried in order to detect deception. None of them found a single highly reliable predictor of deception. Nevertheless several studies suggested that the performance can be improved when clusters of cues are taken into consideration.

However, to define and to identify these cues is often not easy. Systems conceived to detect physiological variables require the availability of the due equipment and of the subjects themselves. Furthermore, it is not simple to standardize the interpretation of their

results. In studies relying on non-verbal behavior analyses, frequently the problem is double: not only the interpretation of the results, but also the detection of the cues is difficult to be standardized, as their identification depends on human evaluations. These difficulties are shared also by several studies whereby verbal behavior is evaluated, especially those in which the main focus is on semantic rather than stylistic analyses.

In this scenario, stylometry offers some remarkable advantages for deception detection. The data collection is as simplest as possible, since only texts are required. Feature selection can be largely carried out in an objective and automatic way. Data analysis is equally objective. Last but not least, the performance in detecting deception is not worse, and possibly better than the performance of other methods (the orienting reflex-based polygraph test reached accuracy levels clearly better than others, but the application of this method is quite limited, since it requires to present significant stimuli to the subjects).

To some extent, this outcome can result counterintuitive. In fact, as acutely observed by Vrij (2008), 'police manuals, and people in general, tend to neglect paying attention to verbal cues to deception, because it is assumed that liars are able to control their speech well and therefore are unlikely to leak deception through their speech'. By contrast, Newman et al. (2003) claims: 'Although liars have some control over the content of their stories, their underlying state of mind may leak out through the style of language used to tell the story'. The same concept is expressed by Vrij (2008): 'it is incorrect to assume that liars always control their speech well. [*...omissis...*] although people will be aware of what they are conveying, they may be less aware of their exact wording. As a result, they may not notice minor changes in their speech when they start lying. People are unlikely to attempt to control their speech if they don't notice changes in their speech'. Stylometry looks exactly for these 'minor changes'. This is a precise application of the Morellian method, mentioned in Section 1.2.

In this theoretical framework, the present dissertation should represent a novelty for two reasons. First, this is the first study of such kind carried out on the Italian language, allowing cross-lingual comparisons. Second, DeCour is a homogeneous collection of texts coming from a high-stakes setting, whereby the ground truth is known with a great degree of confidence. As abundantly discussed, the fulfillment of these requirements is rare in the field of deception detection. Nonetheless, DeCour allows to study deceptive language as it is produced by the subjects in conditions of high psychological involvement.

Part II

Corpus and methods

Chapter 3

Dataset

3.1 Data collection

In order to study deceptive language, we tried to build a corpus of texts:

- coming from a real life scenario;
- characterized by a strong psychological involvement of the speakers;
- collected in standard conditions;
- of which the truthfulness or truthlessness was known.

We found a way to fulfill these requirements in a legal context.

It happens in criminal proceedings that investigators interview, more or less formally, several subjects who consequently have the possibility to issue true or false statements. In most cases the reports in which the testimonies are collected do not bring back the words exactly pronounced by the subjects, but represent a synthesis of their declarations, carried out by the police officer who hears and records them. These reports are not a faithful mirror of the linguistic behavior of the subjects, therefore they are not useful from the point of view of the present work.

In some particularly serious cases, it is also possible that the interrogation in front of the public prosecutor is recorded and transcribed word by word. These interrogations could be useful, but they are relatively rare and also difficult to find because in the proceedings where they could be carried out, they are not always. Above all, even when theoretically possible, to find external and objective evidences of the truthfulness or deceptiveness of statements would be very difficult from a practical point of view. In fact, these evidences are usually dispersed in a lot of different and various investigative data, often in a huge amount.

Therefore the point was to find testimonies not only recorded word by word, but also of which the truthfulness or deceptiveness was easily verifiable.

3.2 Hearings

In Italian criminal proceedings there is a specific moment in which all the testimonies are imperatively recorded word by word: that is, the hearings that take place during the debate in front of the judge. Furthermore, in some proceedings the truthfulness or deceptiveness of the testimonies is easily verifiable. It is the case of criminal proceedings ex art. 368 and 372 of the Italian Criminal Code, which concern the crimes of 'calumny' and 'false testimony'[1]. While the concept of false testimony is intuitive, in Italian Criminal Code calumny is a particular kind of false testimony, consisting of the attempt to charge on someone else the responsibility of a crime that has been committed. The distinction makes sense because in the Italian legal system nobody can be forced to say some truth unfavorable to oneself. It means that to lie about a committed crime is not a crime, but it is so if trying to charge the responsibility to someone else. In order to collect this kind of data, contacts have been taken with Courts in four Italian towns, with the aim to be allowed to examine their dossiers and extract information with scientific purposes. Authorizations have been received to collect data, with the only restriction of using them in anonymous form, in respect to the privacy of the involved subjects.

The inquiries for calumny and false testimony usually originate from another previous proceeding, in which the defendant or a witness takes part in a hearing and issues statements that are found not reliable. In these cases, a new criminal proceeding arises, aimed to establish if the subject committed the crime of calumny or of false testimony. More rarely, the proceeding concerns statements which are not issued in a hearing, but in circumstances in which the words of the subjects are not recorded *verbatim*: typically, this is the case of the complaints lodged to the police. Nevertheless, in some cases the subjects, after having issued unreliable statements in front of police, come to the courtroom and confirm in a hearing the same testimony previously given. This is the less frequent situation because people who have lied during a hearing or in some other moment, have often the good sense of not repeating the crime twice and in front of the judge.

In fact, since these proceedings are aimed at verifying if the subject lied or not, they

[1]The art. 368 reads: "Chiunque, con denunzia, querela, richiesta o istanza, anche se anonima o sotto falso nome, diretta all'Autorità giudiziaria o ad altra Autorità che a quella abbia obbligo di riferirne, incolpa di un reato taluno che egli sa innocente, ovvero simula a carico di lui le tracce di un reato, è punito con la reclusione da due a sei anni." In brief, it punishes whoever tries to charge the responsibility of a crime on someone who he knows is innocent.
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." This article punishes someone who, in front of the Judicial Authority, says a falsehood or denies the truth, or does not reveal what he knows about the investigated facts.

imply the development of investigative activities which make highly improbable the possibility of errors or bias in the identification of deceptive statements. Also because in the presence of any reasonable doubt about guilt, the defendant is acquitted. Furthermore, in the Italian Criminal Code an essential part of the crime is the so called 'subjective element', which refers to the fact that not only the not truthfulness of the statements has to be ascertained, but also the precise intent of the subject of deceiving the Judicial Authority. In the end, the outcome of these proceedings is a judgment that summarizes the facts and, when the defendant is found guilty, points out in a certain, organic and exhaustive way the lies which he told.

In this way it has been feasible to create DeCour, a corpus of transcripts that contain the exact words pronounced by the subjects in the hearings, and about which it is possible to reliably know the truthfulness. In particular, in order to allow the study of deceptive language, DeCour is made of hearings where the subjects have effectively been found guilty. To be more precise, in few cases the defendants have been acquitted, but merely for procedural and legal reasons: in every hearing which constitutes the corpus, there are lies told by the defendant, and these lies are recognized and clearly pointed out in the judgment.

3.3 Preprocessing

3.3.1 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.

3.3.2 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.

3.3.3 Lemmatization and POS-tagging

The whole corpus was put in lower-case, and then lemmatized and pos-tagged using a version of TreeTagger[2] (Schmid, 1994) trained for Italian.

3.4 Annotation

3.4.1 Mark up format

Hearings in Court are events strongly ritualized, with rules determined by the Code of Criminal Procedure. It means that the development of every hearing is highly regular, almost like in an experimental design, giving the opportunity of collecting data in relatively homogeneous conditions, even when the actors differ. The protagonist of each hearing is the subject who gives the testimony. He answers the questions posed by three other figures, who cannot be absent from any hearing: the judge, the public prosecutor and the defendant's lawyer. Therefore, the considered testimonies have the form of a dialogue, in which at least four actors are present. It is possible that other actors intervene, for example more than one public prosecutor, or more than one defendant lawyer, or a lawyer for the victim of the crime, or a police officer: but these are less frequent cases.

