PROCEEDINGS

3rd European Conference of the International Association of Forensic Linguists

Bridging the Gap(s) between language and the law

Edited by
Rui Sousa-Silva
Rita Faria
Núria Gavaldà
Belinda Maia

Bridging the Gap(s) between Language and the Law

Proceedings of the
3rd European Conference of the
International Association of Forensic Linguists

Contents

Editors' introduction

In October 2012, the Faculty of Arts and the Faculty of Law of the University of Porto jointly organised the 3rd European Conference of the International Association of Forensic Linguists on the theme of **Bridging the Gap(s) between Language and the Law**. The event took place in the Faculty of Law. Although the adjective 'European' in the title suggests a regional conference, the participation of delegates from over 20 countries demonstrates how international the conference was. The range of topics in the field of Forensic Linguistics / Language and the Law was equally diverse. The high number of papers presented by academics, lawyers and practitioners alike, not only from the field of linguistics, but also from the fields of law, criminology, engineering and natural language processing, demonstrates the multidisciplinary nature of this conference, which is also evident in the background of the conference Local Organising team. This was a successful accomplishment of one of the main objectives of this conference: bridging the gap(s) between language and the law. For the first time in the history of the International Association of Forensic Linguists, a meeting of the Association included a strand in Portuguese, although English was, as usual, the most widely used language in the conference.

We hope that this collection of articles does justice to the conference and adds to the increasing body of literature and helps establish how diverse the research into Forensic Linguistics / Language and the Law can be. This volume is not formally divided into sections or parts according to the themes discussed. However, readers will be able to identify the main themes approached: forensic authorship analysis and plagiarism; legal translation, interpreting and multilingualism; and interaction between language and the law. The book opens with the **Presidential address** of Maria Teresa Turell, in which the then President of the IAFL discusses the role(s) of the Association and the direction(s) of Forensic Linguistics. The volume then continues with two articles on computer-assisted authorship attribution and plagiarism detection. First, in **A semi-automatic authorship attribution technique applied to real forensic cases involving Judgments in Spanish**, Sheila Queralt Estevez and Maria Teresa Turell present the result of their analysis of real forensic data for authorship

attribution. Maria Teresa Turell and Paolo Rosso, in **Computational approaches to plagiarism detection and authorship attribution in real forensic cases**, show how computational linguistics applied to real data can assist both authorship attribution and plagiarism detection. Next Andrea Nini explores the applications of **Codal variation theory as a forensic tool**. In his paper, he demonstrates how the theoretical framework of codal variation described by Systemic Functional Linguistics (SFL) can contribute to authorship analysis. David García-Barrero, Manuel Feria and Maria Teresa Turell subsequently discuss the results and implications of **Using function words and punctuation marks in Arabic forensic authorship attribution**. The two articles that follow add to the set of papers on authorship analysis and attribution. In the first paper, Firuza Bayramova presents her view of the **Complex Research on speech interference characteristics**. Then, in **Forensic Linguistics accredited: Four years of experiences with ISO 17020 in authorship analysis**, Sabine Ehrhardt initiates the much needed discussion of the issues related to accreditation in the field of Forensic Linguistics. The experience of the four years described will certainly encourage new directions in Forensic Linguistics.

The next set of articles approaches the issues of multilingualism, as well as translation and interpreting, not only in contemporary European Union, but also in the United States. Firstly, Belinda Maia investigates the EU Directive on the right to translation and interpreting in the EU. She explores **The problems and practicalities of training translators and interpreters for the future envisaged by the EU Directive 2010/64/EU** to unveil some of the less evident problems underlying the application of the EU Directive. Sandra Silva contributes to this discussion from a legal perspective, by approaching **The right to interpretation and translation in Criminal Proceedings: the situation in Portugal**. Despite concluding that Portuguese law meets the minimum standards outlined in the Directive, she makes some recommendations to remedy the gaps identified and improve the effective application of the right to interpreting and translation. Thirdly, Karolina Paluszek presents the results of an analysis of several cases taken to the European Court of Justice to discuss **Multilingualism and certainty of law in European Union**. Lastly, in **And Justice for All: Non-Native Speakers in the American Legal System**, Leah M. Nodar discusses the problems faced by non-native speakers of English in the United States when dealing with the legal system to challenge the principle that all have an equal right to justice.

The final set of articles includes a diverse range of contributions. Miriam Jiménez Bernal first approaches the topic of **Fear as a key element in deceptive and threatening narratives** by presenting the results of her analysis of real data. Then, in **Investigating legal language peculiarities across different types of Italian legal texts: an NLP-based approach**, Giulia Venturi reports on her research applied to Italian to explain how a natural language processing approach can contribute to identify and investigate legal language peculiarities across different types of legal texts. In **The Linguistic Functions of 'Knowingly' and 'Intelligently' in Police Cautions**, Margaret van Naerssen then explores the meanings of 'knowingly' and 'intelligently' to demonstrate that a linguistic analysis can not only help determine whether civil rights of police interviewees are being respected, but also that there is room for improvement of acceptability of linguistic evidence. This set of articles and the volume end with

Language and Law: ways to bridge the gap(s). In this article inspired in the conference main theme, Virgínia Colares discusses some of the issues faced by linguistics in legal contexts and makes suggestions of how the existing gaps can be bridged.

Of course, this collection of diverse and inspiring articles would not have been possible without the contribution of all the authors, whom the editors would like to thank. We hope that the readers will enjoy reading these articles as much as we did. Secondly, we would like to thank the Scientific Committee, who worked so hard to select the best papers submitted in the first place to this conference. The conference would not have been such a success without their work and support. We would also like to thank the reviewers, to whom we owe the selection of the articles to be published in this volume. Their detailed comments and suggestions to authors significantly contributed to the work presented in this book. Last, but not least, we would like to dedicate these proceedings to Professor Maria Teresa Turell, who – sadly – was unable to join us and attend the conference in Porto, but who supported us from the very beginning. Her support was priceless.

Rui Sousa-Silva

Rita Faria

Núria Gavaldá

Belinda Maia

Porto, December 2013

Presidential address

M. Teresa Turell Julià [*]

As president of the International Association of Forensic Linguists, it gives me a great pleasure to address you in the opening session of the second regional IAFL conference taking place this year. The other regional conference, also sponsored by IAFL, took place in Kuala Lumpur (Malaysia) last July.

I'm very grateful to Ed Finegan, our IAFL vice-president, for reading my presidential address on my behalf.

This regional conference entitled "Forensic Linguistics: Bridging the Gap(s) between Language and the Law" is jointly hosted by the Faculty of Arts and the Faculty of Law at the University of Porto here in Portugal and is co-organised by scholars from several centers at this same institution: *Centro de Linguística da Universidadedo Porto* (CLUP) and *Laboratório de Inteligência Artificial e Ciência de Computadores* (LIACC).

With fond memories from the last IAFL biennial conference organized by our Aston colleagues a year ago and the very succesful Kuala Lumpur conference last July, I'm sure that this regional conference will be a great success, thanks to the local organizing committee led by Belinda Maia and Rui Sousa-Silva. I would like to thank all colleagues involved in the organization of this event for their efficiency and "savoir faire". My heartfelt thanks also go to the plenary speakers invited to this conference, for contributing to it with substantial topics and areas of expertise, and to all participants for choosing this conference to disseminate their research results in the many areas of study around which our discipline is structured.

The seventy-one papers, two round-tables, and the colloquium on "Making Linguistics Relevant to the Law" open for discussion at this conference indeed reflect the multidisciplinary nature and the multi-dimensionality of Forensic Linguistics. We are very pleased to welcome participants from all five continents and from nineteen countries around the world: Australia, Austria, Belgium, Brazil, Check Republic, China, Germany, Cyprus, Italy, Japan, the Netherlands, Poland, Portugal, Romania, Russia, Spain, South Africa, the UK and the US.

---

[*]President of the IAFL – International Association of Forensic Linguists

As a representative of the IAFL, I would like to say a few words about our discipline, Forensic Linguistics/Language and Law, and our association.

I usually refer to our discipline as Forensic Linguistics in its "broadest" sense, and this means covering all areas where law and language overlap: Language and the Law, Language and the Legal Process, and Language as Evidence, both in criminal cases (authorship analysis and attribution, forensic speaker identification/voice comparison) and civil cases (contract disputes, defamation, product liability, consumer product warnings, deceptive trade practices, copyright infringement and plagiarism detection, and trademark litigation).

By referring to "Forensic Linguistics", I purposely emphasize the term Forensic, which stands for "used in Court of law or public discussion and debate" but, most importantly, derives from "Forum", whose third entry in *Webster's Encyclopaedic Unabridged Dictionary of the English Language* reads as "an assembly for the discussion of questions of public interest", that is, in our case all questions that refer to the interface between language and law. And I also emphasize Linguistics because the term refers to the scientific study of language, the task in which we are all engaged.

For those of you who are not familiar with the development of our discipline, suffice to say that, although it was born in the fifties, it was not until the nineties of the twentieth century that Forensic Linguistics emerged very forcefully: the experts' performance became much more professionalized; there was an outstanding increase in the publication of articles and chapters in a number of forensic linguistics topics, whose content was much more methodologically grounded than before; the *International Association of Forensic Phonetics* (now named the *International Association of Forensic Phonetics and Acoustics*) was founded in 1991 and the *International Association of Forensic Linguists* in 1992; a Forensic Linguistics journal was created in 1994, which had different publishers and whose title underwent several changes, its current name being *The International Journal of Speech Language and the Law*. I take this opportunity to thank former and current editors of the journal for carrying out their task rigorously and efficiently.

The present state of the discipline reflects consolidation and vitality. With the turn of the century, Forensic Linguistics has grown to see *The International Journal of Speech Language and the Law* reaching its 19th year of publication and many FL articles appearing in law journals; IJSLL has become one of the facilitators of the international development of Forensic Linguistics, and it can also be said that a vital spark for the discipline is the interdisciplinary stand that this journal has taken. Another sign of international vitality is the number of international and regional conferences that have been celebrated: our 10th biennal IAFL conference took place in Aston in July 2011 and our next biennial conference will be held in Mexico City in June 2013; our discipline has also grown to see the emergence of forensic linguistics laboratories and centres, government-funded police laboratories and agencies around the world; a state of the art in FL literature with volumes on particular topics, edited volumes, introductions to the field, textbooks and handbooks; and finally FL graduate courses and specialist Master's degree courses in several universities around the world.

All this shows that in many countries around the world Forensic Linguistics is now a well-defined, well-established discipline, which seeks to uncover and establish the existing interplay between linguistics and legal issues.

We have a legacy from the past that involves in most cases quality, rigour, validity and reliability in our analysis and opinions. The discipline faces several challenges, however: the integrated study of forensic linguistics/language and law across different judicial systems and geographical boundaries; the development of replicable methods of analysis to be used in expert witness evidence in order to ensure internal and external validity in research; the extensive setting up of codes of good practice and conduct; and also our presence in associations and societies of Forensic Sciences.

When I started my mandate in 2011, IAFL was facing several challenges: a centralised enrolment and payment system made operational through a new website, with a new design and updated content, and the activation of the already existing IAFL-MEMBERS@JISCMAIL.AC.UK list, in order to foster and promote internal debate within the association. I'm very plased to be able to say that some of these challenges are *now achievements on their way*, thanks to the very efficient work of all members of the IAFL Executive Committee without exception: Ron Butters as past president, Ed Finegan as vice-president, Georgina Heydon as secretary, Phil Gaines as treasurer, and our three active additional members, Krzysztof Kredens as publicity officer, and Peter Gray and Azirah Hassim as ordinary members:

a) The design and content of our new website is almost ready, thanks to work done by the steering committee (Ed Finegan, Krzysztof Kredens and myself) that I appointed before the summer. This new IAFL website will include a centralised enrolment and payment system, with a two-tiered option: regular and reduced (student and retired members) and a donation mechanism, a members' directory, and other options. Meanwhile, we invite scholars to join the IAFL or renew their membership through the existing provisional system of enrolment and payment.

b) The IAFL-MEMBERS@JISCMAIL.AC.UK list, through which the EC will encourage the electronic vote of the membership on several issues between general business meetings, is now active.

c) An IAFL Code of practice draft, which has been put together by a committee chaired by Ron Butters, will soon be posted up on our web for members to make suggestions and then have an electronic vote through the IAFL members list.

Finally, I'd like to say that we – IAFL officers and members – are in fact taking advantage of the legacy handed down by former presidents and previous executive committees in order to a) promote the celebration of regional conferences like this one in Porto; b) reflect the multifaceted nature of Forensic Linguistics and of IAFL and develop all areas of work and study; c) encourage rigorous research and work across the different judicial systems around the world; and d) extend membership to other continents and countries where IAFL is not yet present.

On behalf of the IAFL Executive Committee and on my own behalf, let me pronounce my words of friendship and support to the organisers of this conference, and extend my best wishes for a very successful conference to all participants.

M. Teresa Turell

President, International Association of Forensic Linguists

Barcelona, October 11, 2012

A semi-automatic authorship attribution technique applied to real forensic cases involving Judgments in Spanish

Sheila Queralt Estevez and M. Teresa Turell Julià [*]

Abstract. *Recent studies in forensic authorship attribution report on newly developed techniques, although it seems that, in this quest for valid and reliable identification markers, syntactic structure has shown to be less appealing, which is easily explained by the fact that syntactic variables are clearly more complex, more difficult to process and less frequent than variables at other linguistic levels. This paper presents a series of experiments in forensic authorship, whose aim is to evaluate the discriminatory potential of sequences of linguistic categories (POS n-grams) by using real forensic legal texts written in Ecuador Spanish. The hypotheses tested in these experiments are that a) the most frequent tag sequences will discriminate effectively between authors and b) both bigrams and trigrams will both show this discriminatory capacity. Experiments were carried out by making use of two morpho-syntactically annotated corpora from two real forensic cases, consisting of disputed Judgments (the N for case A = 1; the N for case B = 1) and non-disputed texts (Judgments and other legal texts) from two author candidates in each respective case, with five non-disputed texts for each candidate author in each case. In both cases, a control corpus of non-disputed texts was used, with three authors and five texts per author, and a total of fifteen non-disputed texts. All texts were analysed both qualitatively and quantitavely, in the latter case by running Linear Discriminant Analysis. Preliminary results confirm the hypotheses for both bigrams and trigrams in each case.*

Keywords: N-grams, forensic written text comparison, authorship attribution, Spanish, forensic real cases.

Introduction

Language reflects a series of linguistic traits that can be used in authorship attribution contexts. So far there is not a single method or technique that can be used in forensic

---

[*]IULA, Universitat Pompeu Fabra

analyses or in expert witness consulting. In order to understand the status of authorship attribution it is very important to bear in mind the complementary nature of forensic linguistic evidence very much concerned with methodological reliability.

