Lexical vs. Surface Features in Deceptive Language Analysis

Tommaso Fornaciari
Center for Mind/Brain Sciences
Corso Bettini 31, Rovereto
Università di Trento
tommaso.fornaciari@unitn.it

Massimo Poesio
Center for Mind/Brain Sciences
Università di Trento &
Language and Computation Group
University of Essex

ABSTRACT

Methods for identifying deceptive statements in language could be of great practical use in court and in other legal situations. Among the best known proposals in this direction are methods proposed by Pennebaker and colleagues relying on the Linguistic Inquiry and Word Count (LIWC). They used LIWC in different texts or transcriptions of spoken language, in which deception could have been used, but collected in an artificial way. We analyse the performance of these techniques to identify deceptions in genuine court testimonies from criminal proceedings for calumny and false testimony, in which deceptive statements were precisely identified in court judgments, and compare it with that of methods relying exclusively on surface information.

1. INTRODUCTION

Methods for identifying deceptive statements in language could be of great practical use in court and in other legal situations, e.g., to help the work of police forces, which face every day situations, where they have to evaluate questionable testimonies. Detecting deception isn't easy—humans find this task difficult, and their performance recognizing deception is not much better than chance [2]. Worse, it seems that specific training does not improve their skills [6].

Fortunately stylometric techniques have often been shown to be effective at picking up clues that identify aspects of a text or its author which humans can't spot. For example, the authors of anonymous text [7] or particular dimensions of personality [15]. In the case of detecting deception, the hope is to find clues in communication not under conscious control of the person producing the language that might reveal the deceptive character of a statement. The idea that "statements that are the product of experience will contain characteristics that are generally absent from statements that are the product of imagination" is historically known as Undeutsch Hypothesis [16]. In more formal terms, it could be asserted that, from a cognitive point of view, the elaboration of a false narrative is different from a simple memory recovery, so that some evidences of this difference could be found in the communicative outputs.

The major stumbling block in testing the Undeutsch hypothesis with computational methods is the scarcity of appropriate resources - i.e., of corpora of spoken or written language in which deceptive statements have been annotated. Such texts are not easy to come by, and as a result, most deception studies artificially produce language [9, 3, 14].

One of the key characteristics of the work discussed here is that we rely instead on real life data: the (Italian) Corpus of DEception in COURt (DECOUR), currently under construction and consisting of transcripts of criminal proceedings for calumny and false testimony in which the defendant was found guilty.

In the judgments issued in these trials, the events are carefully examined and the defendant's deceptive statements are explicitly listed: in fact, the judgments concern just these statements, and in 9 judgments on 14 the most relevant ones are reported verbatim. These circumstances mean that these criminal proceedings allow us to study deception in court with an unusually high degree of objectivity (keeping in mind of course that human error is always possible in court as well).

The work described in this paper had two objectives. First, we intended to evaluate the effectiveness of lexically-based techniques for deception detection—and in particular, the methods proposed in [9]—on our real-life data, which is in Italian, while these techniques have so far only been used for English, and with artificially produced data. Second, we intended to compare the same techniques with methods relying purely on surface features of the text.

The structure of the paper is as follows. We first discuss the lexical-based approach that we investigated, then our methods and the experimental setting used to compare techniques (included in this section are our sets of data), followed by results and a discussion.

2. BACK GROUND

2.1 Stylometry

Stylometry is the study of linguistic style in text, typically through statistical techniques. In forensic linguistics, typical stylometric tasks include author profiling [4, 12], author attribution [8, 7] and plagiarism analysis [13]; another well-

established type of stylometric analysis is deducing age and sex of authors of written texts [5].