Each text file that contains a testimony is transformed into XML format, with the aim of marking up actions and words of each participant. First, each XML has an **header** that contains some meta-information about the testimony, such as place and date of the event, and about the speaker, such as his age, sex, place of birth and if known - unfortunately, not often - his level of instruction. The hearing properly said begins with an **introduction**: a formal part of the report which gives act to the introduction of the subject in front of the judge and, if needed, of his availability to answer the questions (to issue statements is an option for the defendant, but is a duty for the witnesses). Then, the real dialogue begins and each intervention of the different actors, delimited by the interventions of others participants, is marked as **turn**. Each turn can be constituted of one or more **utterances**,

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

which are delimited by terminal punctuation marks: this is the atomic analysis unit of DeCour. Into each turn, besides, some **action** carried out by the speaker can be inserted, according to what is minuted in the report. In the end, a **conclusion** can be present or not, with some last ordering of the judge or some ending formulas.

3.4.2 Coding scheme

Each utterance issued by the speaker receives a label, which concerns his degree of truthfulness. This annotation is carried out by hand, on the basis of the information found in the Court's final judgment. Obviously, between the white of the truth and the black of the falsity, there are wide gradations of gray, and the judgment, that describes the facts and points out the lies of the defendant, cannot give reason for each statement issued in the courtroom. This is the reason why the process of labeling the utterances of DeCour followed a path which represented the research of a trade-off between opposite demands: the analytical representation of their degree of truthfulness and the achievement of a satisfying degree of agreement between different annotators, regarding this evaluation.

First step, DeCour has been labeled according to the following annotation scheme.

False. Utterances pointed out in the judgment as false, or of which the falsity is a logic consequence of some ascertained lie, are taken as false. For example, if the subject claims to have not been somewhere with someone, but actually he was, he also has to lie when he denies having known the same person.

Anyway, even though the judgment gives certain guidelines, it is not always easy to decide whether to assign to an utterance the label *false* or not. In fact, sometimes the meaning of linguistic behavior forces the focus on the function of the utterance, rather than on its literal sense. For example, from a theoretical point of view, questions do not represent any fact, therefore in strict sense they cannot be considered either true or false. But if a subject, pretending to not know a person, and asked "Do you know Mr. Rossi?", answers "Is Mr. Rossi the person in front of me?", the function of his answer/question is to generate in the judge a false representation of the reality, according to which he would not know Mr. Rossi. So, this utterance is considered a lie and labeled as false.

On the other end, some utterances which seem to be false, from a logical point of view are instead true. In one proceeding, for example, a subject claimed to be an electronic engineer, while he had had a simple high school education. In front of the lawyer who was saying "You said you are an engineer...", the witness completed the

sentence saying "electronic"[3]. Obviously, the fact that he was an engineer was false; but it was true that he had said that he was an engineer: then this answer was true, regarding the question posed.

True. Utterances which are found coherent with the reconstruction of the facts contained in the judgment, are considered as true. Also the utterances which concern something not considered in the sentence, because they are uninfluential in respect to the investigated facts, are generally considered true.

For example, if the public prosecutor asks "For how long have you been married to Mrs. Bianchi?", and the subject answers "For eight years", this answer is considered true, even though the judgment says nothing about that, because there are no logical reasons to lie on this detail.

Not reliable. Utterances which concern the investigated facts, but of which the truthfulness or deceptiveness is not established by the judgment, are considered not reliable. These utterances are related to some point about which the speaker could have some interest to lie, but the judgment does not provide the necessary information to evaluate them.

An interesting fact is that some judgments establish that the defendant was lying, when he had been claiming to not remember something. In these cases, the statements in which the subject says to not remember some specific event, are considered false. On the other hand, obviously, there are (many) proceedings in which it is not considered (or, at least, proved) as a lie the fact that the subject claims to not remember something. Also in this case, if the lack of memory is related to something that does not concern the topic of investigation, the utterance is considered sincere and labeled as true; otherwise, if it is related to the object of investigation, and the subject lying could defend his own interests, it is considered not reliable.

True or not reliable. This class of utterances is similar to the not reliable ones. They are also related to the topic of investigation, and the judgment does not provide information about their truthfulness. Nevertheless, according to the event and/or on the basis of a weak connection with the interests that the speaker tries to defend, common sense induces one to believe that these utterances could be truth. The

[3] In Italian, unlike the English language, often adjectives follow the noun (for example, "electronic engineer" is "ingegnere elettronico"). Therefore the lawyer was just waiting for the witness to complete his sentence.

boundaries of this class of utterances resulted in being too subjective, and this caused problems of agreement between annotators, as discussed in the next Subsection.

False or not reliable. This is the specular situation in respect to the previous point: the only difference is that the utterances seem to be false, even though their deceptiveness has not been clearly established by the inquiries.

Undecidable. Utterances that, from a logical point of view, cannot be either true or false, are considered undecidable. Belonging to this class are questions, such as "Excuse me, can you repeat?", but also of several utterances stopped in mid-sentence. This is also the case of utterances which have meta-communicative function, and regulate the relations between actors, such as "Now I'll explain." or "I would like to see you, if you were me..." and so on.

3.4.3 Agreement evaluation

The first studies carried out on DeCour (Fornaciari and Poesio, 2011a,b) concerned preliminary analyses carried out on data collected in only three Courts, which represented the first nucleus of DeCour. Since the study regarding the agreement between different annotators was not completed, these studies relied only on utterances held as surely true or false, having discarded the other ones.

The agreement study regarding the coding scheme described above was carried out employing three coders, each of whom marked 605 utterances, which meant about 20% of the final size of DeCour. Kappa has been used as metric to evaluate their agreement (Artstein and Poesio, 2008), and its value was calculated under four different conditions, as follows:

6 classes. The agreement was calculated on the previous coding scheme as it has been described;

4 classes. The utterances marked up as *true or not reliable* were collapsed into the class *true*, while the *false or not reliable* utterances, in turn, were joined to the *false* ones. Then the four classes became *true*, *false*, *not reliable* and *undecidable*.

3 classes. In this condition *true* and *true or not reliable* utterances, *false* and *false or not reliable*, and lastly *not reliable* and *undecidable* were respectively collapsed together into the classes *true*, *false*, *uncertain*.

2 classes. In the last condition, the *false or not reliable* utterances were joined to the class *false*, while all the remaining utterances were collapsed into the generic class *not false*.

The values of Kappa under the different conditions are shown in Table 3.1. The values of K for two classes indicate a moderate to substantial agreement depending on whether we choose the interpretation of K values proposed by Carletta (1996) or that proposed by Landis and Koch (1977). Given that the fine-grained original annotation scheme was not suitable to reach a satisfying agreement between coders, in the end the whole DeCour was annotated according to the only three collapsed classes: **true**, **uncertain** and **false**.

Table 3.1. Kappa values of the agreement studies.

Classes evaluated	Kappa values
6 classes	.40
4 classes	.56
3 classes	.57
2 classes	.64

3.5 Corpus statistics

DeCour has been collected in the Courts of four Italian towns: Bologna, Bolzano, Prato and Trento. It is constituted of 35 hearings, issued by 31 subjects. They appear 19 times as witnesses, 14 times as defendants, one time as expert witness and one time as victim of another crime. Their mean age at the time of the hearing is slightly higher than 36. 23 are men, 7 women and one transgender. The region of birth is northern Italy for 12 of them, center for 2, south for 9, while 8 subjects were foreigner but good Italian speakers. Lastly, the education is known only for six subjects: four of them having a high school education, one middle school and the last one elementary school.

Table 3.2 shows the number of turn and utterances of the participants in the hearings. While the utterances of other figures are not taken into consideration, the 3015 utterances of the speakers have been labeled as shown in Table 3.3: that is DeCour contains 31.34% of false, 39.87% of true and 28.79% of uncertain utterances.

Table 3.2. Turns and utterances in DeCour.

Figure	Turns	Utterances
Speakers	2094	3015
Public prosecutors	1002	1323
Judges	921	1201
Defendant lawyers	388	527
Police officers	3	4
Tot.	4408	6070

Table 3.3. Labels of DeCour's utterances.

Label	Nr.
True	1202
Undecidable	868
False	945
Total	3015

In terms of tokens, the size of DeCour, with and without punctuation, is shown in Table 3.4. As stated above, punctuation marks are considered in blocks: this means, for example, that a single dot and the three dots of the ellipsis are both considered as a single token.

Table 3.4. DeCour's size.

Utterances	Tokens			
	With punct.		Without punct.	
	Mean	Tot.	Mean	Tot.
True	12.86	15456	10.67	12847
Uncertain	12.02	10439	9.99	8669
False	16.85	15924	14.15	13376
Total		41819		34892

Chapter 4

Methods

In the next Chapter we will present several experiments concerned with the development of computational models for deception detection based on machine learning techniques. In this Chapter 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. In fact, since our utterances are transcriptions of spoken language and the punctuation marks were inserted by the transcriber, it was judged opportune to keep trace both of the exact number of words that the subject meant to issue and of the meaningful pauses detected by the transcriber.