The cumulative evidence considered in forensic authorship attribution has involved the use of several methods and techniques such as the use of reference corpora, type-token ratios, hapax legomena (de Vel *et al.*, 2001), vocabulary analysis (Hoey, 2005; Turell, 2004a,b; Woolls and Coulthard, 2007), sequences of linguistic categories, also called Part-Of-Speech n-grams, in forensic analyses (Bel *et al.*, 2012; Queralt *et al.*, 2011; Queralt and Turell, 2012; Spassova and Turell, 2007; Spassova and Grant, 2008; Spassova, 2009; Turell, 2010), among others.

The forensic linguist's role in forensic text comparison is to observe those linguistic variables and data which might be decisive for determining, among several candidates, who the author of a particular spoken or written text is, which is the research question addressed in the cases presented in this paper. Forensic linguists should base their analyses on valid and reliable methods and techniques by a) undertaking experimental research on real world texts, outside case work, b) applying the same techniques to real forensic case texts, c) using statistical analyses to establish the significance of results, and d) making use of corpus linguistics, and many other approaches, both qualitative and quantitative. Finally, forensic linguists should make this information much more comprehensible to the judge and court.

Recent studies in forensic authorship attribution report on newly developed techniques, although it seems that, in this quest for valid and reliable identification markers, syntactic structure has shown to be less appealing, which is easily explained by the fact that syntactic variables are clearly more complex, more difficult to process and less frequent than variables at other linguistic levels.

This paper presents a series of experiments in authorship attribution, whose aim is to evaluate the discriminatory potential of sequences of linguistic categories (POS n-grams) by using real forensic legal texts written in Ecuador Spanish.

Aim

The theoretical aim of this paper is to show the usefulness of the concept "idiolectal style" (Turell, 2010) in forensic text comparison in its application to the study of non-discrete variables such as sequences of linguistic categories. The methodological aim behind the analysis of two sets of real forensic texts considered in this paper is to establish among several candidates who the author of a written disputed text is in order to help the Court in their decision. The protocol established by our lab, once determined that there is enough linguistic evidence to proceed with the analysis, involves a qualitative and a quantitative approach. In this paper we only focus on the quantitative approach by using the sequences of linguistic categories and its subsequent analysis to establish the statistical significance of the results and to ensure research validity and reliability.

The variable

The variable for our study has to do with sequences of linguistic categories. In previous publications (Bel *et al.*, 2012; Queralt *et al.*, 2011; Queralt and Turell, 2012; Spassova and Turell, 2007; Spassova and Grant, 2008; Spassova, 2009; Turell, 2010) the term Morpho-syntactic Annotated Tag Sequences – MATS was used, but in order to be coherent with the current literature the term Part-of-Speech n-grams (POS-ngrams) is adopted.

Figure 1. Sequences of linguistic categories (POS n-gram).
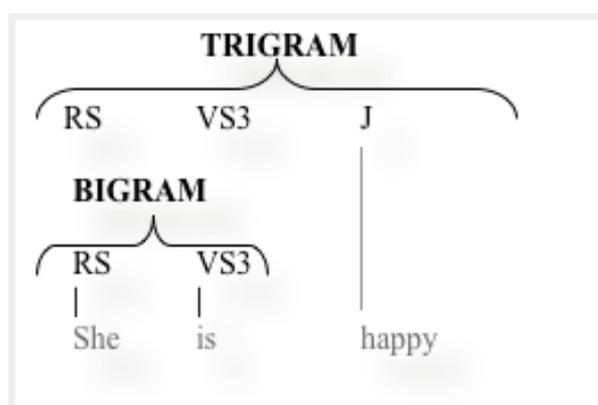
The sentence: "ella es feliz siempre" (She is happy always) is tagged in Figure 1. Two examples of POS n-grams can be observed, the bigram RS-VS3 and the trigram RS-VS3-JQ where RS stands for singular pronoun, VS3 for third person singular verb and JQ, for qualifying adjective.

Hypotheses

Results reported in previous experiments with real world texts and the application of the Linear Discriminant Analysis (LDA) methods showed the efficient discriminatory potential of POS n-grams. Thus, the hypotheses formulated for both cases are that most frequent tag sequences will discriminate effectively between authors and that bigrams and trigrams will both show this discriminatory capacity, more effectively than other sequences.

The method

This method involves several phases: firstly, a pre-processing phase, in which texts are revised for misspellings or any other possible errors. Secondly, a morpho-syntactic tagging phase, during which the text is converted into a row of token types and tags. Thirdly, a disambiguation stage, through which texts are disambiguated and errors are corrected. Fourthly, a tag extraction phase – during which the information obtained, refers to the number of POS n-grams types and tokens and on to POS n-grams frequency values to be used in the subsequent statistical analysis. And finally, once the tags have been extracted, a last stage involves the application of LDA, in order to have the different text sets classified by author, and have the results projected onto graphs. LDA could be defined as a multivariate statistical technique with three main purposes:

a) To describe whether the use of the n-grams under analysis (bigrams or trigrams) is statistically significant.

b) To determine which are the n-grams that exhibit the highest potential to discriminate between different authors.

c) To predict group membership when we have an unknown text. For example we may have an anonymous text and we want to know who is the most probable author to have produced this text with regards to n-grams use. In order to predict group membership, this technique creates a discriminant function that is the result of a combination of the

n-grams weighted to maximize the difference between the idiolectal style of several authors.

Cases and results

The cases reported in this paper implied the consideration of a Judgment whose authorship was being questioned and two possible author candidates in each case. Thus, the aim of the work undertaken in this analysis was to help the court decide whether the written style observed in the disputed Judgment showed linguistic similarities with the non-disputed Judgments of Candidate 1 or Candidate 2 for each respective case.

Case A

Corpus

As Table 1 summarises, the data for this case consist of a disputed judgment that was divided into 5 excerpts and 5 different judgements of two possible male authors (Candidate 1 and Candidate 2), written in Ecuador Spanish. For accountability and reliability purposes a corpus of three sets of five anonymous Judgements from three judges used in another case was considered, with a similar length and textual structure.

Table 1. Corpus Case A.

Writers	Gender	Genre	Text information 1600 words
Candidate 1	M	Judgment	5
Candidate 2	M	Judgment	5
Control 1	M	Judgment	5
Control 2	M	Judgment	5
Control 3	M	Judgment	5
Disputed text	M	Judgment	5

Results

Bigrams

Figure 2 shows the LDA projection for bigrams applied to the corpus of Case A. In this figure it can be observed that the disputed excerpts are located in the same area of Candidate 1, while the centroid of Candidate 2 occupies a different side of the graph.

Table 2 shows that the LDA classification method successfully classified 100% of the texts by authors within their own group, while the cross-validation method confirmed that the analysis was 96% correct. The five excerpts of the disputed text were attributed to Candidate 1.

Trigrams

Figure 3 shows the LDA projection for trigrams. This figure illustrates that the excerpts of the disputed text are close to the centroid of Candidate 1 while Candidate 2 is placed far from the disputed texts set.

Figure 2. Linear Discriminant Function Analysis based on bigrams – Case A.
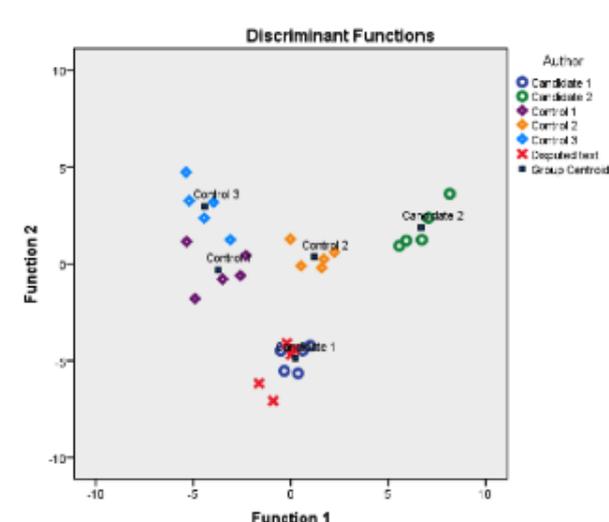
Table 2. Classification and cross-validation results for bigrams – Case A

			Predicted Group Membership					
		Autor	Candidate 1	Candidate 2	Control 1	Control 2	Control 3	Total
Original	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	5	0	0	0	5
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5
		Disputed	5	0	0	0	0	5
Cross-validation	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	5	0	0	0	5
		Control 1	0	0	5	0	0	5
		Control 2	0	0	1	4	0	5
		Control 3	0	0	0	0	5	5

100,0% of original grouped cases correctly classified.

96,0% of cross-validated grouped cases correctly classified.

As shown in Table 4, 100% of the texts were classified successfully while the cross-validation confirmed that the analysis was 92.0% correct. Trigrams classified four excerpts of the disputed text to Candidate 1 and one excerpt to one of the control authors.

The interpretation of these statistical results using LDA for POS n-grams revealed that there was a quite high probability that the author of the disputed Judgment was Candidate 1.

Case B

Corpus

As Table 4 summarises, the corpus for this second case consisted of a disputed judgment fragmented into five excerpts from two possible male authors (Candidate 1 and Candidate 2), written also in Ecuador Spanish. In order to optimize the discriminatory potential of the variable under analysis, a set of anonymous Judgements from another case was also used.

Figure 3. Linear Discriminant Function Analysis based on trigrams – Case A.
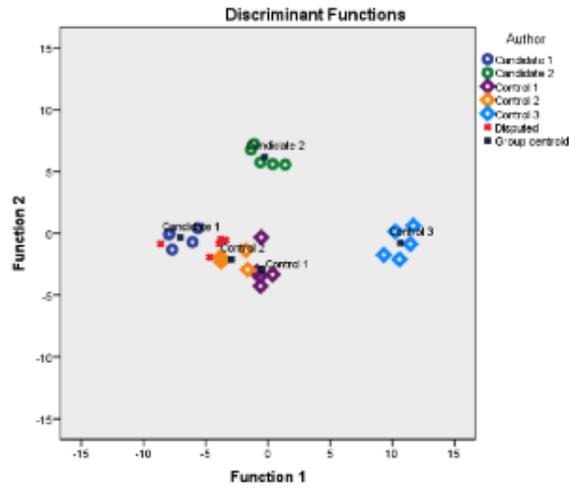
Table 3. Classification and cross-validation results for trigrams – Case A

		Autor	Predicted Group Membership					Total
			Candidate 1	Candidate 2	Control 1	Control 2	Control 3	
Original	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	5	0	0	0	5
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5
		Disputed	5	0	0	0	0	5
Cross-validation	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	5	0	0	0	5
		Control 1	0	0	5	0	0	5
		Control 2	0	0	1	4	0	5
		Control 3	0	0	0	0	5	5

100,0% of original grouped cases correctly classified.

96,0% of cross-validated grouped cases correctly classified.

Table 4. Corpus Case B.

Writers	Gender	Genre	Text infor-mation 800 words
Candidate 1	M	Judgment	5
Candidate 2	M	Judgment	6
Control 1	M	Judgment	5
Control 2	M	Judgment	5
Control 3	M	Judgment	5
Disputed text	M	Judgment	5

Results

Bigrams

Figure 4 presents results for bigrams of the LDA applied to Case B text sets. This figure shows that the disputed excerpts are closer to Candidate 2 than to Candidate 1.

Figure 4. Linear Discriminant Function Analysis based on bigrams – Case B.
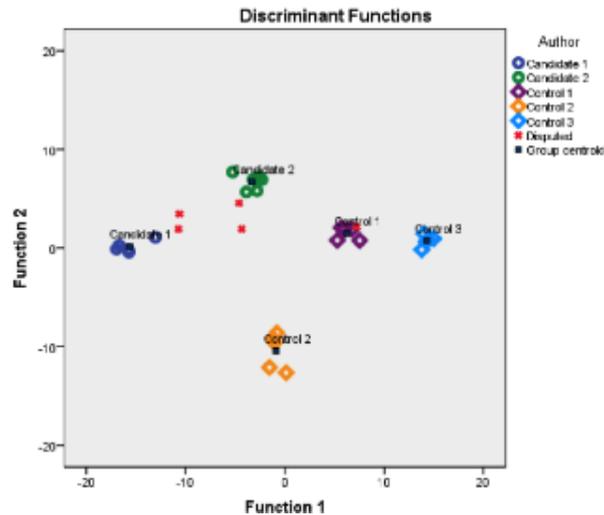
As shown in Table 5, the LDA classification method classified 100% of the texts correctly and confirmed that the analysis was 100% correct. However, this table illustrates that three of the five disputed excerpts were attributed to Candidate 2, one to Candidate 1 and another one to Control author 3.

Table 5. Classification and cross-validation results for bigrams – Case B

		Autor	Candidate 1	Candidate 2	Control 1	Control 2	Control 3	Total
			Predicted Group Membership					
Original	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	6	0	0	0	6
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5
		Disputed	**1**	**3**	**0**	**0**	**1**	**5**
Cross-validation	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	6	0	0	0	6
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5

100,0% of original grouped cases correctly classified.

100,0% of cross-validated grouped cases correctly classified.

Trigrams

In Figure 5 the LDA results for trigrams are projected. Most of the disputed excerpts are classified near the centroid of Candidate 2.

The LDA classification method successfully classified 100% of the texts by authors within their own group and the cross-validation method confirmed that the analysis

Figure 5. Linear Discriminant Function Analysis based on trigrams – Case B.
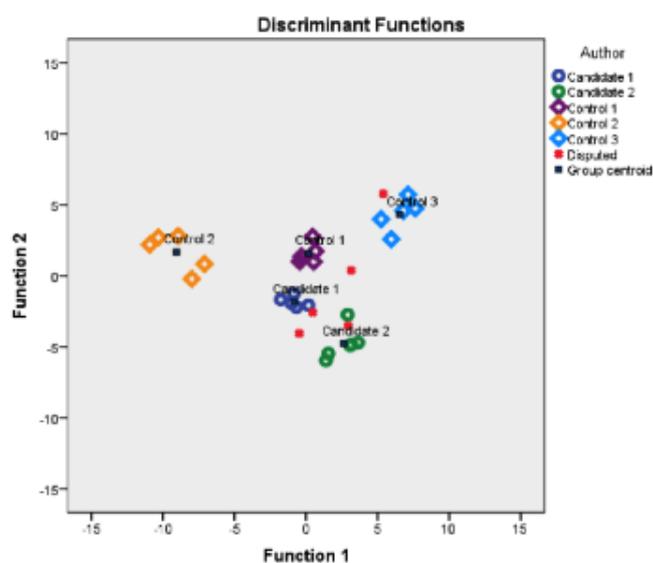
was 100% correct. As Table 6 illustrates, three of the disputed excerpts were attributed to Candidate 2 and two of the excerpts were classified near the control author centroid.

Table 6. Classification and cross-validation results for trigrams – Case B

			Predicted Group Membership					
		Autor	Candidate 1	Candidate 2	Control 1	Control 2	Control 3	Total
Original	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	6	0	0	0	6
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5
		Disputed	0	**3**	**1**	0	**1**	5
Cross-validation	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	6	0	0	0	6
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5

100,0% of original grouped cases correctly classified.