As Koppel *et al.* (*op.cit.*) point out, the features used in stylometric analysis belong to two main families: surface-related and content-related features. The first type of feature includes the frequency and use of function words or of certain n-grams of words or part-of-speech. Such features have been shown to be surprisingly effective in work, e.g., by Daelemans and his lab [7]. The second kind of feature specifies information about the semantic content of words, accessed from dictionaries and lexical resources. Perhaps the best-known lexical resource for deception detection is the Linguistic Inquiry and Word Count (LIWC), created by Pennebaker [10] and used by his group for a number of studies on deceptive language [9]. In addition LIWC has been employed in studies on deceptive language carried out by other groups, such as Strapparava and Mihalcea [14], who obtained good results at classifying into "sincere" or "deceptive" texts collected with the Amazon Mechanical Turk service. Strapparava and Mihalcea used the LIWC for post-hoc analysis only, measuring several language dimensions, such as positive or negative emotions, self-references, and so on. In this way, they were able to identify some distinctive characteristics of deceptive texts, but only in descriptive terms: they did'nt make use of the LIWC outputs to distinguish the deceptive texts from sincere ones. Newman *et al.* [9], by contrast, used LIWC to carry out the classification itself.

LIWC also includes dictionaries of languages other than English, amongst which is Italian. We were therefore able to employ the categories of the Italian LIWC dictionary [1] as features to train models aimed at estimating if the statements of our Italian corpus were deceptive or sincere. Our corpus and our analysis units are different from the work of Newman *et al.*, but we followed an analogous methodological path.

2.2 Newman et al.

Newman *et al.* collected a corpus of sincere and deceptive texts through five different 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, in the fifth study, which was also videotaped, it 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 texts were given a preliminary analysis 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 following list of variables:

- 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.

So, when in one text the LIWC recognizes a word belonging to a category, for example "I" or "you" for the category "pronoun", or "no", "neither", "never" for the category "negation" and so on, the count of that category grows. Therefore, the output of the LIWC is a profile of the text, based on the presence of the different categories in it.

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 a 5-fold cross validation was performed, training a logistic regression on the texts of four studies and testing on the fifth. 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: an interesting starting point to demonstrate that the deceptive language is different from the sincere. 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%.

3. METHODS

In this work, we first aimed to adapt to Italian the deception detection methods proposed by Newman *et al.*; and secondly, to compare the results obtained in this way with those obtained using only surface features. We discuss each method in turn in this Section, and present the results in the next.

3.1 Adapting Newman et al.'s Techniques to Italian

In order to use the LIWC for deception detection, we collected, for each utterance, feature vectors based on the categories of the Italian LIWC dictionary. We did not directly employ the LIWC software for tokenization, preferring to make use instead of our tokenization rules. We simply counted out the correspondences in our corpus with the items of the Italian LIWC dictionary, incrementing the scale of the corresponding categories and then normalizing the frequencies so obtained.

We built five kinds of vectors, with the following features:

"Newman 29" First, for uniformity with the work of Newman *et al.*, we selected the features of the Italian LIWC dictionary corresponding to the categories of the English dictionary employed in the cited work. Due to the fact that the Italian categories for pronouns are larger than the English ones, the 29 categories of Newman *et al.* became 35. These categories are listed in Table 1.

"All" A second model featured all 85 categories of the Italian LIWC dictionary, plus the first three features of the "Newman 29" vector, that is the words counted and the percentage of words both longer than six letters[1] and captured by the dictionary, for an amount of 88 features.

"Our 29" Third, we selected the best 29 features on the basis of the *beta* weights of all variables, as obtained by the models trained with the "All" set of features. These were the LIWC variables with $beta > 1$. Table 2 shows the features and their weight.

"Newman 5" Then, a vector was built reproducing the 5 categories which Newman *et al.* employed to evaluate all their corpus simultaneously. Also in this case, to pass to the Italian categories implied to collect more categories, which is 10. The variables are shown in table 3.

[1]According to the choice of the Italian LIWC, we considered the words longer than six letters, regardless of possible differences between Italian and English language in the length of the words.