4.1.2 LIWC features

Our first experiments were devoted to replicate Newman et al. (2003)'s study, employing the Italian version of LIWC software (Alparone et al., 2004).[1] The LIWC software outputs a few types of surface information about utterances in addition to the lexical 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**

[1] The LIWC for several languages can be obtained from www.liwc.net.

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"[2] were discarded: "Loro" is used to categorize only one lexical item in the dictionary, whereas "Passivo" and "Formale" are simply the Italian translation of English dictionary's categories, but they are not related to any Italian 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. All *n*-grams include punctuation marks. Since these were inserted in the texts by the transcribers, pilot experiments were carried out employing *n*-grams with and without punctuation marks, in order to ascertain how they affect the performance in the classification task. Punctuation marks turned out to be useful to improve the performance of the trained models, therefore the *n*-grams of all our experiments include them.

In each experiment, the frequency lists of *n*-grams 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 4.1.

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

Table 4.1. 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

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 4.1 were arrived at on the basis of extensive empirical experimentation. The figure of 196 features in Table 4.1 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, 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:

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

where e = entropy, tp = true positives,[3] fp = false positives, tn = true negatives, fn = false negatives,

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

and

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

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

To compute the Information Gain of a feature again 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 most 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:

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

[3] Because the scientific focus of this work is to verify if it is possible to identify deceptive statements, the 'positives' are the false utterances.

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 thesis we will omit the $_f$ indices except when required.

The models' performance was measured in experiments based on n-fold cross-validation. In the Chapter 6, the metrics mentioned above are showed as the overall result of the performance of all the folds constituting the experiment. However, in every experiment we show the mean accuracy of each fold as well. The mean accuracy of each fold is usually lower than the total accuracy of all the folds, therefore this is assumed as the most cautious measure of the performance of the models.

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 3015 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 2070 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 100000 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. 2147 utterances remained, of which 1202 true and 945 false, as above. Out of the 100000 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 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 a previous work (Fornaciari and Poesio, 2011b) 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 100000 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.

Chapter 5

Practical realization

This chapter describes from a practical point of view how the data constituting DeCour were collected and how the experiments were carried out.

5.1 Data collection

To collect data to create DeCour was quite complicated and time-consuming, and perhaps this explains why this kind of study was never carried out before. The first step was to get in touch with the Presidents of the Courts in which the data were collected. The Courts - that is those of Bologna, Bolzano, Prato and Trento - were identified according a simple criterion of logistic opportunity in order to hold down the costs of the research, and in one case because of a previous personal knowledge of a Public Prosecutor who allowed us to be directly introduced to the President of the Court.

The cooperation which was requested from the Presidents of the Court consisted in the authorization to examine the files of criminal proceedings for 'calumny' and 'false testimony' and to collect copy of the records reckoned to be interesting for the research, namely the transcripts of the hearings and the judgments. In spite of, or maybe thanks to his peculiarity and novelty, the request was favorably welcomed in every Court. As stated above, the only condition posed was to safeguard the privacy of the subjects involved in the criminal proceedings used to create DeCour.

However, the request was conceived to be as less demanding as possible, from the point of view of the support needed from the personnel of the Courts. In fact, starting from the end of the past century, in each Italian Court the so called 'Registro Generale - Re.Ge.' began to be fed. This is a database that allowed, since that moment, to carry out archive researches according to different parameters, among which the type of crime. Without this system, it would not have been possible to find the dossiers, if not through a manual research file by file: and this is the reason why DeCour contains only criminal proceedings held starting from 1999 (and until 2008, since few years are usually necessary to know the conclusion of all the degrees of judgment of the proceedings). As far the Court's personnel was concerned, their task was simply to carry out the research on Re.Ge. in order to

provide us with the list of files to be consulted. Apart from the Court of Bolzano, we got directly to the archives in order to find the files.

5.2 Text processing

The Courts' files are stored in hard copy archives. This meant to scan directly in the Courts the documents of interest. All the documents were saved as images in pdf format. The transcripts of the hearings were also saved as text files, through the conversion of the images by means of Optical Character Recognition (OCR). These tasks were carried out using the software Omnipage.[1] Due to the low quality of the hard copies and to the frequent presence of underlinings and notes on the pages, the OCR's output was often unsatisfactory. Therefore all the texts of the hearings were reread and manually corrected, in accordance to the content of the original documents. This manual correction was also used to insert in the text some simple markers, which in turn were employed as cues for the transformation of the texts in XML files. The transformation from text to XML files was realized using Perl Programming Language[2] in the ActivePerl Business and Enterprise Edition.[3]

In addition to the transcripts of the hearings, into the XML files was inserted also the output of TreeTagger[4] - lemmas and POS (in this way it was possible to collect these features directly from the XML files having run TreeTagger just once, rather than every time needed). Figure 5.1 shows an example of an XML file of DeCour.

5.3 Datasets' creation

Perl was also used to collect the features from the XML files. The frequency lists of n-grams and their Information Gain were extracted through Perl scripts employing the package XML::DOM.[5] Then the frequencies of the selected features in each utterance were used to fill the dataset files. These files were matrices in which every row represented an entity - in the case of DeCour an utterance issued by the subject interrogated in the

[1] http://www.nuance.com/for-business/by-product/omnipage/index.htm

[2] http://www.perl.org

[3] http://www.activestate.com/activeperl

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

[5] http://search.cpan.org/~tjmather/XML-DOM-1.44/lib/XML/DOM.pm

Figure 5.1. An example of XML file of DeCour.

```
<hearing>
  <header birtharea="N" birthplace="xxxxx" birthyear="xxxxx" court="BZ" day="xxxxx" idsub="xxxxx" month="xxxxx" name="xxxxx"
      nrdoss="xxxxx" nrhear="xxxxx" sex="M" study="unk" typesub="defwit" typetest="false" year="xxxxx" yeardoss="03"/>
  <intro> giudice monocratico - dott.ssa xxxxx xxxxx viene introdotto il testimone ; questi viene avvertito dal giudice dei suoi obblighi e rende la
      dichiarazione ex art. 497 c.p.p. . fornisce le generalità : xxxxx xxxxx , nato a xxxxx il xx xxxxx xxxxx , ivi residente . </intro>
  <turn nrgen="1" nrpros="1" speaker="pros">
    <utterance class="x" nrgen="1" nrpros="1">
      lei nella primavera del 2001 ci può dire come ha conosciuto xxxxx xxxxx , in quali circostanze ?
    </utterance>
    <lemmas class="x" nrgen="1" nrpros="1">
      lei nella primavera del @card@ ci potere dire come avere conoscere <unknown> <unknown> , in quale circostanza ?
    </lemmas>
    <pos class="x" nrgen="1" nrpros="1">
      PRO:pers ARTPRE NOUN ARTPRE NUM CLI VER2:fin VER:infi WH AUX:fin VER:ppast NOUN ADJ PUN PRE DET:wh NOUN SENT
    </pos>
  </turn>
  <turn nrgen="2" nrsub="1" speaker="defwit">
    <utterance class="uncertain" nrgen="2" nrsub="1">
      adesso non mi ricordo come l' ho conosciuto , comunque ci siamo conosciuti ...
    </utterance>
    <lemmas class="uncertain" nrgen="2">
      adesso non mi ricordare come l' avere conoscere , comunque ci essere conoscere ...
    </lemmas>
    <pos class="uncertain" nrgen="2">
      ADV NEG CLI VER:fin WH CLI AUX:fin VER:ppast PUN WH CLI AUX:fin VER:ppast SENT
    </pos>
    <utterance class="uncertain" nrgen="3" nrsub="2">
      non mi ricordo , in giro , al xxxxx anche , perché prendevo il metadone tempo fa .
    </utterance>
    <lemmas class="uncertain" nrgen="3">
      non mi ricordare , in giro , al <unknown> anche , perché prendere il metadone tempo fa .
    </lemmas>
    <pos class="uncertain" nrgen="3">
      NEG CLI VER:fin PUN PRE NOUN PUN ARTPRE NOUN ADV PUN WH VER:fin ART NOUN NOUN ADV SENT
    </pos>
    <utterance class="uncertain" nrgen="4" nrsub="3">
      adesso sono due anni che sono a posto , quasi due anni .
    </utterance>
    <lemmas class="uncertain" nrgen="4">
      adesso essere due anno che essere a posto , quasi due anno .
    </lemmas>
    <pos class="uncertain" nrgen="4">
      ADV VER:fin DET:num NOUN CHE VER:fin PRE NOUN PUN ADV DET:num NOUN SENT
    </pos>
  </turn>
```

hearing. The columns of the datasets contained:

- Some meta-information about the subjects themselves, such as age, sex and so on;
- The class of the utterance, that is 'false', 'true' and 'uncertain';
- The features properly said, which describe the utterance in the vector space.