100,0% of cross-validated grouped cases correctly classified.

Results help us conclude that Candidate 1 can be rejected as a possible author of the disputed text and that there exists a moderate probability that the author of the disputed text could be Candidate 2.

Conclusions

We hope that the application of this technique can help forensic linguists to base their analyses on valid and reliable methods and techniques by using, in this case, sequences of linguistic categories in their analyses to be included in their expert witness reports, and to make the information much more comprehensible to the judge and the court, since 80% of the information included in the expert witness's report is usually incorporated

by the judges in their judgments and since the expert witnesses' most important duty is to assist the judge and give reliable forensic linguistic evidence in court.

Our view is that we need to refine our linguistic methods and techniques, and make them as valid and as reliable as possible, so that the unfortunately existing "room for maneuver" is reduce and, in turn, our opinions as forensic linguists are more scientifically grounded.

References

Bel, N., Queralt, S., Spassova, M. S. and Turell, M. T. (2012). The use of sequences of linguistic categories in forensic written text comparison revisited. In *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*, 192–209, Birmingham.

de Vel, O., Anderson, A., Corney, M. and Mohay, G. (2001). Mining e-mail content for author identification forensics. *SIGMOD Record*, 30(4), 55–64.

Hoey, M. (2005). *Lexical priming: A new theory of words and language.* London and New York: Routledge.

Queralt, S., Spassova, M. and Turell, M. T. (2011). L'ús de les combinacions de seqüències de categories gramaticals com a nova tècnica de comparació forense de textos escrits. *Llengua, Societat i Comunicació*, 9, 59–67.

Queralt, S. and Turell, M. T. (2012). Testing the discriminatory potential of sequences of linguistic categories (n-grams) in Spanish, Catalan and English corpora. In *The Regional Conference of the International Association of Forensic Linguists*, Kuala Lumpur, Malaysia.

Spassova, M. S. (2009). *El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español.*

Spassova, M. S. and Grant, T. D. (2008). Categorizing Spanish written texts by author gender and origin by means of morpho-syntactic trigrams: Some observations on method's feasibility of application for linguistic profiling. In *Curriculum, Language and the Law Inter-University Centre*, University of Zagreb, Dubrovnik (Croatia).

Spassova, M. S. and Turell, M. T. (2007). The use of morpho-syntactically annotated tag sequences as forensic markers of authorship attribution. In M. T. Turell, M. S. Spassova and J. Cicres, Eds., *Proceedings of the second european IAFL conference on forensic linguistics, language and the law*, 229–237, Barcelona: Publicacions de l'IULA.

Turell, M. T. (2004a). The disputed authorship of electronic mail: Linguistic, stylistic and pragmatic markers in short texts. In *First European IAFL Conference on Forensic Linguistics, Language and Law*, Cardiff: Cardiff University.

Turell, M. T. (2004b). Textual kidnapping revisited: The case of plagarism in literary translation. *International Journal of Speech, Language and the Law*, 11(1), 1–26.

Turell, M. T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *International Journal of Speech Language and the Law*, 17(2), 211–250.

Woolls, D. and Coulthard, M. (2007). Tools for the trade. *International Journal of Speech, Language and the Law*, 5(1), 33–57.

Computational approaches to plagiarism detection and authorship attribution in real forensic cases

Maria Teresa Turell & Paolo Rosso *

Abstract. *This paper integrates work done in the fields of forensic and computational linguistics by applying computational approaches to plagiarism detection and authorship attribution in real forensic cases. It reports on findings obtained through the application of computational tools, which have been useful in plagiarism detection and analysis of real forensic cases, and computer-aided queries of annotated corpora, which have allowed forensic linguists to test the statistical significance of new morpho-syntactic markers of forensic authorship attribution such as non discrete linguistic variables (i.e. Morpshosyntactically Annotated Tag Sequences) occurring in fairly long texts.*

*Keywords: **Forensic linguistics, computational linguistics, plagiarism detection, authorship attribution, real forensic cases.***

Introduction

One of the many areas of study of forensic linguistics has to do with written text comparison used to identify unique and idiosyncratic parameters of an individual's idiolectal style, working with the assumption that language can reveal a writer's socio-individual and socio-collective factors (age, gender, occupation, education, religion, geographical origin, ethnicity, race and language contact situation) and focusing on what the texts say and whether two or more texts have been written by the same author (authorship) or have been plagiarised from each other (plagiarism). Over the last two decades, forensic linguists have been claiming their ability to assist the court in civil and criminal proceedings and act as expert witnesses by meeting different evidentiary requirements depending on the diverse legal systems (Common and Civil Law).

---

*ForensicLab, Universitat Pompeu Fabra & Natural Language Engineering Lab, Universitat Politècnica de València

However, forensic linguists nowadays face two fundamental challenges. One has to do with the nature of written forensic texts, either long, but unique, in the sense that there may not be any other texts to compare with - so that there is only linguistic evidence within the text itself in order to establish possible plagiarism and/or come up with a linguistic profile in authorship attribution -, or very short texts (handwritten letters, annonymous electronic or type-written documents), for which the empirical evaluation of markers of disputed authorship is not easily allowed.

The second challenge is framed around the verification that the major part of the forensic written text comparison conducted these days is still quite unsystematic and unreliable. Therefore, there is a need to subject it to scrutiny in terms of methodologically-incorrect selections of the universe of non-disputed texts, ignorance of base rate knowledge information, lack of scientific design and analytical methods and, above all, in terms of the existing speculation as to the actual reliability that should be involved in evaluating whether there is more inter-writer than intra-writer variation and, furthermore, whether or not an individual's idiolectal style varies throughout his life span and beyond different genres.

Notwithstanding, during the last decade, scientific and reliable approaches to the kind of text comparison involved in plagiarism detection and authorship attribution, both in forensic and non-forensic cases, have responded to the need to rise to the challenges mentioned above. These approaches include stylometric measurements of an individual's style (Baayen *et al.*, 1996; Love, 2002; Feiguina and Hirst, 2007; Spassova and Turell, 2007; Grant, 2007; Chaski, 2001); identification of idiolectal styles (Chaski, 2001; Grant and Baker, 2001); stylistic methods (McMenamin, 2001), and vocabulary analytical techniques (Coulthard, 2004; Turell, 2004), with consideration of core lexical elements, hapax legomena, hapax dislegomena, lexical density and lexical richness and the use of corpus of reference in order to establish the low/high frequency of words in disputed text sets, by taking into account the concepts of markedness and saliency.

Several are the investigations carried out in order to deal with the above-mentioned difficulties and problems from a computational linguistics perspective. In Shivakumar and Garcia-Molina (1995) a copy detection approach based on word frequency analysis was introduced. In Kang *et al.* (2006) an approach based on word comparison at sentence level which takes into account vocabulary expansion with Wordnet[1] was described. A few methods attempt to solve plagiarism detection on the basis of word n-grams comparisons (Lyon *et al.*, 2004; Muhr *et al.*, 2010) or also character n-grams (Schleimer *et al.*, 2003; Grozea *et al.*, 2009). Recently, researchers have also approached the issue of cross-language plagiarism (Potthast *et al.*, 2011; Barrón-Cedeño *et al.*, 2010; Gupta *et al.*, 2012; Franco-Salvador *et al.*, 2013). When there may not be any suspicious text to compare the suspicious document with, the linguistic evidence may have to be provided on the basis of stylistic changes found in the document itself (intrinsic plagiarism) (Stein and Meyer zu Eissen, 2007). Another method for intrinsic plagiarism detection is the one described in Stamatatos (2009), where character n-gram profiles have been used. Computational linguists have also considered the somewhat related issue of authorship attribution, where linguistic profiles need to be investigated in order to try to determine who the real author of a text is (Stamatatos *et al.*, 2000; Koppel *et al.*, 2009).

This article argues that forensic plagiarism detection and authorship attribution can benefit from the complementary interplay between approaches used in forensic linguis-
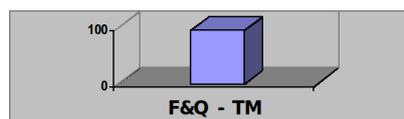
Figure 1. Overlapping vocabulary (F&Q-TM) Activities (Turell, 2008: 288)

tics (textual qualitative analysis, observation of semantic and pragmatic markers that cannot be analysed by automatic procedures, the use of reference corpora to set up the rarity or expectancy of the writers' idiolectal choices) and those used in computational linguistics, automatic or semiautomatic natural language processing techniques that would allow researchers to establish the statistical significance of results, something which is becoming more and more necessary when acting as linguistic expert witnesses in court. The studies presented in this article report on findings from the application of both computational tools and computer-aided queries of annotated corpora. They have been useful in plagiarism detection and analysis of real forensic cases, allowing forensic linguists to test the statistical significance of new morpho-syntactic markers of forensic authorship attribution such as non discrete linguistic variables (i.e. Morpshosyntactically Annotated Tag Sequences) occurring in fairly long texts.

In-tandem forensic and computational approaches to plagiarism detection

Forensic linguistics makes a distinction between copying of ideas and linguistic plagiarism. Copying of ideas can exist without linguistic plagiarism but if the latter is detected, the former occurs as well by the nature and definition of the linguistic sign itself. This discipline has devised a methodology which includes several qualitative and quantitative tools and is used to establish a) the nature and degree of plagiarism, b) plagiarism directionality, and c) the threshold level of textual similarity between texts above which this similarity becomes suspicious. For the purposes of this article, we will consider point c) in particular and explain what tools are used to analyse this textual similarity.

CopyCatch[2], one of the many existing concordance tools used to detect linguistic plagiarism, allows researchers to calculate the threshold level of textual similarity which becomes suspicious in a forensic case. This program incorporates several measurements such as threshold of overlapping vocabulary, hapax legomena, hapax dislegomena, unique and exclusive vocabulary and shared-once phrases. It has a visual output, with plagiarised sections marked in red, which is very useful when presenting evidence in court. In order to establish this threshold level of textual similarity forensic linguists can count on empirical evidence which suggests that "up to 35% similarity is normal and up to 50% is not unusual, although the further above 50% the more likely it is to indicate that the texts under consideration have not been produced independently and that there exists a borrowing relationship between the texts under consideration". Empirical research has also proved that this threshold level should be increased up to 70%, in the case of plagiarism between translations (Turell, 2004).

Out of the eight (8) Spanish plagiarism detection cases, in which our forensic linguistics laboratory has been involved in the last 7 years, three (3) present verbatim plagiarism and the other five (5) reflect partial plagiarism, with varying degrees of paraphrase and overlapping vocabulary. The domains of occurrence of these cases are education (text

Table 1. Decontextualization of F&Q (Bruño) in TM (Temario Magister) (Turell, 2008: 286)

Física y Química Bruño (2002) (page 194)	Temario Magister (2005) (Unit 11, pages 8 & 9)
ACTIVITY 2	ACTIVITY 1
2. Completa en tu cuaderno la siguiente Tabla 11.2 consultando la Tabla Periódica.	1. Completa la siguiente Tabla:

ENGLISH TRANSLATION

Física y Química Bruño (2002) (page 194)	Temario Magister (2005) (Unit 11, pages 8 & 9)
ACTIVITY 2	ACTIVITY 1
2. Complete in your exercise book the following Table 11.2 consulting the Periodic Table.	1. Complete the following Table:

books), tourism (guides and brochures), scientific research, music (lyrics) and literature (novels). For example, Figure 1 shows the threshold level of overlapping vocabulary (96%) found when comparing the sections on Activities in the two text books under analysis: the non-disputed text Física y Química Bruño (F&Q, 2002) and the disputed text Temario Magister (TM, 2005) (Bruño vs. Magister),a percentage which indicates that Activities in F&Q have been reproduced almost verbatim in TM.

One of the other equipment facilities supplied by CopyCatch is that the program takes you to the Sentences Page automatically. This Sentences Page, which presents verbatim, or almost verbatim, phrases/sentences, facilitates the identification of the uncoherent/uncohesive segments and the plagiarist's strategies within the whole text, which may lead to a) meaningless sequences due to the 'cut & paste' technique used, b) inconsistency in referential style, c) decontextualisation and d) inversion in the grading of structural elements, among others.

Table 1 illustrates one example of decontextualisation produced by the fact that one part of the directions given in Activity 2 (page 194) in the non-disputed text (F&Q, 2002), namely, "consultando la Tabla Periódica", has been deleted in Activity 2 (Unit 11, page 8) in the disputed text (TM, 2005).

Pl@giarism[3], developed by the University of Maastricht, is another system used in plagiarism detection. The system returns the percentage of textual similarity between two documents (A and B), the percentage of the number of matches with document A versus document B, the percentage of the number of matches with document B versus document A and the total amount of matches between documents A and B. This system performs the comparison on the basis of word trigrams. Like CopyCatch, Pl@giarism has a visual output, with plagiarized sections marked in red. WCopyFind[4], developed by the University of Virginia, is another tool made available for plagiarism detection (Vallés Balaguer, 2009). The system allows researchers to introduce various parameters such as the size of word n-grams, the minimum number of matching words to be reported as possible plagiarism, the maximum number of non matches between perfectly matching portions of a sentence, etc. Apart from highlighting the substitution of words

Table 2. Comparison of results: CopyCatch v. WCopyFind v. Pl@giarism

	CopyCatch	WCopyFind	Pl@giarism
Case: Bruño v. Magister			
Page 5	27 %	19 %	21 %
Page 32	79 %	92 %	92 %
Pages 33-37	95 %	96 %	96 %
Pages 40-46	94 %	86 %	83 %
Case: XXX (for anonymity) v. Signes			
Activities	96 %	87 %	86 %
Cuestionario	98 %	87 %	78 %
Técnico	86 %	83 %	78 %

by synonyms, one added value of WCopyFind is the provision of a word map (a generalized thesaurus). This system tells researchers the percentage of the number of matches with document A versus document B, the percentage of the number of matches with document B versus document A, and like CopyCatch and Pl@giarism, WCopyFind has a visual output, with plagiarized sections marked in red.

Table 2 shows a comparison of CopyCatch, WcopyFind and Pl@giarism results related to the detection of plagiarized fragments found in the corpus sets of two real forensic cases (Bruño vs. Magister and XXX v. Signes. In WCopyFind we used trigrams as n-gram size, and five as the maximum number of non matches. Results illustrate that the three tools have been able to detect plagiarism in these real cases. However, in most cases, the tool CopyCatch returns the highest percentage of similarity between the texts compared. This is because this tool runs on unigrams, once-shared words and unique vocabulary. WCopyFind returns a higher percentage than Pl@giarism, the main reason for this being that WcopyFind considers the possible words inserted between perfectly matching sentence fragments, whereas Pl@giarism does not.