"Our 5" Last, we collected our five features with highest *beta* weights, that is:

English categories	Italian categories
Feeling	Sentim
You	Tu
Sleep	Dormire
Metaphysics	Metafis
Anxiety	Ansia

Table 1: The features of the "Newman 29" vector

English categories	Italian categories
Word Count	Word Count
% words captured by dic.	% words captured by dic.
% words > six letters	% words > six letters
Total pronouns	Pronomi
First-person singular	Io
	Io Ver
Total first person	Noi
	Noi Verb
Total third person	Lui lei
	Loro
	Se
	Lui Verb
	Loro Ver
Negations	Negazio
Articles	Articol
Prepositions	Prepos
Affective/emotional proc.	Affett
Positive emotions	Emo Pos
Negative emotions	Emo Neg
Cognitive processes	Mec Cog
Causation	Causa
Insight	Intros
Discrepancy	Discrep
Tentative	Inibiz
Certainty	Certez
Sensory/perceptual proc.	Proc Sen
Social processes	Social
Space	Spazio
Inclusive	Inclusi
Exclusive	Esclusi
Motion verbs	Movimen
Time	Tempo
Past tense verb	Passato
Present tense verb	Present
Future tense verb	Futuro

3.2 Surface strings

The surface features were extracted from a training set of 623 utterances, discussed below. First, we lemmatized and part-of-speech tagged these utterances, using a version of TreeTagger[2] [11] programmed for Italian. Then we considered the "true" and the "false" utterances separately, as two independent *corpora*.

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

Table 2: The features of the "Our 29" vector

English features	Italian features	*beta* weights
Feeling	Sentim	1820.3733
You	Tu	1031.9776
Sleep	Dormire	741.7434
Metaphysics	Metafis	674.1313
Anxiety	Ansia	195.5052
Leisure	Svago	64.6542
School	Scuola	30.2712
Affect	Affett	11.3263
He/She	Lui lei	10.8196
Body	Corpo	10.5046
Humans	Umano	10.1749
Down	Sotto	6.0316
Transitive	Transiti	5.8079
Achieve	Raggiun	5.3449
Conditional	Condizio	4.7417
Anger	Rabbia	4.0103
To be	Essere	3.6371
Space	Spazio	3.3979
You verb	Voi Verb	3.36
To have	Avere	2.8534
Senses	Proc Sen	2.4546
Dictionary	Dic	1.9339
Discrepancy	Discrep	1.7454
Social	Social	1.7434
Number	Numero	1.4363
We verb	Noi Verb	1.4317
Negate	Negazio	1.2563
Certainty	Certez	1.0186
Pronouns	pronomi	1.0133

For each set of utterances, we built six frequency lists, selecting their most frequent items, as follows:

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

So we collected 675 surface features for each class of utterances. Afterwards we merged the features of both sets. Therefore, theoretically, we could have had a vector of which the length could vary from 675 features, in case of perfect identity of the features of the two sets of utterances, to 1350 features, in case of no overlap. At the end, we obtained a vector of 1021 features, including two features not related to the frequency lists: the length of the utterances themselves, with or without punctuation.

Just the fact that there was not a lot of overlap between the most frequent surface features of "true" and "false" utterances, seemed promising about the possibility to distinguish the two classes.

Table 3: The features of the "Newman 5" vector

English categories	Italian categories
First-person singular	Io
	Io Ver
Total third person	Lui lei
	Loro
	Se
	Lui Verb
	Loro Ver
Negative emotions	Emo Neg
Exclusive	Esclusi
Motion verbs	Movimen

4. EXPERIMENTS

4.1 The Data

The data used in this work is the (Italian) Corpus of Deception in Court (DECOUR), a collection under construction of transcripts of criminal proceedings for "calumny" and "false testimony", in which the truthfulness or deceptiveness of testimonies is certain and easily verifiable, because when a defendant is found guilty, the trial ends with a judgment which explains the facts and points out the lies told by the subject, often *verbatim*.

At present, DECOUR is constituted of transcripts from 18 testimonies interrogating a total of 17 subjects and collected in the Italian Courts of Trento, Bolzano and Prato. The average age of the subjects was about 36; 14 of our subjects were male, 2 females, and 1 transgender; 8 subjects were from the North of Italy, 2 from the Center, 3 from the South, and 4 from abroad. Finally, we only knew the educational level of five subjects: in four cases they had high school qualifications, and in the last case Italian middle school.