Figure 5.2 shows a fragment of dataset in DeCour.

As far as the LIWC features are concerned, they were also collected by a Perl script based on the LIWC Italian dictionary (Alparone et al., 2004), rather than by the use of the LIWC software itself. This allowed to integrate directly the LIWC features into the

Figure 5.2. A fragment of dataset of DeCour.


```
id,hear,sex,age,birtharea,study,typesub,class,classcollapsed,uttlen,uttlennop,QUOTATION_MARK,COMMA,DOT,DOT_DOT_DOT,
BO000737199611070201 3,BO000737199611070201,F,thirty,N,M,defwit,1,1,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
BO000737199611070201 5,BO000737199611070201,F,thirty,N,M,defwit,1,1,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
BO000737199611070201 7,BO000737199611070201,F,thirty,N,M,defwit,1,1,33,28,0,3,1,0,0,0,2,0,0,0,0,0,0,0,0,0,0,3,0,0
BO000737199611070201 9,BO000737199611070201,F,thirty,N,M,defwit,1,1,8,6,0,1,1,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
BO000737199611070201 11,BO000737199611070201,F,thirty,N,M,defwit,1,1,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 13,BO000737199611070201,F,thirty,N,M,defwit,1,1,54,50,0,3,1,0,1,0,0,1,0,0,1,2,0,0,0,3,2,1,3,2,
BO000737199611070201 15,BO000737199611070201,F,thirty,N,M,defwit,2,1,11,9,0,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0
BO000737199611070201 17,BO000737199611070201,F,thirty,N,M,defwit,1,1,4,3,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 19,BO000737199611070201,F,thirty,N,M,defwit,0,0,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 21,BO000737199611070201,F,thirty,N,M,defwit,2,1,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 23,BO000737199611070201,F,thirty,N,M,defwit,2,1,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 25,BO000737199611070201,F,thirty,N,M,defwit,1,1,7,6,0,0,1,0,2,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
BO000737199611070201 27,BO000737199611070201,F,thirty,N,M,defwit,1,1,13,11,0,1,1,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,1,0,
BO000737199611070201 29,BO000737199611070201,F,thirty,N,M,defwit,1,1,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 30,BO000737199611070201,F,thirty,N,M,defwit,1,1,9,8,0,0,1,0,0,0,0,0,0,1,0,0,0,0,2,0,0,0,0,0,0,
BO000737199611070201 32,BO000737199611070201,F,thirty,N,M,defwit,1,1,5,4,0,0,1,0,3,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,
BO000737199611070201 34,BO000737199611070201,F,thirty,N,M,defwit,2,1,6,4,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,
BO000737199611070201 36,BO000737199611070201,F,thirty,N,M,defwit,2,1,24,21,0,2,1,0,0,0,0,0,0,1,1,2,1,0,0,1,0,0,3,0,
BO000737199611070201 37,BO000737199611070201,F,thirty,N,M,defwit,1,1,15,12,0,2,1,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 39,BO000737199611070201,F,thirty,N,M,defwit,1,1,7,5,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
BO000737199611070201 41,BO000737199611070201,F,thirty,N,M,defwit,0,0,12,9,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1
BO000737199611070201 43,BO000737199611070201,F,thirty,N,M,defwit,1,1,12,10,0,1,1,0,1,0,0,0,0,1,0,0,0,0,0,0,0,1,1,0,
BO000737199611070201 45,BO000737199611070201,F,thirty,N,M,defwit,2,1,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 47,BO000737199611070201,F,thirty,N,M,defwit,0,0,31,26,0,4,1,0,0,0,0,0,0,2,0,2,0,0,0,0,0,0,4,2,
BO000737199611070201 49,BO000737199611070201,F,thirty,N,M,defwit,0,0,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 51,BO000737199611070201,F,thirty,N,M,defwit,1,1,6,4,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
BO000737199611070201 53,BO000737199611070201,F,thirty,N,M,defwit,0,0,6,4,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 55,BO000737199611070201,F,thirty,N,M,defwit,1,1,5,3,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
BO000737199611070201 58,BO000737199611070201,F,thirty,N,M,defwit,1,1,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 60,BO000737199611070201,F,thirty,N,M,defwit,0,0,10,6,0,3,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
BO000737199611070201 62,BO000737199611070201,F,thirty,N,M,defwit,0,0,9,7,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 64,BO000737199611070201,F,thirty,N,M,defwit,0,0,4,3,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 68,BO000737199611070201,F,thirty,N,M,defwit,2,1,6,5,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,
BO000737199611070201 71,BO000737199611070201,F,thirty,N,M,defwit,0,0,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 73,BO000737199611070201,F,thirty,N,M,defwit,0,0,2,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 75,BO000737199611070201,F,thirty,N,M,defwit,2,1,3,2,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 77,BO000737199611070201,F,thirty,N,M,defwit,1,1,10,8,0,1,0,0,2,0,1,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0
BO000737199611070201 79,BO000737199611070201,F,thirty,N,M,defwit,0,0,12,10,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
BO000737199611070201 80,BO000737199611070201,F,thirty,N,M,defwit,0,0,16,15,0,0,1,0,2,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,
BO000737199611070201 82,BO000737199611070201,F,thirty,N,M,defwit,0,0,5,3,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```


datasets described above, applying the LIWC dictionary to texts processed according our tokenization, which took into consideration the specific characteristics of DeCour. Also the linguistic dimensions of LIWC were tuned according to the demands of DeCour, as discussed in 4.1.1 and in 4.1.2.

5.4 Data analysis

Each dataset as depicted above constituted a fold of the n-fold cross-validations carried out in the experiments, since the features were identified on the basis of different training sets of hearings. The datasets were analyzed using the environment for statistical computing R,[6] and its library e1071[7] (Dimitriadou et al., 2011; Meyer, 2004) which implements in R

[6] http://www.r-project.org

[7] http://cran.r-project.org/web/packages/e1071/index.html

the well-known SVM library LIBSVM[8] (Chang and Lin, 2011).

After several empirical trials in which the SVM parameters were tuned in order to optimize the the models' performance in our experiments, the options of 'svm' function in R were settled as follows:

- `type = "C-classification"`, which indicates the task the function is asked for;
- `kernel = "radial"`, that is the kind of kernel employed;
- `cost = 5`, this parameters enhances the complexity of the model in order to reduce the errors;
- `probability = TRUE`, which enables the option of receiving as output the probability according to which an entity is assigned to a class.

[8] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Part III

Experiments and results

Chapter 6

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, that is discarding the uncertain utterances, which in the previous group of experiments were grouped together with the true ones into the generic class of not-false utterances. Since the class of uncertain utterances contains statements lacking of propositional value, this is arguably a more realistic application of the methodology we used, which would only be employed 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.

6.1 Comparing Lexical and Surface Features

6.1.1 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.

6.1.2 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 Section 4.1.1 and the 84 LIWC features from Section 4.1.2.

The results of this experiment are summarized in Table 6.1.[1] The mean accuracy of each fold of the experiment in detecting false utterances was 68.28%, with standard deviation $\sigma = 8.86$. This rate of the mean accuracy is almost 6 points percent higher than that of the heuristic algorithm, but does not exceed the majority baseline. However, the total accuracy of all the folds of the experiment is higher than the majority baseline, being the first 69.35% and the second 68.66%.

Table 6.1. 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	1747	323	74.40%	84.40%	79.09%
Total	2091	924			
Total accuracy	69.35%	30.65%			
Mean accuracy	*68.28%*				
Monte Carlo baseline	60.03%				
Majority baseline	**68.66%**				
Heuristic baseline	62.39%				