For real cases Bruño vs. Magister and XXX v. Signes, some linguistic evidence was given on the basis of the comparison with other texts. However, as mentioned above, one of the challenges forensic linguists have to face is that very often written forensic texts are unique in the sense that there may not be any other texts against which to compare them. Therefore, in order to be able to establish possible plagiarism the linguistic evidence has to be found within the text itself (intrinsic plagiarism). YYY[5] is a new tool the aim of which is to help forensic linguists to come up with a linguistic profile on the basis of a stylistic analysis of the text in order to determine whether or not there are fragments of different writing styles. This tool divides the text into fragments and for each of these fragments, it calculates various vocabulary richness measurements (function K proposed by Yule (Yule, 1944), function R proposed by Honore (Honore, 1979)) and text complexity (Flesch-Kincaid Readability Test (Flesch, 1948), Gunning Fog Index (Gunning, 1952)) as suggested in Meyer zu Eissen *et al.* (2007). The aim behind Stylysis is to identify text fragments with different writing styles, which could indicate that it is a plagiarized fragment or that it has been written by a different author. Thus, this tool could also help in linguistic profiling to attribute authorship of a text.
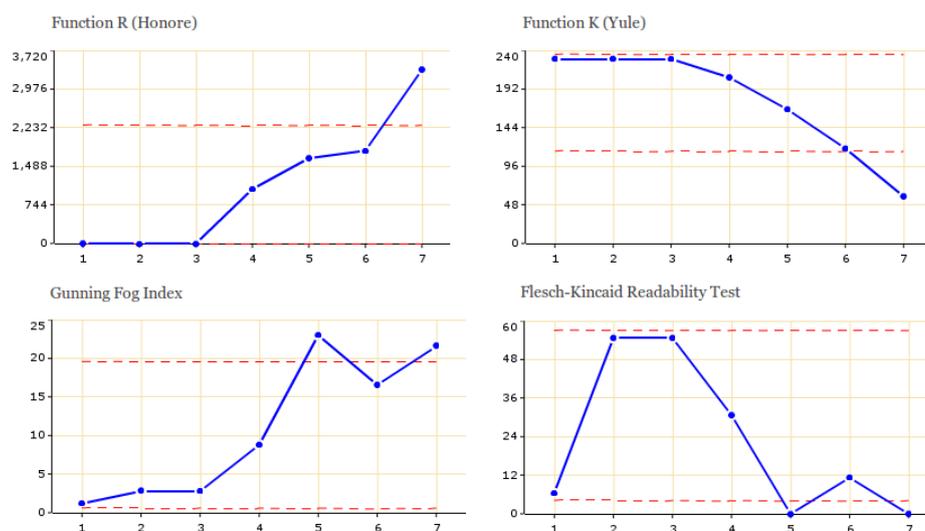
Figure 2. Results obtained with the Stylysis tool in case XXX v. Signes (5)

Figure 2 shows the results for page 5 of the corpus set of case XXX v. Signes. R and K functions show that fragment 7 exceeds the standard deviation (dashed line). This could indicate a possible change in writing style. The Flesch-Kincaid Readability Test and the Gunning Fog Index show that fragments 5 and 7 exceed the standard deviation. On the basis of these measurements, fragment 7 is the only suspicious fragment. And indeed fragment 7 proved to be a case of plagiarism. This example demonstrates the usefulness of Stylysis in providing language evidence for intrinsic plagiarism detection.

The use of computer-aided queries of annotated corpora in forensic authorship attribution

One computational approach used in the kind of forensic text comparison leading to more reliable authorship attribution outcomes is the study of the syntactic characterization of a writer's idiolectal style through computer-aided queries of annotated corpora that can help to establish the statistical significance of sequences of linguistic categories, namely, Morpho-syntactic Annotated Tag Sequences (MATS), as proposed in Spassova and Turell (2007). This approach is not new in non-forensic contexts (Baayen *et al.*, 1996), where these sequences are frequently referred to as n-grams (and depending on the number of categories combined, the terms used are bigrams, trigrams, etc.), but the ForensicLab has been one of the first to apply this method to real forensic cases.

This method is structured around the following activities:

1. A pre-processing phase, in which texts are segmented into their basic components: title, paragraphs, sentences, and paragraph beginnings and ends are marked ($</s></p>$).
2. A morpho-syntactic tagging phase, during which the text is converted into a flow of token types and tags.
3. A disambiguation stage, through which texts are disambiguated and errors are corrected.
4. A tag extraction phase - making use of LEGOLAS 2.0 - during which the information obtained refers to the number of MATS types and tokens and on the MATS frequency values to be used in the subsequent statistical analysis.

\<s>	El	comercial	de	la	empresa	vendía	los	mejores	hoteles	\</s>
...	AMS	N5-MS	P	AFS	N5-FS	VDA6S	AMP	ADJ	N5-MP	...

Bigram   Trigram
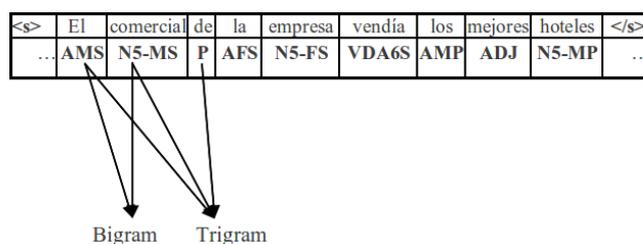
Figure 3. Morpho-syntactically Annotated Tag Sequences (MATS)

5. Once the tags have been extracted, a last stage involves the application of Discriminant Function Analysis (DFA), in order to classify the different text sets, and the projection of results onto graphs.

During the pre-processing and processing phases several processing and disambiguation tools from the IULA's technical corpus[6] were used (Morel *et al.*, 1998). For example, once tagged, the sentence: "el comercial de la empresa vendía los mejores hoteles" ('the firm's salesperson was selling the best hotels') is projected and represented in the way Figure 3 illustrates. Two examples of MATS are marked, namely, "el comercial" (AMS N5-MS), which is a bigram and "el comercial de", (AMS N5-MS P), which is a trigram, and where A stands for article, M for masculine, S for singular, N5-MS for singular masculine common noun, and P for preposition.

Part of the linguistic evidence used to report on six (6), out of nine (9), Spanish real forensic authorship attribution cases considered was drawn by applying this methodological protocol. Outcoming results from all these cases have shown that the discriminatory potential of MATS is higher with long text samples and with a big number of control reference texts (Grant, 2007) and that bigrams and trigrams are more discriminatory than longer sequences of linguistic categories. To test the working hypotheses which are at play in forensic written text comparison - that is, a) that everyone has an 'idiolectal style', as relating to the linguistic system shared by lots of people, but used in a distinctive and unique way by particular individuals, who have different options at their reach in their linguistic repertoire and make selections from these options (Halliday, 1989), and b) that a disputed text (or several disputed texts) can be attributed to the same author who wrote a set of non-disputed texts - sets of anonymous texts from other real forensic cases are used, thus optimizing the discriminatory potential of MATS.

Figure 4 shows the projection for bigrams of the DFA applied to the three text sets under analysis in the forensic case XXX v. SEHRS: the disputed e-mails (+), the non-disputed faxes (△) and the anonymous emails from another case (○, indicated as docs in Figure 4). In this figure, it can be seen that although there is a certain distance between the disputed emails and the non-disputed faxes, the distance of all the emails from another case and the text sets relevant to the case under analysis is even bigger, which indicates more statistically significant differentiation in the idiolectal use of MATS in the emails from another case than the one found in the comparison of the NDTfax and DT@ text sets.

The classification method of DFA classified with success 100% of the texts by authors within their own group while the cross-validation method confirmed that the analysis was 83.4% correct. Two of these emails belong to the non-disputed faxes set, whereas the
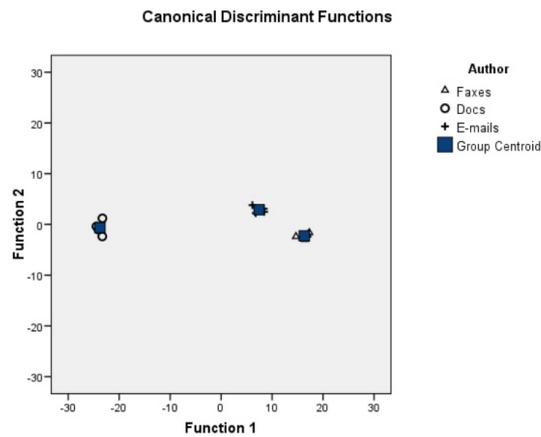
Figure 4. Discriminant Function Analysis (NDTfax and DT@) - Bigrams (Turell, 2010)
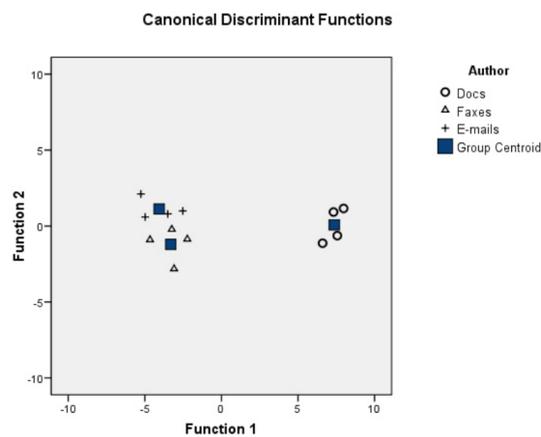
Figure 5. Discriminant Function Analysis (NDTfax and DT@) - Trigrams (Turell, 2010)

other two disputed emails are classified within its original group, and the anonymous emails from another case are all classified within their original group. This outcome seems to confirm that the probability that the author of the disputed emails could be the author of the non-disputed faxes is quite high.

Figure 5 shows the projection of the results for trigrams. This figure illustrates that the centroids of the disputed emails ($+$) and the non-disputed faxes ($\triangle$) are placed in the same area of the graph, while the centroid of the emails from another case ($\bigcirc$) is located in the opposite side of the graph.

However, on this occasion the DFA classification method shows that only 75% of the texts are classified with success; only three of the disputed emails are classified as if produced by the author of the non-disputed faxes; besides, cross-validation confirmed that the analysis was only 63% correct, since two of the disputed emails are attributed to the group of non-disputed faxes and two of the non-disputed faxes are attributed to the group of disputed emails.

These statistical results using DFA reveal that, in spite of the reduced corpus size and the short text length (although in effect the total N for MATS is not that small - 1,589 tokens for bigrams and 660 tokens for trigrams), these structures seem to exhibit a quite

high discriminatory potential and also that bigrams turn out to be more discriminatory than trigrams, as other forensic cases have shown. This would allow us to conclude that sequences of grammatical categories observed in a writer's 'idiolectal style' can be used quite reliably as valid markers of authorship.

Conclusions

In this article we have attempted to show that, for both computational and forensic linguistics, joint work in the areas of forensic plagiarism detection and authorship attribution can be very fruitful. Present-day comparative methods used in forensic plagiarism detection and authorship attribution exhibit limitations; so there is a need to count on intra-evidential complementary evidence which is tested with computational (automatic or semi-automatic) natural language processing techniques. Computational linguistics, on the other hand, needs to be able to use linguistic data from real forensic cases - and not just synthetic data, automatically generated or manually generated through Amazon mechanical Turk (Potthast *et al.*, 2010) - in order to establish the actual performance features of the existing systems of automatic plagiarism detection (Barrón-Cedeño *et al.*, 2013). It is precisely because of the nature and length of forensic texts (usually quite short) and corpora (small-sized), which can be a drawback when trying to establish the statistical significance of results, that computational linguistics must come into play so that forensic linguists are able to refine their comparative methods and techniques. In this article we have reported on the comparative evaluation of three plagiarism detection tools (CopyCatch, extensively used in forensic plagiarism detection, WCopyFind and Pl@giarsm) that are available to forensic linguists, while we are aware that there are other automatic plagiarism detection systems that define the state of the art and are part of the know-how of the PAN competition on Plagiarism Detection (Potthast *et al.*, 2012).

One important empirical question that can be raised, but not answered in this article, is what kind of evaluation results would be drawn when automatic detection tools can be applied to real forensic data, once these systems are commercialized or become public domain tools. This is only one first enriching step towards the establishment of stronger collaboration links between forensic and computational linguists. However, it has not been possible to compare the forensic protocols and automatic authorship attribution techniques used in forensic linguistics with other existing automatic approaches to written authorship such as those devised by computational linguistics, which will be discussed in the context of the PAN competition on Authorship Identication[7].

Acknowledgments

Cases cited

Case: Bruño vs. Magister.
Case: XXX v. Signes.
Case: XXX vs. SEHRS.

Notes

[1]http://wordnet.princeton.edu

[2]CopyCatchGold, CFL Development, http://cflsoftware.com/

[3]http://www.plagiarism.tk

[4]http://www.plagiarism.phys.virginia.edu/Wsoftware.html

[5]http://memex2.dsic.upv.es:8080/StylisticAnalysis/es/index.jsp

[6]http://brangaene.upf.es/plncorpus/index2.htm, http://brangaene.upf.es/plncorpus/faq.html

[7]http://pan.webis.de/

References

Baayen, R. H., van Halteren, H. and Tweedie, F. J. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3), 121–131.

Barrón-Cedeño, A., Rosso, P., Agirre, E. and Labaka, G. (2010). Plagiarism detection across distant language pairs. In *Proc. of the 23rd International Conference on Computational Linguistics, COLING-2010*, 37–45, Beijing, China.

Barrón-Cedeño, A., Vila, M., Martí, A. and Rosso, P. (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4).

Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics. The International Journal of Speech, Language and the Law*, 8(1), 1–65.

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

Feiguina, O. and Hirst, G. (2007). Authorship attribution for small texts: Literary and forensic experiments. In *International Workshop on Plagiarism Analysis, Authorship Identification and Near-Duplicate Detection, PAN 2007*: 30th Annual International ACM SIGIR Conference.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233.

Franco-Salvador, M., Gupta, P. and Rosso, P. (2013). Cross-language plagiarism detection using multilingual semantic network. In *Proceedinfs of 35th European Conference on Information Retrieval, ECIR-2013*, Moscow, Russia.

Grant, T. (2007). Quantifying evidence in forensic authorship analysis. *The International Journal of Speech, Language and the Law*, 14(1), 1–25.

Grant, T. and Baker, K. (2001). Identifying reliable, valid markers of authorship: a response to Chaski. *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 8(1), 66–79.

Grozea, C., Gehl, C. and Popescu, M. (2009). Encoplot: Pairwise sequences matching in linear applied to plagiarism detection. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel and E. Agirre, Eds., *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, 10–18, San Sebastian, Spain.

Gunning, R. (1952). *The technique of clear writing.* McGraw-Hill Int. Book Co.

Gupta, P., Barrón-Cedeño, A. and Rosso, P. (2012). Cross-language high similarity search using a conceptual thesaurus. In *Proceedings of 3rd International Conference of CLEF on Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics, CLEF 2012*, 67–75, Rome, Italy.