Unlike the study of Newman *et al.*, our analysis units were not whole documents, but the single utterances issued by the subjects. We had 1437 utterances issued by the heard subjects, which appeared in the hearings as defendant, witness or expert witness. The utterances of other figures in the hearings, typically the judge, the prosecutor and the lawyer, were by default assumed as true and not considered in this work.

Each utterance of the subject being questioned received a label regarding the truthfulness or not of the utterance itself, on the basis of the information found in the judgment issued by the judge. Obviously, between the white of truth and the black of falsity there were several degrees of grey, and the judgment that describes the facts and points out the lies of the defendant, can't be used to label each statement issued in the courtroom. Therefore, we developed a coding scheme to take these issues into account. The labels used to mark utterances were chosen from amongst these categories:

"False" The utterance is clearly identified in the judgment as false, or its falsity is a logical consequence of some ascertained lie.

"True" The utterances that are consistent with the reconstruction of the facts contained in the judgment, are

considered true. Also the utterances that explain something not considered in the judgment, because uninfluential in respect to the investigated facts, are generally considered true.

"Not reliable" An utterance is considered not reliable if it is related to the facts under investigation, but the judgment does not prove its deceptiveness.

"True or not reliable" Like the "not reliable" utterances, the "true or not reliable" ones are related to the topic of investigation, and the judgment demonstrates nothing about them. The only difference–sometimes hard to make–is that, according to the event and to other statements for certain true or false, and/or on the basis of a weak connection with the interests that the subject tries to defend, it is logical to suppose that they are probably true. In brief, according to common sense those utterances should be true, but the fact is not demonstrated, and ultimately questionable.

"False or not reliable" This is the specular situation in respect to the previous point.

"Undecidable" The utterances that, from a logical point of view, cannot be either true or false, are considered undecidable. This is the case with a lot of questions (like "Excuse me, can you repeat?"), but also of several utterances stopped in mid-sentence, that haven't a complete sense. This is also the case of the utterances that have a meta-communicative function, and regulate the relations between actors, like "Now I'll explain." or "If you think so..." and so on.

The amount of labeled utterances and of their tokens (with and without punctuation) is shown in the following table.

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

Only the utterances labeled as "true" and "false" were used in our study, and the other ones discarded. We obtained therefore a corpus of 870 utterances, of which about 61.7% were "true" and 38.3% are "false".

4.2 The logistic regression

To carry out the analyses, the corpus of 870 "true" or "false" utterances was split in this way:

- 10 hearings were used as a training set, for a total of 623 utterances: meaning about 72% of the corpus, in terms of utterances. It is also the part of the corpus from which we collected the features of the surface vectors;
- 4 hearings were used as a test set, for a total of 148 utterances, equal to 17% of the corpus.
- 4 hearings were used as a development set, for error analysis and so on.

Using the training set mentioned above, we built models performing logistic regression in the Weka package[3]. We employed separately the vectors made by the content features of the Italian LIWC dictionary (*op.cit.*), and the vectors of surface features collected from the training set. The test set was employed for the classification task.

4.3 Chance levels

To evaluate the results of the analyses, we defined our chance level through a Monte Carlo simulation. The test set had 81 "true" utterances and 67 "false", which means respectively 54.73% and 45.27%. 10000 times, a random simulator simply produced 148 previsions, obtaining the result "true" with p = .5473.
Comparing the simulated results with the test set, we found that less than 1% of simulations exceeded the 60% of "correct answers". So we assumed the 60% of correct classifications as the threshold for our test set.

5. RESULTS

5.1 The content feature vectors

The results of the experiments with content feature vectors are shown in Tables 4, 5, 6, 7 and 8. The performances of "Newman 29" and "All" vectors are similar and clearly higher than chance level. The "Our 29" features also did better than chance level, but the results were not as good. "Newman 5" and "Our 5" vectors, instead, did not exceed the chance level. In other words, the feature selection techniques we used did not seem to be very useful—in general, the more features were employed in the vectors, the better the results.