6.1.3 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 6.2. The mean accuracy of the models was 68.29%, with standard deviation $\sigma = 11.13$. As in the previous

[1] Here and in the rest of the dissertation we indicate the highest accuracy achieved in bold. The total accuracy is not considered; the mean accuracy is considered instead, as this is a more prudent estimation of the models' performance.

experiment, the performance is higher than that of the heuristic baseline and random choice, but not than that of the majority baseline. However, also in this case the total accuracy is better than the random 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 4.1.

Table 6.2. 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	1804	266	73.81%	87.15%	79.93%
Total	2109	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 6.3. 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 percentage points, respectively). The feature vectors in this case consisted of 252 features: 250 surface features and the two utterance length features.

6.1.4 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.

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 percentage points) and the majority baseline (although only by a few tenths of point). The

Table 6.3. 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	1723	347	75.74%	83.24%	79.31%
Total	2116	899			
Total accuracy	70.18%	29.82%			
Mean accuracy	**69.89%**				
Monte Carlo baseline	60.03%				
Majority baseline	68.66%				
Heuristic baseline	62.39%				

overall performance of the 35-fold cross-validation is presented in Table 6.4.

Table 6.4. 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	1800	270	74.44%	86.96%	80.21%
Total	2127	888			
Total accuracy	70.55%	29.45%			
Mean accuracy	**68.96%**				
Monte Carlo baseline	60.03%				
Majority baseline	68.66%				
Heuristic baseline	62.39%				

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 percentage points higher than the heuristic baseline, but it is slightly lower than the majority baseline (which in turn is lower than the total accuracy of 69.88%). Table 6.5 summarizes the results.

Table 6.5. 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	1725	345	75.39%	83.33%	79.16%
Total	2107	908			
Total accuracy	69.88%	30.12%			
Mean accuracy	*68.59%*				
Monte Carlo baseline	60.03%				
Majority baseline	**68.66%**				
Heuristic baseline	62.39%				

6.2 Discriminating between clearly False and clearly True utterances

6.2.1 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 dissertation 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, he often mentions 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.

6.2.2 Using the LIWC

Table 6.6 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 percentage 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).

Table 6.6. 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	1465	682			
Total percent	68.23%	31.77%			
Mean accuracy	**66.48%**				
Monte Carlo baseline	54.54%				
Majority baseline	55.98%				
Heuristic baseline	59.57%				

6.2.3 Surface features

Best Frequencies.

Table 6.7 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 percentage 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 6.8.

Table 6.7. 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	1500	647			
Total percent	69.86%	30.14%			
Mean accuracy	**68.62%**				
Monte Carlo baseline	54.54%				
Majority baseline	55.98%				
Heuristic baseline	59.57%				

Table 6.8. 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	1493	654			
Total percent	69.54%	30.46%			
Mean accuracy	**68.25%**				
Monte Carlo baseline	54.54%				
Majority baseline	55.98%				
Heuristic baseline	59.57%				

6.2.4 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 6.9: 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 percentage points.

Table 6.9. 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	1516	631			
Total percent	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-validation experiment, where only true and false utterances were considered. The results are shown in Table 6.10. The mean accuracy is 68.90%, with standard deviation $\sigma = 11.18$: that is more than 8 points percent higher than heuristic baseline.

Table 6.10. 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	1508	639			
Total percent	70.24%	29.76%			
Mean accuracy	**68.90%**				
Monte Carlo baseline	54.54%				
Majority baseline	55.98%				
Heuristic baseline	59.57%				

6.3 Selecting more homogeneous sets of defendants

6.3.1 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: (i) only considering defendants of the same gender (unfortunately we only have enough data to try this on male defendants); (ii) only Italian native speakers; and (iii) defendants of a similar age. We consider each of these in turn.

6.3.2 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 2234 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%.

Table 6.11. 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	1258	264	74.97%	82.65%	78.62%
Total	1550	684			
Total accuracy	69.38%	30.62%			
Mean accuracy	**69.51%**				
Monte Carlo baseline	60.11%				
Majority baseline	68.13%				
Heuristic baseline	62.58%				

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 6.11 presents the overall results in this experiment.

6.3.3 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 2177 utterances, of which 625 (28.71%) were false. Therefore, the Majority Baseline was 71.29%. By contrast, 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. However, the total accuracy of all the folds constituting the experiment was slightly higher than the majority baseline. Table 6.12 summarizes the results.

Table 6.12. 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	1299	253	77.83%	83.70%	80.66%
Total	1554	623			
Total accuracy	71.38%	28.62%			
Mean accuracy	*70.12%*				
Monte Carlo baseline	62.56%				
Majority baseline	**71.29%**				
Heuristic baseline	64.22%				

6.3.4 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

Table 6.13. 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	1088	232	75.92%	82.42%	79.04%
Total	1340	577			
Total accuracy	69.90%	30.10%			
Mean accuracy	**70.28%**				
Monte Carlo baseline	60.93%				
Majority baseline	68.86%				
Heuristic baseline	63.90%				

remaining dataset consisted of 1917 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 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 6.13 shows the overall performance of the model, which is better than both the majority and heuristic thresholds.

Chapter 7

Discussion

7.1 Predicting deception

Our first result is that all models proposed in Chapter 6 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 one percentage point. In the rest of this subsection we discuss more in detail what makes the task so hard.

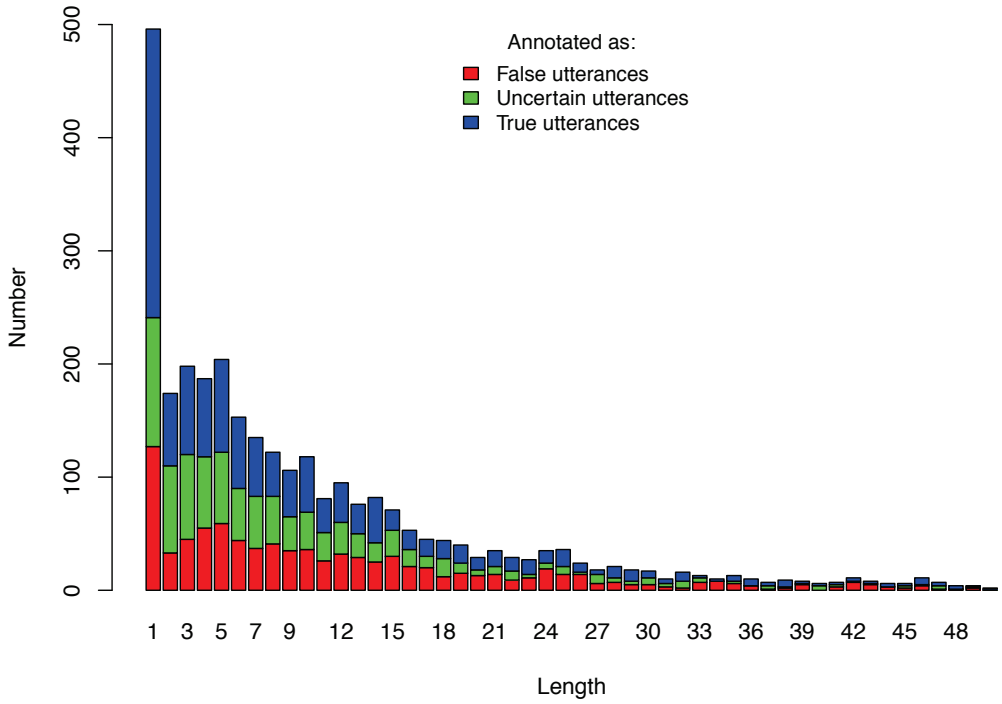
Deception at the utterance level

A first point to note is that being able to achieve a better performance 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 7.1 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 undistinguishable 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

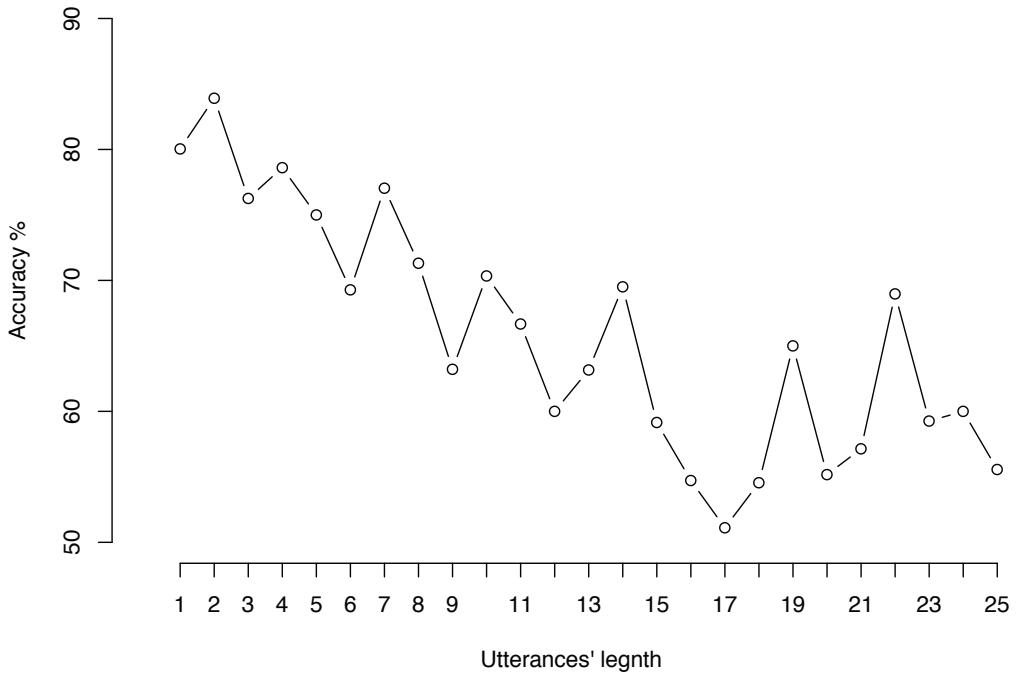
Figure 7.1. The distribution of the lengths of the utterances in DeCour.



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 something, is always around 2 percentage 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 Figure 7.2, in which utterance length and classification accuracy in the experiment using surface features selected through Information Gain (Table 6.3) are charted. Clearly, the longer the utterances, the lower the accuracy. Since short statements are typically conventional, that is made by stereotypical linguistic formulas, this suggests that formulaic language could be a good predictor in order to classify statements as true or false.

Figure 7.2. The relation between utterance length and classification accuracy.

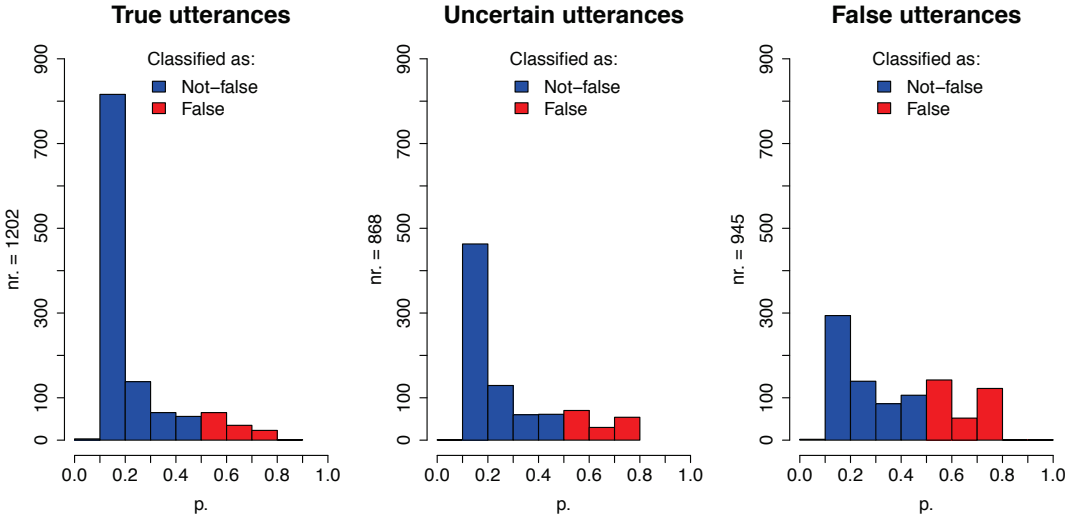


Uncertainty and noise.

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 percentage points. As explained above, the class of uncertain utterances consists of (i) utterances which cannot have a value of true or false (e.g., questions) or (ii) 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 Figure 7.3. 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 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.