Halliday, M. (1989). *Language, context and text. Aspects of language in a social semiotic perspective.* Oxford: Oxford University Press.

Honore, A. (1979). Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2), 172–177.

Kang, N., Gelbukh, A. and Han, S. Y. (2006). Ppchecker: Plagiarism pattern checker in document copy detection. *Lecture notes in computer science.*, 4188, 661–668.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1), 9–26.

Love, H. (2002). *Attributing authorship: An introduction.* Cambridge: Cambridge University Press.

Lyon, C., Barrett, R. and Malcolm, J. (2004). A theoretical basis to the automated detection of copying between texts, and its practical implementation in the ferret plagiarism and collusion detector. In *Plagiarism: Prevention, Practice and Policies Conference*, Newcastle, UK.

McMenamin, G. (2001). Style markers in authorship studies. *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 8(2), 93–97.

Meyer zu Eissen, S., Stein, B. and Kulig, M. (2007). Plagiarism detection without reference collections. In R. Decker and H. Lenz, Eds., *Advances in Data Analysis*, 359–366.

Morel, J., Torner, S., Vivaldi, J. and Cabré, M. T. (1998). *El corpus de l'IULA: Etiquetaris.* Serie Informes, 18 (2nd Edition). Barcelona: Publicacions de l'IULA.

Muhr, M., Kern, R., M., Z. and Granitzer, M. (2010). External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system. In *Notebook Papers of CLEF 2010 LABs and Workshops*, University of Padova, Italy.

Potthast, M., Barrón-Cedeño, A., Stein, B. and Rosso, P. (2010). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics, COLING-2010*, 997–1005, Beijing, China.

Potthast, M., Barrón-Cedeño, A., Stein, B. and Rosso, P. (2011). Cross-language plagiarism detection. *Languages Resources and Evaluation. Special Issue on Plagiarism and Authorship Analysis*, 45(1).

Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P. and Stein, B. (2012). Overview of the 4th international competition on plagiarism detection. In *Notebook Papers of CLEF 2012 LABs and Workshops, CLEF-2012*, Rome, Italy.

Schleimer, S., Wilkerson, D. S. and Aiken, A. (2003). Winnowing: Local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, New York, NY.

Shivakumar, N. and Garcia-Molina, H. (1995). Scam: A copy detection mechanism for digital documents. *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries.*

Spassova, M. S. and Turell, M. T. (2007). The use of morpho-syntactically annotated tag sequences as forensic markers of authorship attribution. In M. T. Turell, M. S.

Spassova and J. Cicres, Eds., *Proceedings of the Second European IAFL Conference on Forensic Linguistics/Language and the Law.*, 229–237, Barcelona.

Stamatatos, E. (2009). Intrinsic plagiarism detection using character n-gram profiles. *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software*, 38–46.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4), 461–485.

Stein, B. and Meyer zu Eissen, S. (2007). Intrinsic plagiarism analysis with meta-learning. In *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN 2007*, Amsterdam, Netherlands.

Turell, M. T. (2004). Textual kidnapping revisited: the case of plagiarism in literary translation. *The International Journal of Speech, Language and the Law*, 11(1), 1–26.

Turell, M. T. (2008). Plagiarism. In J. Gibbons and M. T. Turell, Eds., *Dimensions of Forensic Linguistics*, 265–299, Amsterdam/Philadelphia: John Benjamins.

Turell, M. T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law*, 17(2), 211–250.

Vallés Balaguer, E. (2009). Putting ourselves in sme's shoes: Automatic detection of plagiarism by the wcopyfind tool. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel and E. Agirre, Eds., *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, 34–35, San Sebastian, Spain.

Yule, G. (1944). *The statistical study of literary vocabulary.* Cambridge University Press.

Codal variation theory as a forensic tool

Andrea Nini [*]

Abstract. *This paper addresses how the Systemic Functional Linguistics (SFL) theoretical framework of* codal variation *(Hasan, 1990) can be helpful when applied to the field of forensic linguistics, especially for authorship analysis. Firstly, the framework will be introduced and discussed in the lights of traditional modern sociolinguistics. Then, it will be shown how the concept of* codal variation *can be useful for describing and understanding the idiolect, or, in SFL terms, the personalised meaning potential of an individual. An example of a successful application of this concept will be taken from the Bentley case, where the distinction between two codes proved to be of high evidential value. The discussion will then lead on to the implications that* codal variation *could have for authorship/sociolinguistic profiling, considering other examples from the literature for which an SFL interpretation could lead to an improvement. Combining the theory of* codal variation *with Biber's multidimensional framework can represent a first step towards building a method of authorship analysis that is driven by the knowledge of population base-rate of a number of linguistic variables. An example from a real case will be presented where this kind of analysis proved to be a promising first step towards a theoretically valid methodology for authorship analysis. Possible improvements and directions for more research will be illustrated and discussed.*

Keywords: Forensic linguistics, authorship analysis, authorship attribution, systemic functional linguistics, multidimensional analysis, codal variation, sociolinguistics, corpus linguistics, Bentley case.

Introduction

This paper will present the implications that *codal variation theory*, a particular branch of Systemic Functional Linguistics (SFL), can have for authorship analysis. Authorship analysis is currently dominated by two strands of methodologies: the qualitative stylistic methods and the quantitative stylometric methods. A major problem with these approaches is that they do not address the issue of providing a valid explanation of why

---

[*]Aston University

authorship analysis is possible. This paper proposes the theory of codal variation as a way of introducing a strong theoretical framework for authorship analysis, thus potentially solving the problem of theoretical validity (Grant and Baker, 2001). By providing a hypothesis of why authorship analysis is possible, new theories can be tested and better practice can be developed.

Semantics in SFL

Before introducing the discussion on codal variation a brief digression is needed on the model of semantics in SFL. This is necessary as many concepts introduced below depend on a particular understanding of semantics, which is marginally different from the mainstream sociolinguistic definition of the term.

SFL models the meaning of a lexicogrammar item as being the function that it serves. That being so, SFL does not consider meaning to be truth-conditional, as traditionally conceived by generativists and sociolinguists. For example, the clauses:

a) *Mary eats the apple*

b) *Is Mary eating the apple?*

c) *The apple is eaten by Mary*

in SFL do not mean the same thing. In this framework, meaning corresponds to function and function is in turn organised in a three folded division in major functional strands, or *metafunctions*: these are the *ideational metafunction*, that is, the function of language to represent things and events; the *interpersonal metafunction*, that is, the function of language to communicate interactions or to make people interact; the *textual metafunction*, that is, the function of language of distributing the information so to anchor the text to the context (Halliday and Matthiessen, 2004).

Hence, the clauses presented above are constant in their *ideational* meaning, since the same Agent (Mary) is acting in the same Process (eat) to the same Goal (apple), but do present variation at the *interpersonal* and *textual* meanings. Clause (b), as opposed to clause (a) and (c), is a yes-no question and it therefore realises the meaning of being a request for information, thus situating the speaker as the person who seeks information and the hearer as the person who is assumed by the speaker as having the information. Clause (c), as opposed to clause (a) and (b), is a passive clause which makes the Goal (apple) the starting point of the clause. In clause (c) the speaker thus expresses a *textual* meaning that consists in the assumption that the hearer knows something about an apple and that it is likely to be a new information that it is Mary who eats it (this explanation is rather simplified for reasons of space; Halliday and Matthiessen 2004: 64).

Although traditionally a division is adopted between *meaning* and *style*, where *meaning* corresponds to the *ideational metafunction* and *style* corresponds to the combination of the *interpersonal* and *textual metafunctions*, for the rest of the paper, when the term meaning is employed, it is generally adopted to intend the three metafunctions.

Codal variation

It is possible to explain codal variation by comparing how linguistic variation is modelled in traditional linguistics and in SFL. Two kinds of variation are widely recognised in both SFL and traditional modern linguistics:

1. *Situational variation* (or *registerial variation* in SFL): the variation of meanings that is found in texts produced in different contexts.

2. *Social variation* (or *dialectal variation* in SFL): the variation in the way of real-ising meanings, which originates by the fact that different social groups have alternative ways of expressing the same meanings.

For the sake of further explanation two examples can be used to illustrate this point: (1) *registerial variation* is the variation in the frequency of past tenses between a story and an academic paper, a variation given by the different contexts in which the writer operates and thus independent on the person who is writing; (2) *dialectal variation* is the variation in the ways of making the same meanings within different social groups, such as, for example: for *ideational meaning*: 'pail' vs. 'bucket'; or for *interpersonal meaning*: different ways of realising a tag question: 'isn't it?' vs. 'innit?'.

The assumption underlying the postulation of these two kinds of variation is that there are two causes for the variation (context and social group) and two areas in which the variation appears (meanings and realisations of the meanings). Moreover, another assumption of the models is that there can be only two combinations: context creates variation in meanings *whereas* social group creates variation in the realisations of the meanings. That is, if the context varies then there is production of different meanings but not of different realisations of meanings and, consequently, if the social group varies then there is production of different realisations of meanings but not of different meanings.

SFL, however, adds another level of variation by considering another possible com-bination between these parameters: social group *to* meanings. This is *codal variation*:

3. *Codal variation*: the variation of meaning in relation to the social group, when the texts considered are produced in a comparable context.

The concept of codal variation was originally developed by Hasan (1990) under the term *semantic variation*. The term was later on reframed by Matthiessen (2007) as *codal variation* to avoid ambiguity with other interpretations of the term 'semantic variation'.

The term *codal* originates from the main source of theoretical influence of this con-cept, Bernstein's concepts of *restricted* and *elaborated codes*. Bernstein (1962) proposed that social classes produce different meanings in similar context because of the way so-cial classes interpret the context. In his original proposal, he claimed that this leads to individuals coming from a working class background producing different grammatical structures from individuals coming from a middle class background. Bernstein's work was extremely controversial and generated a debate with other approaches, in particular with Labovian variationist sociolinguistics (Martin, 1992: 573). The debate was gener-ated by Bernstein's suggestion that people from different social classes produce differ-ent meanings. Such a claim was not warmly welcomed in a mainstream linguistics that at the time professed that semantics was universal and based on the truth-conditional meaning. The issues at stake in the debate can be exemplified by discussing Figure 1 below.

Figure 1 represents a series of *realisations*. The table has to be read as each layer being realised by the one below it. Therefore: context is realised by semantics, which in turn is realised by lexicogrammar, which in turn is realised by a form of expression. Although this table represents a model of language typically found almost only in SFL, certain assumptions underlying the table are indeed shared by traditional modern so-ciolinguistics as well. Indeed, the theoretical stance underlying this table can be also found in the concept of *sociolinguistic variable*.
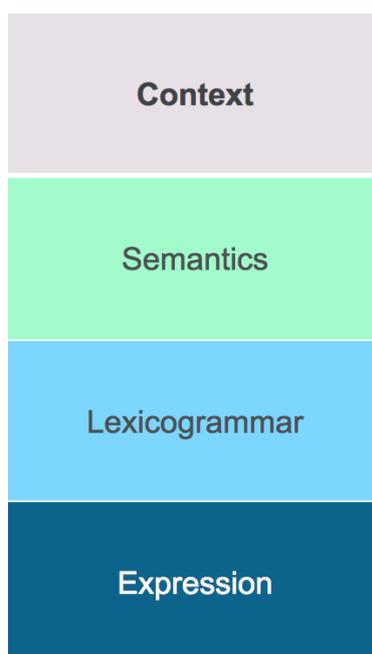
Figure 1. Levels of language according to SFL.

It is possible to reformulate the definition of a sociolinguistic variable as a socially distributed linguistic variable that measures variation at one level of language by keeping constant the level that is realised by it (or, diagrammatically speaking, the level above it). For example, a variation at expression level, such as the often studied *g clipping*, can be studied only in cases where the lexis (or lexicogrammar, in SFL terms) is constant (e.g. *singing* vs *singin'* or, grammatically speaking, *present continuous realised phonetically by* [ŋ] vs *present continuous realised phonetically by* [n]). For the same reason, lexicogrammatical variation, in the form of either lexical alternations or syntactic/morphological alternations, can be studied only when the semantics is kept constant, that is, when you have 'different ways of saying the same thing' (e.g. *car* vs *automobile*; copula deletion: *he is working* vs *he working*).

It follows that it is indeed possible to model semantic variation *only* if something can be held as constant above the semantics. In SFL, this something is the context, which is modelled as *a level of language* that is realised by the semantics. Traditional Labovian sociolinguistics lacked a model of context. However, as Hasan (2009) pointed out, a model of context is necessary to create an integrated sociolinguistic theory. The level of context on top of semantics is not just an addition to the face structure of the model. The theoretical significance of modelling an additional layer is indeed that semantics is not constrained to be the truth-conditional universal that generativists presuppose but that it is another level of language as a whole. That being so, as all the other levels of language, this level can vary sociolinguistically and it is arbitrary in the same way as lexicogrammar. According to Hasan (2009), this understanding of meaning allows the development of a new kind of sociolinguistics that takes into account how meanings together with form vary across social groups.

Hasan's (1990) work represents an empirical demonstration of this concept. Her experiment showed that the variation at the level of semantics is correlated with social

groups in the same way as the variation of Labovian sociolinguistic variables at the other levels are correlated with social groups. In her study, Hasan recorded samples of Australian working class and middle class mothers in the process of talking and playing at home with their children. She then analysed these samples using SFL and used Principal Component Analysis to group her variables. The components obtained can be thought of as semantic styles in the context of mother-child home conversation in Australia. After assigning component scores to the dyads mother-child, an ANOVA revealed that the differences between the two social classes were significant. Hasan interpreted the results as pointing to the fact that the speech of mothers talking to their children is influenced by the family's social positioning in terms of social class.

In other words, Hasan's experiment is evidence that different social classes operated in the same context in different ways and therefore produced different meanings when dealing with the context of regulating children's behaviour. These different ways of dealing with the context point to a difference in how the context is interpreted and understood by the two social groups.

Similar studies were then reproduced in the SFL community by other researchers (Martin (1992: 578) cites many; Hasan (1996); Rochester and Martin (1979)).

Codal variation in the forensic context

The usefulness of this theory for forensic purposes has already been shown, although codal variation has never been defined as such. The most significant example of an application of this theory is the Derek Bentley case (Coulthard and Johnson, 2007).

The analysis of the statements involved in the case showed that police officers and lay people differed in the position of the word *then* in the context of a police statement. Coulthard's analysis showed that, in the context of a police statement, the phrase *I then* is used significantly more by police officers than by lay people, who in turn prefer *then I*. Analysing this semantically using SFL indicates that these two constructions are indeed different regarding their textual meanings.