Always, the fluctuations in performance are due to different levels of effectiveness in detecting deceptive utterances. "Newman 29", "All" and "Our 29" vectors, indeed, have exactly the same accuracy detecting "true" utterances. But the worst models are increasingly blind to deceptiveness, and tend to evaluate all utterances as "true": the "Our 5" vector, for example, judges "true" 146 of the 148 utterances in the test set. Also for this reason, the recall of "true" utterances is always high. The crucial challenge, therefore, is to discover the "false" utterances: the recall of the best vectors is a little less than .5, up to about 0 for the worst ones.

However, the best vectors reach high levels of precision in detecting deception, close to .9. This means that, if on the one hand it is not simple to recognize deceptive utterances, on the other one, when models judge an utterance as deceptive they are unlikely to be wrong. The same precision is not found regarding the "true" utterances: it is due to the tendency of the models to see "true" utterances, with advantage for the recall, and disadvantage for the precision.

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

Table 4: Logistic regression - "Newman 29" vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	33	34	0.868	0.493	0.629
True utterances	76	5	0.691	0.938	0.796
Total	109	39			
Total per cent	**73.65**%	**26.35**%			

Table 5: Logistic regression - "All" vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	32	35	0.865	0.478	0.615
True utterances	76	5	0.685	0.938	0.792
Total	108	40			
Total per cent	**72.97**%	**27.03**%			

Table 6: Logistic regression - "Our 29" vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	21	46	0.808	0.313	0.452
True utterances	76	5	0.623	0.938	0.749
Total	97	51			
Total per cent	**65.54**%	**34.46**%			

Table 7: Logistic regression - "Newman 5" vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	4	63	0.5	0.06	0.107
True utterances	77	4	0.55	0.951	0.697
Total	81	67			
Total per cent	**54.73**%	**45.27**%			

Table 8: Logistic regression - "Our 5" vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	1	66	0.5	0.015	0.029
True utterances	80	1	0.548	0.988	0.705
Total	81	67			
Total per cent	**54.73**%	**45.27**%			

Table 9: Logistic regression - Surface feature vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	35	32	0.729	0.522	0.609
True utterances	68	13	0.68	0.84	0.751
Total	103	45			
Total per cent	**69.59**%	**30.41**%			

5.2 The surface feature vectors

The results of the experiments with surface feature vectors are shown in Table 9. The model trained with surface features also achieves results well above chance level—indeed, almost as good as those with the best content feature vectors. This difference is mainly due to "false positive" errors: more utterances are classified as "false" even if they are not. In fact, in this model the precision in detecting "false" utterances is lower (although consequently the recall is slightly better).

6. DISCUSSION

Even if the 29 features of Newman *et al.* were selected for English texts, they are very effective with Italian testimonies, as well. The "Newman 29" vector is the best, but performs better than "all" only because it classified well a single "false" utterance more than the other one: so their results are largely equivalent. This confirms Newman *et al.*'s hypothesis that to exclude from the vectors the features related to the topic of the texts does not result in worse performance. This also suggests that, if typical features of deceptive language exist, they should not be found in the topic of the speech. Moreover, the "Our 29" vector, which collects the most weighty features of the "All" vector, contains six features clearly content-related (Sleep, Metaphysics, Leisure, School, Body, Humans), and their performances are inferior to "All" and "Newman 29" vectors. It is therefore possible that to exclude selectively some content-related features is damaging. It creates imbalances in evaluating specific topics since vectors, which include all or none of content-related features, perform better and very similar to each other.