Figure 7.3. The probabilities with which the utterances are classified as false or not-false, in each class of utterances.



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.

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

7.2 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 7.1 shows the 48 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,

Table 7.1. 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
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

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 7.2 contains the lists of the twenty most frequent tokens, bigrams, trigrams and tetragrams of true and false utterances.[1] 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 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.

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

Table 7.2. N-grams Frequency in DeCour

True utterances							
Tokens	Freq.	Bigrams	Freq.	Trigrams	Freq.	Tetragrams	Freq.
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' è 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							
Tokens	Freq.	Bigrams	Freq.	Trigrams	Freq.	Tetragrams	Freq.
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 è 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

The mean values of the LIWC dimensions with the greatest differences in value for true and false utterances are shown in Tables 7.3 and 7.4, 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).

Table 7.3. 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

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, 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

Table 7.4. 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

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.[2] Minor errors in the output of the parser were then hand-corrected using simple heuristic 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 7.5.

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.

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

Table 7.5. First person pronouns and verbs in true and false utterances

	False Utterances	True Utterances
Number	945	1202
Tokens	15924	15456
Pronoun "Io"-"I"	291	191
First person pronouns ("Io"-"I", "me/mi"-"me")	393	257
First person verbs	1057	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

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.

7.3 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 documents. 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.

Appendix

Appendix A

Istructions for coders

Annotazione delle false testimonianze

Caro annotatore,

ti consegnerò una serie di sentenze di diversi Tribunali, emesse in procedimenti penali per "calunnia" e "falsa testimonianza". In queste sentenze, gli imputati vengono giudicati per aver mentito o calunniato qualcuno, nel corso di un'udienza in cui erano chiamati a testimoniare.

Le sentenze si chiudono con la condanna dell'imputato oppure, molto raramente, con un'assoluzione dovuta a motivi procedurali. Sempre, tuttavia, in queste sentenze vengono ricostruiti i fatti su cui il soggetto ha testimoniato, e vengono individuate le menzogne che ha pronunciato in aula.

Dopo aver letto ogni sentenza, che potrai sempre consultare a tuo piacimento, dovrai leggere il verbale dell'udienza a cui la sentenza si riferisce. Durante la lettura, alla luce dei fatti così come ricostruiti nella sentenza, che vengono considerati rispondenti al vero, dovrai esprimere un giudizio su ciascuna frase pronunciata dal soggetto che viene sentito.

Potrai etichettare le frasi scegliendo una delle seguenti categorie:

- Le categorie della certezza:

  False La frase è chiaramente indicata nella sentenza come falsa, o la falsità è una conseguenza logica dei fatti, così come sono stati ricostruiti dal Giudice.

  Ad esempio, se un soggetto, mentendo, afferma di non essersi incontrato con una persona in un determinato luogo, necessariamente mente anche quando afferma di non conoscerla affatto.

  In ogni caso, non è sempre facile distinguere una frase falsa da una vera. La menzogna, infatti, può essere definita come "una falsa dichiarazione resa con la deliberata intenzione di ingannare", ma anche più in generale come "qualche cosa tesa a creare una falsa impressione". In questa ottica, la corretta interpretazione del comportamento linguisitco costringe a concentrarsi più sulla funzione delle singole frasi, piuttosto che sul loro significato latterale.

Ad esempio, le frasi interrogative di per sè non rappresentano fatti, e dunque non potrebbero avere valore nè di veridicità , nè di falsità . Ma se un testimone, di nuovo fingendo di non conoscerne un altro, alla domanda del Giudice "Conosce il signor Rossi?", risponde "Sarebbe il signore qui davanti a me?", il soggetto con la sua risposta espressa in forma interrogativa sta cercando di ingenerare nel Giudice il falso convincimento che non lo conosce. Pertanto la risposta dovrebbe essere etichettata come falsa.

In altri casi, tuttavia, occorre al contrario fare molta attenzione al significato delle parole. In un processo, un soggetto si spaccia per ingegnere, mentre è soltanto diplomato. Nel corso del dialogo, l'avvocato dice: "Lei ha detto di essere ingegnere elettronico?", ed il soggetto risponde "Sì". In questo caso il soggetto, pur non essendolo, aveva veramente affermato di essere un ingegnere. Pertanto la risposta è vera, rispetto alla formulazione della domanda.

Infine, a volte la difficoltà può nascere dal fatto che una singola frase può contenere più proposizioni, delle quali alcune vere e altre false: se nella frase è presente anche un solo elemento di falsità , essa deve essere valutata come falsa.

True La frase descrive i fatti in modo coerente con quanto ricostruito dalla sentenza.

Possono essere considerate vere le frasi su cui la sentenza non si pronuncia, in quanto vertono su argomenti la cui veridicità non incide sulla dinamica degli eventi e sugli interessi del soggetto. Ad esempio, se il Pubblico Ministero chiede al soggetto "Per quanto tempo è stato sposato con la signora Bianchi?", e questi risponde "per otto anni", se tale risposta non incide sui fatti oggetto di indagine può essere considerata vera, anche se la sentenza non si pronuncia su tale punto.

- Le categorie delle opinioni. Quando ritieni che le informazioni presenti nella sentenza non siano sufficienti per esprimere giudizi certi di verità o falsità , sei chiamato a pronunciarti secondo la tua opinione, secondo queste categorie:

  Probably false Sono le frasi inerenti ai fatti oggetto di indagine, su cui appunto la sentenza non si pronuncia. Tuttavia, secondo la tua interpretazione dell'evento, ritieni che esse siano probabilmente false.

  Probably true La medesima situazione di cui sopra, ma stavolta la tua opinione è che probabilmente si tratti di una frase vera.

  Not reliable Sono le frasi su cui secondo te la sentenza non offre certezze, e tu non sapresti pronunciarti sulla loro probabile veridicità o falsità . È in pratica la risposta "non so".

- Le frasi non decidibili:

  Undecidable Queste sono le frasi che, secondo te,da un punto di vista logico non possono assumere nè valore di verità nè di falsità . È il caso delle domande ("Scusi, può ripetere?"), delle frasi lasciate a metà ("Veramente, io..."), delle frasi con funzione meta-comunicativa ("Lei può pensare quello che vuole!"), e così via.

Grazie per la collaborazione e buon lavoro!

Bibliography

Bibliography

Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., and Sartori, G. (2011). Detecting fakers of the autobiographical iat. *Applied Cognitive Psychology*, 25(2):299–306.

Alparone, F., Caso, S., Agosti, A., and Rellini, A. (2004). The Italian LIWC2001 Dictionary. Austin, TX: LIWC.net.

Arntzen, F. (1970). *Psychologie der Zeugenaussage*. Hogrefe, Göttingen, Germany.

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

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

Backster, C. (1962). Methods of strengthening our polygraph technique. *Police*, 6:61–68.