In a clause, the structure that is used to express textual meaning is the Theme-Rheme structure. As Halliday and Matthiessen (2004: 64) propose, the Theme of a clause is '[the element of the clause] which locates and orients the clause within its context'. In the example of the Bentley case, the opposition between *I then* vs. *then I* corresponds to a shift between which element starts the clause and therefore in how the writer wants to orient the reader. For example, *I then followed* the man shifts the start of the clause on *I*, the Subject (and therefore Agent, if the clause is transitive), thus relegating the temporal orientation in the segment of the clause that is presupposed to be already known by the hearer. On the other hand, *then I followed the man* takes the adjunct *then* inside the Theme, thus adding more emphasis on the temporal orientation of the clause. Although the difference in meaning between these two clauses is rather subtle, the opposition between these two variants is undoubtedly one of meaning.

Since this difference in meaning seems to be correlated with social group, in the form of 'policemen' vs. 'lay people', it is possible to conclude that this variation found in statements is similar to what Hasan (1990) found in her studies on socialisation processes in Australia. What Coulthard and Johnson (2007) refers to as *police register* when describing the fact that police officers use *I then* could probably be regarded as an instance of codal variation and it is therefore possible to substitute the term *police register*

with *police code*. Replacing the term is not just an ideological position but also a meaningful theoretical tool. If a feature is recognised as an instance of codal variation, the explanations and predictions theorised within SFL within the theory of codal variation can be extended to the particular instance under analysis. Since codal variation exists because of the different interpretation of the context that social groups develop as part of their sub-culture and the experience they have with a particular genre, in this case it is possible to hypothesise that the difference between *I then* and *then I* is originated from a different interpretation of the context of a police statement given by the social group *police officers*. Lay people who do not experience statements in the same way as policemen seem to interpret the genre as a form of narrative, thus providing elements of narrative like the focus on temporal sequences such as the thematisation of *then*. On the other hand, the experience of policemen and their community of practice trains them to focus on more important things, such as the Subject (most of the times also Agent) of the clause.

The usefulness of this theoretical construct is immediately apparent for authorship profiling and attribution and it is indeed already applied in forensic contexts, although never recognised as such. Chaski (2001) and her syntactic markers are an example of codal variation used for attribution, as the differences between preferences of determiners or use of different verb phrases in the context of emails is indeed an instance of codal variation. Studies such as Koppel *et al.* (2002) or Argamon *et al.* (2009) are again finding out differences in coding orientations between social groups such as age or gender. By knowing why these differences are found and by contextualising this practice in a broader theory, practice can be informed and improved in order to produce more accurate analyses.

Towards a method: Biber's multidimensional analysis

Whether the theory produces valid hypotheses or not is only partially tested. However, the results so far are promising enough to justify the proposal of a general method of authorship analysis. The method can then be tested to validate the theory or reformulate it.

In general, the method consists in the analysis of the known sets of texts for semantic features and then in the identification of those features that do not vary because of registerial reasons but because of codal reasons. In other words, the method consists in finding those linguistic variables that present more intra-author variation than intra-genre variation. If the theory is correct, those variables will represent how the social group(s) to which the writer belongs understand(s) or interact(s) with that particular genre.

Since this hypothetical method requires a set of linguistic variables, the ones that are most obviously suitable would seem to be the classic SFL ones, such as frequencies of types of transitivity, frequencies of types of mood, frequencies of types of theme and so forth. This, however desirable and optimal in theory, has turned out to be unpractical in recent pilot studies (Nini and Grant, 2013). There are two major reasons for this impracticality: (a) at the present time, there are no reliable parsers for SFL features; this implies that the analysis has to be carried out manually, which in turn implies that at times the analysis can be too subjective to be used for forensic purposes; (b) there is no knowledge of how SFL variables systematically vary in different contexts to allow the

analyst to understand to what extent the variation observed for a particular variable in a particular genre is distinctive or normal for that genre. This prerequisite for the variables is necessary as otherwise an assessment of the intra-genre variation is difficult to obtain. Mainly for these two reasons, the option considered in this paper is to introduce another framework that is compatible with codal variation theory as well as satisfying the two above-mentioned criteria. This framework is Biber's multidimensional analysis (Biber, 1988).

In a large scale experiment, Biber (1988) applied a multivariate statistical test called factor analysis to study how linguistic variables (such as frequency of past tenses, frequency of nouns, frequency of mental verbs) that are known to vary from register to register co-vary altogether to create functional orientations. Biber (1988) examined a general corpus representative of the most important genres of the English language and measured automatically 68 linguistic variables. Once the frequencies for these variables were calculated, they were arranged by the factor analysis along factors. The process consists in trying to explain the co-variation between these variables so that the variables that contribute to the same function can be grouped together. In this way it is possible to reduce a set of 68 variables to a more manageable smaller set of factors that can be interpreted for the linguistic function that they realise. In Biber (1988), the analysis generated six factors, which therefore created a six-dimensional space where the genres of the English language can be located. Mapping a genre on this space means calculating the score for each of the factors for each of the texts in a genre, finding the average for the genre and then comparing the figure obtained with the averages calculated for other genres.

Theoretically, the compatibility between this analytical framework and codal variation theory can be noticed in many points of overlap between the two. First of all, in both of these works, it seems evident that the authors start from a model that is based on an analysis of underlying functional orientations. Secondly, it is possible to notice that both Biber (1988) and Hasan (1990) employ a multivariate statistical analysis to find the semantic styles or functional orientations of some language varieties. Finally and most importantly, codal variation theory and the analysis presented in Biber (1988) are compatible with Finegan & Biber's (2001) *register axiom*. Finegan and Biber (2001), in a fashion very similar to Hasan's codal variation, postulate that different social groups possess different degrees of competence of different registers. This competence is formed by exposure to the registers and it therefore varies from social group to social group because different social groups are exposed to different registers.

These theoretical and practical advantages make Biber's multidimensional analysis a good candidate for transforming codal variation theory into a tool for authorship analysis. The method here proposed can be explained by looking at an example taken from a real case.

An example from a real case

The case examined in the present paper is an inclusion/exclusion attribution case where the analyst is provided with an email spreading malicious information (500 words) and four known emails (500 words in total) authored by the main suspect.

The first assumption that has to be met for the method to work is that the known set and the questioned set be compatible in terms of register. This can be assessed quali-

tatively, by looking at the recipient(s) and the medium, for example. It can also be tested quantitatively, by measuring Biber (1988) variables, calculating the factor scores, plotting the texts in the multidimensional space and finally verifying to what extent the texts analysed fall within the same genre. Both pieces of evidence can be collected to conclude whether the contextual comparability assumption is met.

If this assumption is met, it is reasonable to assume that the variation observed in the two sets is mainly given by the coding orientations of the authors, that is, on the way the authors of the questioned set and the known set interpreted the context in which they operated. In this particular inclusion/exclusion case, the question that is asked is: 'is the questioned set compatible with the known set?', which is a subset of the question: 'is there linguistic evidence for common authorship?'. Showing that the coding orientation of the author of the questioned set is compatible with the coding orientation of the known set is a piece of evidence that contributes to answering this question.

Using a notion introduced by Grant (2010), the coding orientation of one author could be thought of as those variables that are *consistently* and *distinctively* used in the author's texts when compared to the genre analysed. To find these variables, one can simply replicate Biber's (1988) study for the known set and questioned set and determine the values for each of the 68 variables, as well as the factor scores for each text. Furthermore, what is needed to validate these counts is a theoretical explanation of why these variables vary in accordance with codal variation theory.

In the case presented as example, after examining all the variables, including the factor scores, two variables were found that equally occurred in the known set and in the questioned set and that, in addition, occurred significantly more often than expected for the genre, that is, that showed *consistency* and *distinctiveness*. These variables were *sentence relatives*, defined as the normalised frequency of: <COMMA + which>, and *pied-piping relatives*, defined as the normalised frequency of: <PREPOSITION + (who|whom|whose|which)>. Without a base-rate knowledge of these variables, comparing the raw or normalised frequencies of these two variables between the questioned and known set and obtaining similar figures is not enough on its own to claim compatibility. In other words, if it is impossible to know what the normal frequency for these variables is, it is equally impossible to gather a piece of evidence to claim common authorship. Looking at Biber's (1988) analysis, the two variables considered present the following frequencies in the genre that is closest to emails, 'professional/personal letters':

Assuming that the emails examined for the case and the personal/professional letters used by Biber (1988) are compatible in their registers, a comparison of those features with the observed frequencies for those two variables for the known and questioned sets is reported below:

The difference between a typical personalprofessional letter and the two known and questioned sets is strikingly significant. Whereas in a typical letter it is normal to find about 1 instance of both variables every 1,000 words, in the known and questioned sets an average of 0.6 per 100 words is observed.

By confronting the observed frequencies with the frequencies expected for the genre it is possible to assess how much *distinctive* the known and questioned sets are when

Figure 2. Frequency for pied-piping relatives and sentence relatives. Frequencies per 1000 words.

	Linguistic feature	Mean	Minimum value	Maximum value	Range	Standard deviation
Biber (1988) professional/personal letters	pied-piping relatives	0.2	0.0	1.0	1.0	0.4
	sentence relatives	0.2	0.0	1.0	1.0	0.4

compared to the norm. The shared *distinctiveness*, given the fact that register has been controlled for, is a piece of evidence to corroborate the hypothesis of common authorship of the two sets.

Qualitatively speaking, an assessment of these two variables in the texts shows that the authors of the known and questioned sets use far more *pied-piping relatives* and *sentence relatives* because they employ a more convoluted and complex syntax full of sub-specification and interpersonal comments on previous sentences. This analysis allows for a quantitative estimation of the variation observed, as well as the possibility of qualitatively explaining the stylistic difference.

An objection to this analysis is that *pied-piping relatives* and *sentence relatives* are two variables that rarely occur in texts. They do not follow a linear distribution that increases with the size of the corpus and therefore plenty of data is needed to establish the typical distributions of these variables (Biber, 1993). Nonetheless, the theoretical qualitative explanation of the difference observed in the two samples corroborates the quantitative finding and compensates for the problem of non-linearity of these two variables.

The sample gathered by Biber (1988) is biased towards middle class educated writers and reflects the coding orientation of that social group. The convoluted syntactic style given by a high frequency of relatives seems to be stigmatised in high education, as it increases sentence length but focusing on clausal complexity rather than nominal complexity (Hunt, 1983). On the other hand, previous studies have shown that individuals with lower education background tend to use more clauses per sentence and more sentences per t-units (Hunt, 1971, 1983). In other words, the figures given above by Biber are skewed towards a particular social group and do not represent a normal distribution for the English language as a whole. However, having this knowledge of how these particular variables are distributed in terms of genre and social groups, that is, in terms of registerial and codal variation, can be useful in forensic cases. In this example, for instance, if the research on relatives and level of education is confirmed, there is reason to believe that both the author of the known set and the author of the questioned set belong to a low education level social group and this, in turn, can be used as a piece of evidence for the sets being produced by the same author.

Figure 3. Observed frequencies for pied-piping relatives and sentence relatives for the Q and K sets. Frequencies per 100 words.

	Variable	Observed frequency
Questioned set	pied- piping relatives	0.6
	sentence relatives	0.6
Known set	pied- piping relatives	0.2
	sentence relatives	1.4

Conclusions

In conclusion, it is proposed that codal variation theory can be a valid tool for forensic authorship analysis, provided that there is enough research on the variables for the genres that are typically involved in a forensic scenario. A method based on this theory is theoretically grounded in SFL but analytically based on Biber's (1988) multidimensional framework.

The example considered might seem just another application of corpus methods to a forensic case and this is indeed true. However, the significant difference with other case reports of this kind is that the corpus analysis carried out is entirely theory based and theory driven. The lack of theory in forensic authorship analysis might become a danger that could threaten the field. Without the knowledge that explains why a particular system or tool works, it is neither possible to be sure that the results given by this tool can be replicated in other cases apart from the experimented ones, nor is it possible to improve them. It is likely that the field can move forward only thanks to theories that can explain why a particular tool is useful and, most importantly, that generate hypothesis that can be validated or rejected in future research.

References

Argamon, S., Koppel, M., Pennebaker, J. W. and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), 119.

Bernstein, B. (1962). Linguistic codes, hesitation phenomena and intelligence. *Language and Speech*, 5(4), 221–240.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257.

Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8(1), 1–65.

Coulthard, M. and Johnson, A. (2007). *An Introduction to Forensic Linguistics*. London: Routledge.

Finegan, E. and Biber, D. (2001). Register variation and social dialect variation: The register axiom. In P. Eckert and J. R. Rickford, Eds., *Style and Sociolinguistic Variation*, 235–267. Cambridge: Cambridge University Press.

Grant, T. (2010). Txt 4n6: Idiolect free authorship analysis. In M. Coulthard and A. Johnson, Eds., *Routledge Handbook of Forensic Linguistics*, 508–523. London: Routledge.

Grant, T. and Baker, K. (2001). Identifying reliable, valid markers of authorship: a response to Chaski. *Forensic Linguistics*, 8(1), 66–79.

Halliday, M. A. K. and Matthiessen, C. M. I. M. (2004). *An Introduction to Functional Grammar*. London: Arnold.

Hasan, R. (1990). A sociolinguistic interpretation of everyday talk between mothers and children. In J. Webster, Ed., *The Collected Works of Ruqaiya Hasan Vol. 2: Semantic Variation: Meaning in Society and in Sociolinguistics*, 73–118. London: Equinox.

Hasan, R. (1996). Ways of saying: ways of meaning. In C. Cloran, D. Butt and G. Williams, Eds., *Ways of Saying, Ways of Meaning: Selected Papers of Ruqaiya Hasan*, 191–242. London: Cassell.

Hasan, R. (2009). Wanted: a theory for integrated sociolinguistics. In J. Webster, Ed., *The Collected Works of Ruqaiya Hasan Vol. 2: Semantic Variation: Meaning in Society and in Sociolinguistics*, 5–40. London: Equinox.

Hunt, K. (1971). Teaching syntactic maturity. In *Applications of Linguistics: Selected Papers of the Second International Congress of Applied Linguistics*, 287–301. Cambridge: Cambridge University Press.

Hunt, K. (1983). Sentence combining and the teaching of writing. In M. Martlew, Ed., *The Psychology of Written Language: Developmental and Educational Perspectives*, 99–125. New York: John Wiley.

Koppel, M., Argamon, S. and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), 401–412.

Martin, J. R. (1992). *English Text: System and Structure*. Philadelphia: John Benjamins.

Matthiessen, C. M. I. M. (2007). The "architecture" of language according to systemic functional theory: developments since the 1970s. In R. Hasan, C. Matthiessen and J. Webster, Eds., *Continuing Discourse on Language*, 505–561. London: Equinox.

Nini, A. and Grant, T. (2013). Bridging the gap between stylistic and cognitive approaches to authorship analysis using systemic functional linguistics and multidimensional analysis. *International Journal of Speech, Language and the Law*, 20(2), 173–202.

Rochester, S. and Martin, J. R. (1979). *Crazy Talk: a Study of the Discourse of Schizophrenic Speakers*. Norwood, N.J.: Ablex.