Unlike what reported by Newman *et al.*, the smaller feature sets do not perform well in our corpus. This is probably due to the fact that our analysis units - the utterances - are considerably shorter than the texts of their study, and therefore they need to be defined by a lot of features, to be adequately identified. As suggested by one of our reviewers, it would also be interesting to know the performances of the small feature sets on texts longer than the sentences of our corpus. It is possible that the results would be better, even if evaluating the deceptiveness the focus is inevitably on each atomic analysis unit, able to provide a single, complete point of information in the communication.

Our results show that using LIWC does in fact result in slightly better performance than when using surface features alone, but not by much, which suggests that reasonable results at deception detection could be obtained with resource-poor languages as well. On the other hand, experiments in progress combining both content and surface features suggest that this combination may result in improved performance.

Our results could also show that our subjects did spend some effort to conceal their lies. In the Monte Carlo simulation, less than 1% of the simulation had a recall of "true" utterances better than 63%. Our models based on the Italian LIWC dictionary categories, instead, show a clear bias, so that they tend to judge as "true" a lot of utterances, and their recall is never lower than 93.8%... at the expense of the recall of "false" utterances. This means that several false utterances are extremely similar to the true ones. It is possible that this is simply due to the structure of the answers of the subjects to the question posed, but of course it fits with their interest to hide the lies.

The good news is that when an utterance is recognized as "false", the models trained with content features are probably right. It would be crucial in a real life scenario, where it would be very important to be confident about the previsions carried out. This could be a practical reason why the content features seem to be better than the surface ones, regardless of their overall accuracy.

The moral could be that, in the context of the hearings in front of the judge, there are "false" utterances that are linguistically similar - or identical - to the "true" ones. Maybe they can not be recognized with tools of textual statistics, but there is also a portion of "false" utterances - maybe about 50%, like our results suggest? - which are different in style from the "true" ones. We hope that this portion can be used to support and to orientate police investigations and judges's decisions, especially in cases in which other kinds of evidence are scarce or absent.

Last but not least, the lack of topic-dependence of the features we identified so far, and the fact that our results confirm Newman *et al.*'s results at detecting deception in other types of language, suggest that the approach adopted here could find application in detecting deception in other types of text as well.

7. ACKNOWLEDGMENTS

The data collection involved in the creation of DECOUR is a complex task that could not have been done without the help of a lot of people. Many thanks to Dr. Heinrich Zanon, President of the Court of Bolzano; to Dr. Sabino Giarrusso, President of the Court of Trento; and to Dr. Francesco Antonio Genovese, President of the Court of Prato. Many thanks also to Dr. Piero Tony, Chief Prosecutor of the Public Prosecutor's Office of Prato, to Dr. Sandro Pettinato of the Court of Trento, to Dr. Biagio Mazzeo, Prosecutor in the Public Prosecutor's Office of Genova, to Dr. Michela Guidi, Prosecutor in the Public Prosecutor's Office of Prato, and to Rita Fava of the Public Prosecutor's Office of Prato.

8. REFERENCES

[1] A. Agosti and A. Rellini. The italian liwc dictionary. Technical report, LIWC.net, Austin, TX, 2007.

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

[3] J. Burgoon, J. Blair, T. Qin, and J. Nunamaker. Detecting deception through linguistic analysis. In H. Chen, R. Miranda, D. Zeng, C. Demchak, J. Schroeder, and T. Madhusudan, editors, *Intelligence and Security Informatics*, pages 958–958. Springer Berlin / Heidelberg, 2003.

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

[5] M. Koppel, J. Schler, S. Argamon, and J. Pennebaker. Effects of age and gender on blogging. In *In AAAI*

2006 Spring Symposium on Computational Approaches to Analysing Weblogs, 2006.

[6] T. R. Levine, T. H. Feeley, S. A. McCornack, M. Hughes, and C. M. Harms. 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, 2005.

[7] K. Luyckx and W. Daelemans. 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, 2008. Association for Computational Linguistics.

[8] F. Mosteller and D. Wallace. *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, 1964.

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

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

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

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

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

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

[15] S. A. Sushant, S. Argamon, S. Dhawle, and J. W. Pennebaker. Lexical predictors of personality type. In *In Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

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