Backster, C. (1963). The backster chart reliability rating method. *Law and Order*, 1:63–64.

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

Bond, C. F. and Uysal, A. (2007). On lie detection "wizard". *Law and Human Behavior*, 31(1):109–115.

Bond, G. and Lee, A. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, 19:313–329.

Brett, A., Phillips, M., and Beary, J. (1986). Predictive power of the polygraph: can the "lie detector" really detect liars? *The Lancet*, 327(8480):544–547.

Bui, C. (2006). *Morte tra le rovine. I segreti dell'indagine criminale*. Analisi Criminale. Centro Scientifico Editore.

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

Burgoon, J., Buller, D., Ebesu, A., White, C., and Rockwell, P. (1996). Testing interpersonal deception theory: Effects of suspicion on communication behaviors and perception. *Communication Theory*, 6:243–267.

Burgoon, J., Buller, D., White, C., Afifi, W., and Buslig, A. (1999). The role of conversation involvement in deceptive interpersonal interactions. *Personality and Social Psychology Bulletin*, 25:669–685.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22(2):249–254.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

Chartrand, T. and Bargh, J. (1999). The chameleon effect: The perception- behavior link and social interaction. *Journal of Personality and Social Psychology*, 76:893–910.

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

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Academic Press, New York.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20.

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25(4):431–447.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D., Acharyya, M., Loughead, J., Gur, R., and Langleben, D. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, 28(3):663 – 668.

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

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A. (2011). r-crane1071. http://mloss.org/software/view/94/.

Ekman, P. (1981). Mistakes when deceiving. *Annals of the New York Academy of Sciences*, 364:269–278.

Ekman, P. (1989). Why lies fail and what behaviors betray a lie. In Yuille, J., editor, *Credibility assessment*, pages 71–82. Kluwer, Dordrecht, the Netherlands.

Ekman, P. (2001). *Telling lies: clues to deceit in the marketplace, politics, and marriage*. W.W. Norton.

Ekman, P., Friesen, W., and Simons, R. (1985). Is the startle reaction an emotion? *Journal of Personality and Social Psychology*, 49:1416–1426.

Ekman, P. and O'Sullivan, M. (2006). From flawed self-assessment to blatant whoppers: The utility of voluntary and involuntary behavior in detecting deception. *Behavioural Sciences & the Law*, 24:673–686.

Fisher, R. and Geiselman, R. (1992). *Memory-Enhancing Techniques for Investigative Interviewing: The Cognitive Interview*. Charles C. Thomas, Springfield, IL, England.

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

Fitzpatrick, E. and Bachenko, J. (2012). Building a data collection for deception research. In Fitzpatrick, E., Bachenko, J., and Fornaciari, T., editors, *Proc. of the EACL Workshop on Computational Approaches to Deception Detection*, pages 31–38.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305.

Fornaciari, T. and Poesio, M. (2011a). Lexical vs. surface features in deceptive language analysis. In *Proceedings of the ICAIL 2011 Workshop Applying Human Language Technology to the Law*, AHLTL 2011, pages 2–8, Pittsburgh, USA.

Fornaciari, T. and Poesio, M. (2011b). Sincere and deceptive statements in italian criminal proceedings. In *Proceedings of the International Association of Forensic Linguists Tenth Biennial Conference*, IAFL 2011, Cardiff, Wales, UK.

Fornaciari, T. and Poesio, M. (2012a). DeCour: a corpus of DEceptive statements in Italian COURts. In Calzolari, N. C. C., Choukri, K., Declerck, T., Uäÿur Döäÿn, M., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Fornaciari, T. and Poesio, M. (2012b). On the use of homogenous sets of subjects in deceptive language analysis. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 39–47, Avignon, France. Association for Computational Linguistics.

Frank, M. and Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology*, 72:1429–1439.

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

Frank, M. G., Menasco, M. A., and O'Sullivan, M. (2008). Human behavior and deception detection. In Voeller, J. G., editor, *Wiley Handbook of Science and Technology for Homeland Security*. John Wiley & Sons, Inc.

Freud, S. (1913). *Il Mosè di Michelangelo*. Bollati Boringhieri Editore, Torino, 1976 edition.

Gamer, M., Rill, H., Vossel, G., and Gödert, H. W. (2006). Psychophysiological and vocal measures in the detection of guilty knowledge. *International Journal of Psychophysiology*, 60:76–87.

Ganis, G., Kosslyn, S., Stose, S., Thompson, W., and Yurgelun-Todd, D. (2003). Neural correlates of different types of deception: An fmri investigation. *Cerebral Cortex*, 13(8):830–836.

Garrido, E., Masip, J., and Herrero, C. (2002). Police officers' credibility judgments: Accuracy and estimated ability. *International Journal of Psychology*, 39:276–289.

Gokhmann, S., Hancock, J., Prabhu, P., Ott, M., and Cardie, C. (2012). In search of a gold standard in studies of deception. In Fitzpatrick, E., Bachenko, J., and Fornaciari, T., editors, *Proc. of the EACL Workshop on Computational Approaches to Deception Detection*, pages 23–30.

Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. (2008). On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes*, 45(1):1–23.

Hauch, V., Blandón-Gitlin, I., Masip, J., and Sporer, S. L. (2012). Linguistic cues to deception assessed by computer programs: A meta-analysis. In Fitzpatrick, E., Bachenko, J., and Fornaciari, T., editors, *Proc. of the EACL Workshop on Computational Approaches to Deception Detection*, pages 1–4, Avignon.

Hess, U. and Kleck, R. (1990). Differentiating emotion elicited and deliberate emotional facial expressions. *European Journal of Social Psychology*, 20:369–385.

Hill, M. and Craig, K. (2002). Detecting deception in pain expressions: The structure of genuine and deceptive facial displays. *Pain*, 98:135–144.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., and Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44.

Jensen, M. L., Meservy, T. O., Burgoon, J. K., and Nunamaker, J. F. (2010). Automatic, Multimodal Evaluation of Human Interaction. *Group Decision and Negotiation*, 19(4):367–389.

Johnson, M., Hashtroudi, S., and Lindsay, D. (1993). Source monitoring. *Psychological Bulletin*, 114:3–29.

Johnson, M. and Raye, C. (1981). Reality monitoring. *Psychological Review*, 88:67–85.

Johnson, M. and Raye, C. (1998). False memories and confabulation. *Trends in Cognitive Sciences*, 2:137–146.

Karatzoglou, A., Meyer, D., and Hornik, K. (2006). Support vector machines in r. *Journal of Statistical Software*, 15(9):1–28.

Kassin, S. and Fong, C. (1999). "i'minnocent!": Effects of training on judgments of truth and deception in the interrogation room. *Law and Human Behavior*, 23:499–516.

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

Köhnken, G. and Steller, M. (1988). The evaluation of the credibility of child witness statements in german procedural system. In Davies, G. and Drinkwater, J., editors, *The child witness: Do the courts abuse children?*, number 13 in Issues in Criminological and Legal Psychology, pages 37–45. British Psychological Society, Leicester, England.

Lamb, M., Sternberg, K., Esplin, P., Hershkowitz, I., Orbach, Y., and Hovav, M. (1997). Criterion-based content analysis: A field validation study. *Child Abuse and Neglect*, 21:255–264.

Landis, R. J. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Langleben, D., Schroeder, L., Maldjian, J., Gur, R., McDonald, S., Ragland, J., O'Brien, C., and Childress, A. (2002). Brain activity during simulated deception: An event-related functional magnetic resonance study. *NeuroImage*, 15(3):727 – 732.

Levine, T. R., Feeley, T. H., McCornack, S. A., Hughes, M., and Harms, C. M. (2005). Testing the Effects of Nonverbal Behavior Training on Accuracy in Deception Detection with the Inclusion of a Bogus Training Control Group. *Western Journal of Communication*, 69(3):203–217.