Using function words and punctuation marks in Arabic forensic authorship attribution

David García-Barrero, Manuel Feria and Maria Teresa Turell [*]

Abstract. *This paper presents an approach to written Modern Standard Arabic forensic authorship attribution drawn for the first time on a full tokenization of the text and based on the analysis of several variables such as type/token ratio, word length in characters, punctuation, conjunctions, the combination of punctuation and conjunctions, and the standard deviation of sentence length in words, among others. These variables have been tested in a sample corpus of three Moroccan writers producing two genres (short stories and literary reviews) in two measurement times. The hypotheses to be tested are: (a) There will be more inter-author than intra-author variation; (b) A writer's idiolectal style will stay stable throughout time; and (c) This idiolectal style will not be so stable when constrained by genre. The texts are segmented (tokenized) with TOKAN, part of MADA 3.2 (Morphological Analysis and Disambiguation for Arabic – CADIM group, Columbia University), based on Tim Buckwalter's Aramorph 1.2.1, which allows simplifying the detection of variables involving agglutinated clitics. Preliminary observations show that some of these variables may be discriminant markers for authorship attribution.*

Keywords: Arabic, authorship attribution, Arabic variation, inter- and intra-author variation, Arabic tokenization.

Introduction

This paper reports on the use of some quantitative techniques in authorship attribution applied to written Modern Standard Arabic (MSA). So far, several semi-automatic and quantitative approaches have been applied to authorship attribution of real-world and real forensic case texts. These approaches include linear discriminant analysis (Bel *et al.*, 2012), the use of reference corpora, Bayesian likelihood ratio methods (applied to both

---

[*]IULA, Universitat Pompeu Fabra; Departamento de Traducción e Interpretación, Universidad de Granada; and IULA, Universitat Pompeu Fabra

oral and written texts), and measurements such as lexical density analysis, among others (See Grieve 2007 and references therein for an evaluation of these techniques applied to authorship attribution.)

However, despite the prominence given to quantitative methods in languages such as English and Spanish, rather less attention has been paid to Arabic. To the best of our knowledge, previous literature on quantitative measurements applied to Arabic authorship attribution is limited to Abbasi and Chen (2005a,b, 2006); Estival *et al.* (2007) and more recently, Ouamour and Sayoud (2012) and Sayoud (2012).

With the purpose of bridging the gap between the state of the art in other languages and in Arabic authorship attribution studies, this article presents an Arabic-driven approach. This approach consists specifically in a quantitative analysis drawn, for the first time, on a full tokenization of the Arabic texts. For this purpose, all clitics have been segmented. Tokenization of clitics is of the utmost importance for two main reasons:

1. Clitics and affixes have an overwhelming statistical impact when dealing with Arabic. Using an Arabic corpus containing 600 million words, Alotaiby *et al.* (2010) showed that the corresponding lexicon size was reduced by 24.54% when applying clitic tokenization produced by AMIRA 2.0 (Habash, 2010: 89) on Arabic Gigaword (Graff, 2007).

2. Among the clitics we can find the three one-letter non-derived prepositions (Ryding, 2005: 366 and following); the ubiquitous conjunctions و $w\#^2{}_1$ and ف $f\#$, which are the scaffolding of the Arabic text (Warraki and Hassanein, 1994); the even more ubiquitous definite article ال $Al\#$, pronouns and +k. We must take into account as well that several clitics can appear together in one word. All these function words are crucial for any quantitative approach to Arabic authorship attribution.

Table 1 shows clitics tokenized by Alotaiby *et al.* (2010) on Arabic Gigaword (Graff, 2007)[2].

On the other hand, 'there was no Western-style punctuation in Classical Arabic [...]. In general, the coordinating conjunctions and discourse markers served as punctuation [...] Modern Written Arabic has adopted[3] and adapted Western punctuation, without abandoning certain features of the Classical Arabic system (especially noticeable in co-ordination)'

This fact, which poses a major problem for text segmentation (Ameur *et al.*, 2008), teaching translation from Arabic into Western languages (Ghazala, 2004) and (Dendenne, 2010) and teaching English as a L2 for target learners whose L1 is Arabic (Awad, 2012) can be efficiently exploited in authorship attribution studies, especially considering punctuation and conjunctions together. This kind of analysis can give us insight into variation in discourse structure, dialectal and diachronic variation in written MSA.

For the purpose of this study, sequences such as a comma followed by a copulative conjunction, or by any other connector, can be measured in an n-gram based approach taking characters, graphic words or tokens, or parts of speech (POS) as basic units. We assume that punctuation marks and conjunctions operate in MSA at the same discursive level and define an Arabic token as 'a space delimited unit in clitic tokenized text' (Diab *et al.*, 2004). Tokenization also solves, to a great extent, the homography issue, which has an enormous impact when dealing with Arabic, as explained below.

Table 1. Clitics' Statistics according to Alotaiby *et al.* (2010)

Clitics	Transliteration	Frequency	%
ال	Al#	103,015,016	57.02
و	w#	31,014,498	17.17
ل	l#	11,470,854	6.35
ه	+h	10,021,855	5.55
ب	b#	8,857,080	4.9
ها	+hA	7,975,714	4.41
هم	+hm	2,264,578	1.25
ف	f#	1,462,196	0.81
س	s#	1,348,962	0.75
نا	+nA	1,144,143	0.63
ك	k#	737,638	0.41
هما	+hmA	449,528	0.25
ي	+y	302,246	0.17
ك	+k	267,685	0.15
ني	+ny	147,805	0.08
كم	+km	125,782	0.07
هن	+hn	35,215	0.02
ا	+A	21,342	0.01
كما	+kmA	5,611	0
كن	+kn	1,057	0

Objectives

This work is part of an on-going research project whose purposes are as follows:

1. To validate in written MSA hypotheses in authorship attribution already verified for other languages such as English, Spanish or Catalan.
2. To shed some light on which variables are potentially discriminant for Arabic authorship attribution.
3. In the long term, this research aims for the establishment of a reliable attribution methodology that makes it possible to conduct scientifically rigorous forensic text comparison in Arabic.

Hypotheses

Attribution studies rely on the notion of idiolectal style, as defined by Turell (2010). This means that each author or speaker has his/her own style that can be measured and characterizes his/her use of language. The main hypothesis that has already been tested for other languages is that:

1. For certain linguistic variables, there is more inter-author variation than intra-author variation, i.e. more variation in texts that have been written by different authors than in texts written by the same author.

     There are obviously other factors affecting variation within texts written by the same author, such as communicative purposes, register, genre and the span of time separating two samples of writing. Therefore, the above-mentioned leading hypothesis can be further elaborated into two more specific null hypotheses:

2. For the same variables, there is less variation between authors writing in the same genre (inter-author intra-genre) than between texts of two different genres written by the same author (intra-author inter-genre), and

3. For the same variables, there is less variation between authors writing within a short period of time (inter-author intra-time) than within the same author in two sufficiently separated measurement times (intra-author inter-time).

Corpus

1. In order to avoid diatopic variation, 3 Moroccan authors produced the writing samples: أحمد المديني Ahmed Al-Madini (author 1), محمد عز الدين التازي Mohamed Azzedine Tazi (author 2) and محمد برادة Mohamed Berrada (author 3). The rationale for choosing Morocco is no other than the authors' familiarity and the interest of the Spanish scholars in the country for obvious geostrategic reasons. We expect to be able to increase the number of authors in the future.

2. In order to avoid diastratic variation and context-based variation in the broad sense, the selected authors belong to the same educational level. The samples were randomly selected from comparable reputed sources, such as monographs (compilations of short stories or literary criticism articles of an author or several authors published by reputed publishers) and national newspapers (mainly the weekly cultural supplements of العلم Al-Alam and الاتحاد الاشتراكي Al-Ittihad al-Ichtiraki). Quotes (in literary criticism articles) and dialogs (in short stories), when present, were removed from the analysis.

3. In order to control inter-genre and inter-author variation, the selected texts include no quotations and the corpus was stratified according to genre into two sets: short story, a genre strongly related to narrative discourse, and literary criticism, a genre strongly related to expository-argumentative discourse. We use here the term *genre* in the traditional, literary sense (and not in the narrower sense defined by Swales 1990).

4. In order to control inter-time variation, we have included samples in two measurement times with a time lapse of nine or more years for each author and each genre.

We have collected 5 equally sized samples for each author, genre and time, which make up a total of 60 samples. Dealing with equally sized samples allows us to avoid text-length dependency problems in this stage of the research process. The sample size has been determined in approximately 650 graphic words (always including complete sentences). This Arabic-driven sample size has been calculated adapting the English 800 words standard, considered an average for forensic real-world texts, by taking into account the agglutinative character of the Arabic script in the light of the conclusions drawn by Alotaiby *et al.* (2010).

Processing

The texts were processed using MADA 3.2, a tool developed by Columbia's Arabic Dialect Modelling Group (CADIM), at Columbia University. This tool provides disambiguation, lemmatization, POS tagging and full tokenization for Arabic.

MADA is based on the information provided in BAMA (Buckwalter's Arabic Morphological Analyzer), or its latest version SAMA 3.1 (Standard Arabic Morphological Analyzer). We have used a free version of BAMA: Aramorph 1.2.1,[4] that allows MADA

to be run since its newest release (MADA 3.2) in February 2012. The free access to this tool represents a noteworthy step forward in Arabic NLP and in Arabic linguistics.

In this research, only TOKAN (the tokenization module of the MADA tool) has been used. In Table 2, we show an example of MADA's capabilities, with an input of a sequence of two Arabic *graphic words* that are split into core words and clitics in TOKAN. MADA also provides POS information on each token.

Table 2. MADA-TOKAN example

Input	وخصوصياتها wxSwSyAthA			العربية AlErbyp	
Segmentation (TOKAN)	‍ها +hA	خصوصيات xSwSyAt	و w#	عربية Erbyp	ال‍ Al#
Part of speech	Pronoun (enclitic)	Noun	Conjunction (proclitic)	Noun	Def. article (proclitic)

Linguistic variables

Tables 3 to 9 describe the linguistic variables used in the analysis: definite article and pronouns (3), prepositions (4), conjunctions (5), demonstratives (6), negative particles (7), punctuation marks (8) and the combination of punctuation marks and conjunctions (9). Despite the fact that the size of the samples has been strictly controlled, data, with the exception of ratios, are expressed as relative frequencies (variable / total number of tokens). In addition to the frequencies of specific words, several combinations have been taken into account, namely sums of words of the same part of speech or sharing similar features, and ratios measuring the relationship between tokens (or sums of tokens) of related parts of speech or features. For instance, we have linked definite articles and enclitic pronouns (variable [6]), since they are mutually exclusive when occurring with nouns; we have also linked the substantive subordinating conjunction and the relative pronouns [47] in order to measure the substantive and relative subordinations.

The variables shown above are not all possible Arabic function words, but only those that had a considerable frequency in the text samples. For that reason, we have only considered third person singular masculine and feminine pronouns and demonstratives.

After the tokenization process, the ambiguity of the Arabic script is solved to a great extent: most of the output tokens represent disambiguated, specific words and parts of speech. However, some ambiguity still affects two important tokens due to homography:

1. The lack of short vowels affects من *mn* [25], which has been considered here only as a preposition ('of', 'from'), but which can also stand for the relative *wh*-pronoun ('who', 'whom'). This kind of ambiguity will be solved in the future by simply introducing short vocalization: مِن *min* (preposition), مَن *man* (pronoun). The preposition لِ *li* [30] has also a counterpart verbal particle لَ *la*, though it is very unlikely to be tested in MSA.

2. In a much lower degree of ambiguity, orthographic variation affects أن *>n* [46], that can stand for أنْ *>ano* (similar to the infinitive particle 'to') and أنَّ *>an a*

Table 3. Variables: Definite article and pronouns

No.	Variable	Description
1	ال	Proclitic definite article *Al#*
2	ـه	Enclitic pronoun 3rd p. sg. masc *+h*
3	ـها	Enclitic pronoun 3rd p. sg. f. *+hA*
4	$\frac{ـه}{ـها}$	Ratio enclitic pronouns 3rd p. sg. m. / f.
5	ـه + ـها	Total enclitic pronouns 3rd p. sg.
6	$\frac{ال}{ـه+ـها}$	Ratio definite article / total enclitic pronouns [1, 5]
7	هو	Independent pronoun 3rd p. sg. m. *hw*
8	هي	Independent pronoun 3rd p. sg. f. *hy*
9	هو + هي	Total independent pronouns 3rd p. sg.
10	ـه + ـها + هو + هي	Total enclitic + independent pronouns [5, 9]
11	$\frac{ـه+ـها}{هو+هي}$	Ratio enclitic / independent pronouns [5, 9]
12	ـه + هو	Total m. pronouns [2, 7]
13	ـها + هي	Total f. pronouns [3, 8]
14	$\frac{ـه+هو}{ـها+هي}$	Ratio m. / f. pronouns [12, 13]
15	الذي	Relative m. pronoun *Al*y*
16	التي	Relative f. pronoun *Alty*
17	الذي + التي	Total relative pronouns m. + f.
18	ـه + هو + الذي	Total m. pronouns [2, 7, 15]
19	ـها + هي + التي	Total f. pronouns [3, 8, 16]
20	$\frac{ـه+هو+الذي}{ـها+هي+التي}$	Ratio m. / f. pronouns [18, 19]
21	ـه+ـها+هو+هي+الذي+التي	Total pronouns (enclitics + independents + relatives) [5, 9, 17]

('that'), both considered here as substantive subordinating conjunctions. Furthermore, the first letter أ > can be spelled as ا *A*, which is a non-normative wide-spread orthographic variant of both أ > and إ < (see Parkinson 1990 and Buckwalter 2004, and therefore MADA+TOKAN normalizes both into ا *A* in the preprocessing. Thus, in this case, أن *>n* is a homograph for إن *<n*, which itself stands for إنْ *<ino* (conditional subordinating conjunction) and إنَّ *<in a* (emphasis particle).

On the other hand, some function words have not been considered in this particular study due to their structural lexical ambiguity, such as ما *mA*, which stands for a negative particle or a relative pronoun. Interestingly, several highly ambiguous words appear combined in compound function words. Testing if discriminatory potential increases with increasing disambiguation provides an important direction for future research.