Logothetis, N. K. (2008). What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878.

Lord, R. D. (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 stylométrie. *Revue des études grecques*, 41:61–81.

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

Lykken, D. (1959). The gsr in the detection of guilt. *Journal of Applied Psychology*, 43:385–388.

Lykken, D. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, 44:258–262.

Lykken, D. (1988). The case against polygraph testing. In Gale, A., editor, *The polygraph test: Lies, truth, and science*, pages 111–126. Sage, London.

Lykken, D. (1991). Why (some) americans believe in the lie detector while others believe in the guilty knowledge test. *Integrative Physiological and Behavioral Science*, 126:214–222.

Lykken, D. (1998). *A tremor in the blood: Uses and abuses of the lie detector*. Plenum Press, New York.

Maffei, S. (2007). Ipnosi, poligrafo, narcoanalisi, risonanza magnetica: metodi affidabili per la ricerca processuale della verità? In De Cataldo Neuburger, L., editor, *La prova*

scientifica nel processo penale, Atti e documenti / Istituto Superiore Internazionale di Scienze Criminali ; 18, pages 420 – 428. Cedam, Padova.

Meissner, C. and Kassin, S. (2002). "he'sguilty": Investigator bias and judgments of truth and deception. *Law and Human Behavior*, 26:469–480.

Merikangas, J. R. (2008). Commentary: Functional mri lie detection. *J Am Acad Psychiatry Law*, 36(4):499–501.

Meyer, D. (2004). *Support Vector Machines: The interface to libsvm in Package e1071*. Techische Universitat Wien, Austria.

Morelli, G. (1880). *Die Werke Italienischer Meister in den Galerien von Muenchen, Dresden und Berlin: ein kritischer Versuch / von Ivan Lermolieff; aus dem Russischen uebersetzt von Johannes Schwarze*. E.A. Seemann, Leipzig.

Moriarty, J. (2009). Visions of deception: Neuroimaging and the search for evidential truth. *Akron Law Review*, 42(3):739–761.

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

Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M. (2003). Lying Words: Predicting Deception From Linguistic Styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.

Niederhoffer, K. G. and Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Olsson, J. (2008). *Forensic Linguistics*. Continuum International Publishing Group, London.

O'Sullivan, M. and Ekman, P. (2004). The wizards of deception detection. In Granhad, P. and Stromwall, L., editors, *Deception detection in forensic context*, pages 269–286. Cambridge Press, Cambridge, UK.

Paceri, R. and Montanaro, S. (1995). *La Polizia Scientifica*. Laurus Robuffo, Roma.

Pavlidis, J., Eberhardt, N., and Levine, J. (2002). Human behaviour: Seeing through the face of deception. *Nature*, 415:35–35.

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

Pepe, G., editor (1996). *La falsa donazione di Costantino.* Tea storica. TEA.

Polizia di Stato, P. d. S., editor (2003). *La Polizia Scientifica 1903-2003.* Laurus Robuffo, Roma.

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

Priori, A., Mameli, F., Cogiamanian, F., Marceglia, S., Tiriticco, M., Mrakic-Sposta, S., Ferrucci, R., Zago, S., Polezzi, D., and Sartori, G. (2008). Lie-specific involvement of dorsolateral prefrontal cortex in deception. *Cerebral Cortex*, 18(2):451–455.

Raskin, D. C. (1979). Orienting and defensive reflexes in the detection of deception. In Kimmel, H., Van Olst, E., and J.F., O., editors, *The orienting reflex in humans*, pages 587–605. Erlbaum, Hillsdale, NJ.

Raskin, D. C. (1982). The scientific basis of polygraph techniques and their uses in the judicial process. In Trankell, A., editor, *Reconstructing the past*, pages 317–371. Norsted & Soners, Stockholm, Sweden.

Raskin, D. C. (1986). The polygraph in 1986: Scientific, professional, and legal issues surrounding acceptance of polygraph evidence. *Utah Law Review*, 29:29–74.

Reid, J. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law, Criminology, and Police Science*, 37:542–547.

Rosenfeld, J. (2002). Event-related potential in the detection of deception, malingering, and false memories. In Kleiner, M., editor, *Handbook of polygraph testing*, pages 265–286. Academic Press, San Diego, CA.

Sapir, A. (2000). *The LSI course on scientific content analysis (SCAN).* Laboratory for Scientific Interrogation, Phoenix, AZ.

Sasaki, Y. (2007). The truth of the f-measure. *October*, pages 1–5.

Saxe, L. and Ben-Shakhar, G. (1999). Admissibility of polygraph tests: The application of scientific standards post- daubert . *Psychology, Public Policy, and Law*, 5(1):203 – 223.

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

Simpson, J. R. (2008). Functional mri lie detection: Too good to be true? *Journal of the American Academy of Psychiatry and the Law*, 36(4):491–498.

Smirnov, V. (1988). Internal and external logic. *Bulletin of the Section of Logic*, 17(3):170–181.

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

Spence, S. A. (2008). Playing devil's advocate: The case against fmri lie detection. *Legal and Criminological Psychology*, 13(1):11–25.

Sporer, S. and Schwandt, B. (2006a). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law*, 13:1–34.

Sporer, S. and Schwandt, B. (2006b). Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology*, 20:421–446.

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

Steller, M. (1989). Recent developments in statement analysis. In Yuille, J., editor, *Credibility Assessment*, pages 135–154. Kluwer, Deventer, The Netherlands.

Stern, P. (2003). The polygraph and lie detection. In *Report of the National Research Council Committee to Review the Scientific Evidence on the Polygraph*, pages 340–357. The National Academies Press, Washington, DC.

Strapparava, C. and Mihalcea, R. (2009). The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In *Proceeding ACLShort '09 - Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*.

Svartvik, J. (1968). The evans statements - a case for forensic linguistics. *Gothenburg studies in English*, 20.

Trankell, A. (1963). *Vittnespsykologins Arbetsmetoder*. Liber.

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

Undeutsch, U. (1984). Courtroom evaluation of eyewitness testimony. *Applied Psychology*, 33(1):51–66.

Vaassen, F. and 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. John Wiley & Sons, 2nd edition.

Vrij, A., Edward, K., Roberts, K., and Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior*, 24:239–263.

Vrij, A., Winkel, F., and Akehurst, L. (1997). Police officers' incorrect beliefs about nonverbal indicators of deception and its consequences. In Nijboer, J. and Reijntjes, J., editors, *Proceedings of the first world conference on new trends in criminal investigation and evidence*, pages 221–238. Koninklijke Vermande, Lelystad, the Netherlands.

Vrji, A. (2005). Criteria-based content analysis - A Qualitative Review of the First 37 Studies. *Psychology, Public Policy, and Law*, 11(1):3–41.

Walczyk, J. J., Roper, K. S., Seemann, E., and Humphrey, A. M. (2003). Cognitive mechanisms underlying lying to questions: response time as a cue to deception. *Applied Cognitive Psychology*, 17(7):755–774.

Wang, J. T.-y., Spezio, M., and Camerer, C. F. (2010). Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *The American Economic Review*, 100(3):984–1007.

Yang, Y. and 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, pages 42–49, New York, NY, USA. ACM.

Yang, Y. and Pedersen, J. O. (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., Burgoon, J. K., Twitchell, D., Qin, T., and Nunamaker, J. F. (2004). A comparison of classification models for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20:139–169.

Zhou, L., Shi, Y., and Zhang, D. (2008). A Statistical Language Modeling Approach to Online Deception Detection. *IEEE Transactions on Knowledge and Data Engineering*, 20(8):1077–1081.

Zuckerman, M., De Paulo, B., and Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In Berkowitz, L., editor, *Advances in experimental social psychology, volume 14*, pages 1–57. Academic Press, New York.