Table 4. Variables: Prepositions

No.	Variable	Description
22	بـ	Proclitic preposition $b\#$
23	في	Preposition fy
24	مع	Preposition mE
25	من	Preposition mn
26	عن	Preposition En
27	على	Preposition ElY
28	كـ	Proclitic preposition $k\#$
29	إلى	Preposition $<lY$
30	لـ	Proclitic preposition $l\#$
31	إلى + لـ	Total 'dative' prepositions $<lY + l\#$
32	بـ+في+مع+من+عن+على+كـ+إلى+لـ	Total prepositions [22−30]
33	$\dfrac{إلى}{لـ}$	Ratio 'dative' prepositions $<lY/l\#$
34	$\dfrac{preposition}{بـ+في+مع+من+عن+على+كـ+إلى+لـ}$	Ratio preposition / total prepositions [22−31, 32]

Table 5. Variables: Conjunctions

No.	Variable	Description
35	و	Proclitic copulative conjunction $w\#$
36	فـ	Proclitic copulative conjunction $f\#$
37	ثم	Copulative conjunction vm
38	و + فـ + ثم	Total copulative conjunctions [35−37]
39	أو	Disjunctive conjunction $>w$
40	و + فـ + ثم + أو	Total coordinating conjunctions [35−37, 39]
41	$\dfrac{و}{و+فـ+ثم}$	Ratio $w\#$ / copulative conjunctions [35, 38]
42	$\dfrac{و}{و+فـ+ثم+أو}$	Ratio $w\#$ / coordinating conjunctions [35, 40]
43	$\dfrac{فـ}{و}$	Ratio $f\#$ / $w\#$ [35, 36]
44	$\dfrac{فـ}{و+فـ+ثم}$	Ratio $f\#$ / copulative conjunctions [36, 38]
45	$\dfrac{فـ}{و+فـ+ثم+أو}$	Ratio $f\#$ / coordinating conjunctions [36, 40]
46	أن	Subordinating conjunction $>n$
47	$\dfrac{أن}{الذي+التي}$	Ratio $>n$ / relative pronouns [46, 17]

Table 10 shows additional descriptive data used, such as type-token ratio, standard deviation of sentence length in words and word-length distribution (i.e. frequencies of the total of words of n-characters). All these measurements have been provided automatically by WordSmith analysing both raw and tokenized texts, with the exception of variable [79] (ratio of tokens in raw and tokenized texts). Except for the n-character

Table 6. Variables: Demonstratives

No.	Variable	Description
48	هذا	Proximal demonstrative m. sg. $h*A$
49	هذه	Proximal demonstrative f. sg. $h*h$
50	ذلك	Distal demonstrative m. sg. $*lk$
51	تلك	Distal demonstrative f. sg. tlk
52	هذا + هذه + ذلك + تلك	Total demonstratives [48−51]
53	$\frac{هذا+هذه+ذلك+تلك}{sentences}$	Ratio demonstratives / sentences [52, 81]

Table 7. Negative particles

No.	Variable	Description
54	لا	Negative particle lA
55	لم	Negative particle lm
56	لا + لم	Total negation [54, 55]
57	$\frac{لا+لم}{sentences}$	Ratio negative particles / sentences [56, 81]

Table 8. Punctuation marks

No.	Variable	Description
58	.	Single dot (stop)
59	..	Two dots (ellipsis)
60	...	Three dots (ellipsis)
61	.. + ...	Total ellipsis [59, 60]
62	. + .. + ...	Total stop + ellipsis [58, 61]
63	،	Comma
64	. + .. + ... + ،	Total stop + ellipsis + comma [62, 63]
65	$\frac{،}{.}$	Ratio comma / stop [63, 58]
66	$\frac{.}{sentences}$	Ratio stop / sentences [58, 81]
67	$\frac{،}{sentences}$	Ratio comma / sentences [63, 81]
68	¶	Newline
69	$\frac{.}{¶}$	Ratio stop / newline [58, 68]
70	$\frac{،}{¶}$	Ratio comma / newline [63, 67]
71	"	Quotation marks (double quotes)

word frequency [84], that shows significantly different data on raw and tokenized texts for obvious reasons, these measurements appear fairly stable when their relative frequencies are obtained dividing the variable in question by either the number of 'tokens' in the raw or in the tokenized texts.

Table 9. Punctuation marks and conjunctions

No.	Variable	Description
72	و ،	String comma + copulative conjunction , $w\#$
73	$\frac{و ،}{،}$	Ratio string , $w\#$ / comma [72, 63]
74	و .	String dot + copulative conjunction , $w\#$
75	$\frac{و .}{.}$	Ratio string . $w\#$/ dot [74, 58]
76	و . + و ،	Total strings [72, 74]
77	$\frac{و . + و ،}{و}$	Ratio strings / copulative conjunction $w\#$ [76, 35]
78	$\frac{و . + و ،}{. + .. + ... + ،}$	Ratio strings / stop + ellipsis + comma [76, 64]

Table 10. Descriptive data (WordSmith)

No.	Variable	Description
79	$\frac{tokens_{sg}}{tokens_{or}}$	Ratio tokens in segmented / original texts
80	$\frac{types}{tokens}$	Type-token ratio
81	$sentences$	Number of sentences
82	$\frac{tokens}{sentences}$	Sentence mean length in words (ratio tokens / sentences)
83	$\sigma(tokens \in sentence)$	Standard deviation of sentence length in words
84	$n-character\ words$	Words with n-characters (letters)

Analysis

In order to test the performance of the measurements, an analysis of variance (ANOVA) and several linear discriminant analyses (LDA) were conducted.

The degree of relation between the variables and the groups of authors, genres and times was established by means of an ANOVA test. Table 11 shows the set of ten variables most strongly related to authorship. Within these top ten variables, the most recurrent are combinations of punctuation marks and conjunctions.

Furthermore, the most discriminant variables for Arabic authorship attribution selected by the LDA are:

1. Combinations of punctuation marks and conjunctions:
   (a) Number of the copulative conjunction و $w\#$ occurring after a comma [72] or a dot [74].
   (b) Sum of the preceding variables divided by the total number of و $w\#$ [77].
2. Punctuation marks:
   (a) Number of commas divided by number of newlines (i.e. average of commas in each paragraphs) [67].
   (b) Number of stops divided by number of sentences (i.e. how much sentences end with a dot) [66].
   (c) Number of ellipses (made of two or three following dots) [61].

Table 11. The 10 variables most strongly related to authorship according to ANOVA

| No.	Variable	Description	p-value
80	$\frac{types}{tokens}$	Type-token ratio	0,00000000007
72	، و	String comma + copulative conjunction ،$w\#$	0,00000003485
63	،	Comma	0,00000008835
77	$\frac{، + و .}{و}$	Ratio strings / copulative conjunction $w\#$	0,00000022231
9	هو + هي	Total independent pronouns 3rd p. sg.	0,00000028884
76	، . + و	Total strings	0,00000078888
73	$\frac{، .}{،}$	Ratio string ،$w\#$ / comma	0,00000174227
84	$3-character\ words$	Words with 3-characters (in raw texts)	0,00000750034
23	في	Preposition fy	0,00001525503
84	$2-character\ words$	Words with 2-characters (in raw texts)	0,00001540769

(d) Quotation marks [71].

3. Other copulative conjunctions: ثم vm [37] and the ratio of ف fa / و $w\#$ [43]. Interestingly, the former variable is a discriminating variable in text samples of literary criticism articles and the latter in text samples of short stories.

4. In addition, data obtained automatically using WordSmith also proved to be discriminatory in these analyses: the type-token ratio [80], some n-character words [84] (namely 2- and 3-character words in the raw samples) and the standard deviation of sentence length in words [83].

5. Less importantly, other function word frequencies were selected, such as:

   (a) The subordinating conjunction أن $>n$ [46].

   (b) The preposition على ElY [26] and its relative frequency with respect to the rest of the prepositions [34].

   (c) The negative particle لم lm [55], the total of negative particles لا lA and لم lm [56] and its relative frequency with respect to the number of sentences [57].

The first test classified the whole corpus by authors with a high accuracy of 59 out of 60 correct assignations using 12 variables (Figure 1). The same test obtained 56 out of 60 in the cross validation classification, in which every sample text is considered as a 'disputed' text to be assigned to an author. The variables were selected using default values of F (3,84–2,71). To gain reliability, we have modified the F values to decrease the number of variables selected, following the norm exposed by Poulsen and French (2004):

> As a "rule of thumb", the smallest sample size should be at least 20 for a few (4 or 5) predictors. The maximum number of independent variables is n - 2, where n is the sample size. While this low sample size may work, it is not encouraged, and generally it is best to have 4 or 5 times as many observations and independent variables.

A summary of the results with default and modified F values with the original and cross validation classifications for this test and the following ones is given in Table 11.

Genre is hypothesized to be another source of variation, as stated before. Therefore, the test was repeated after dividing the corpus by genre. Every single literary criticism

Figure 1. Classification by author (60 samples)
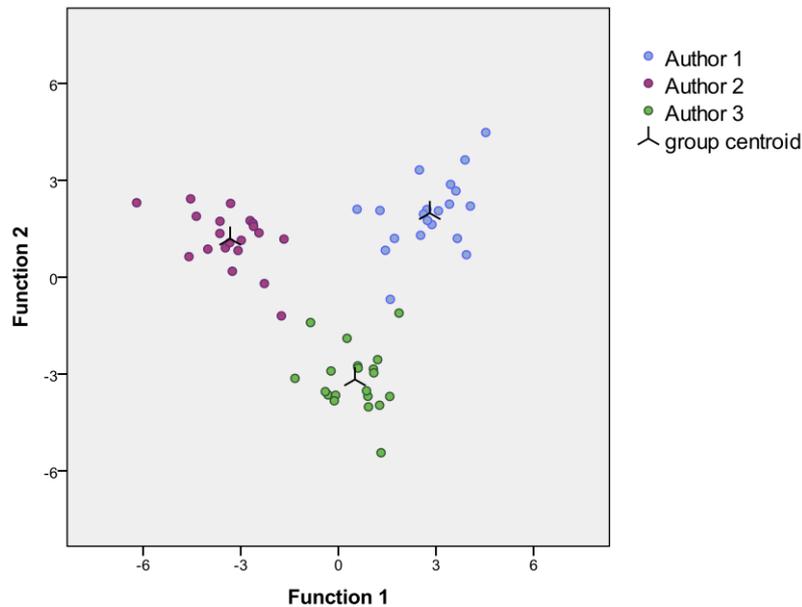
sample was correctly assigned to its author with 10 variables selected (Figure 2); only 1 sample out of 30 was misclassified in the cross validation classification after decreasing the number of variables to 5. Similar results were obtained with short story samples (Figure 3). No interference of the variable "time" was observed in either case.

Figure 2. Classification by author (30 samples of literary criticism articles)
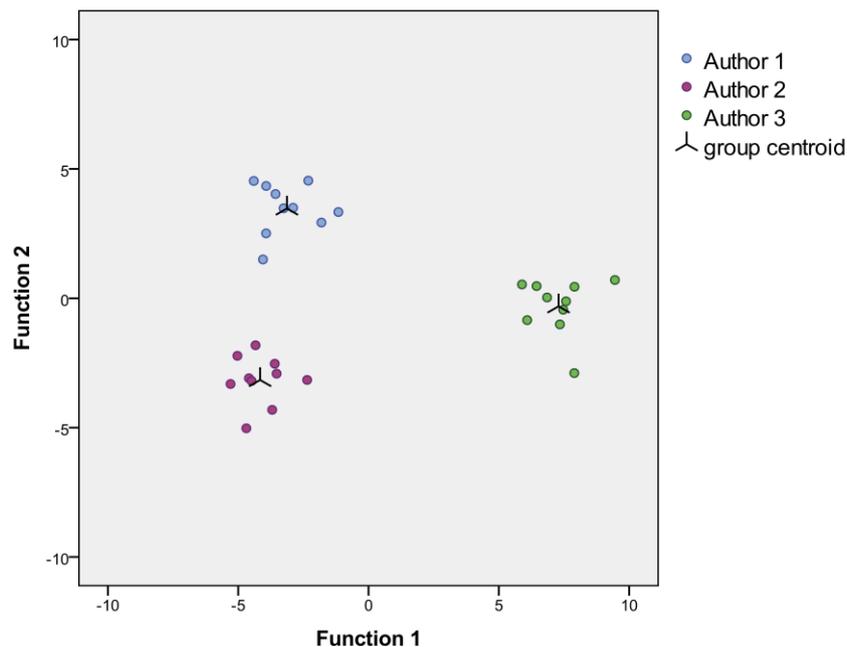
Figure 3. Classification by author (30 samples of short stories)
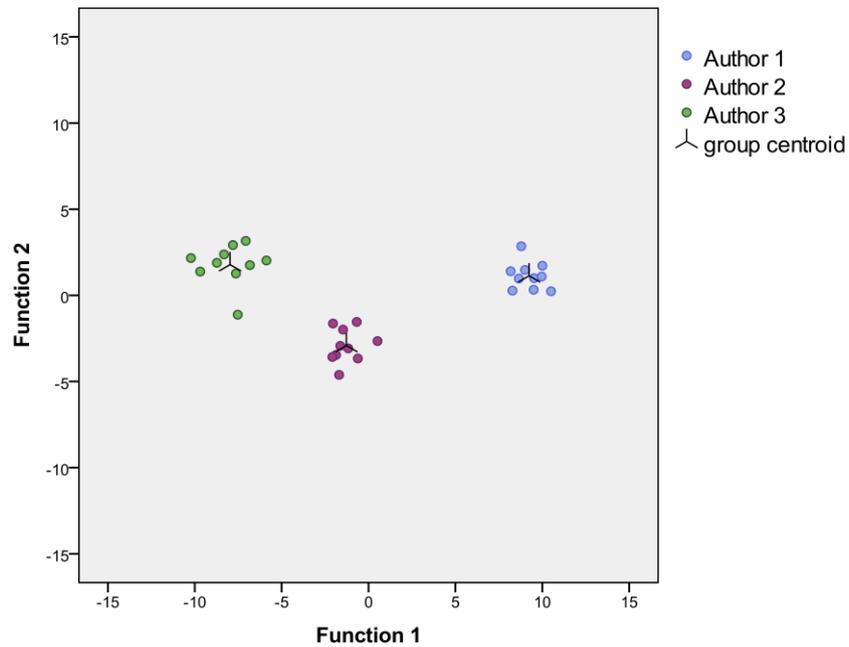
Table 12. Classification results

	F values	Variables selected	Original classification accuracy	Cross validation accuracy
60 samples	3,84-2,71 (default)	12	98,3% (1 error)	93,3% (4 errors)
	5,5-4,5	9	96,7% (2 errors)	95% (3 errors)
	7-6	7	98,3% (1 error)	95% (3 errors)
	8-7	5	93,3% (4 errors)	85% (9 errors)
30 samples of literary criticism articles	3,84-2,71 (default)	10	100%	100%
	5,5-4,5	5	100%	96,7% (1 error)
30 samples of short stories	3,84-2,71 (default)	11	100%	100%
	5,5-4,5	5	100%	96,7% (1 error)

Conclusions and future research

In conclusion, we have obtained very positive results at a not very much sophisticated disambiguation level. As for the hypotheses, we can reject our null hypotheses and consider proved that it is possible to attribute authorship using discrete variables also in Arabic texts. Genre showed to be an influential independent variable, which has to be controlled in the attribution process, whereas time showed no interference on it.

In our view, this research is a pioneer work that contributes to Arabic authorship analysis in particular, and to forensic authorship attribution in general. Moreover, it contains theoretical and methodological proposals that contribute to Arabic linguistics.

Future research will explore other measurements to cover a wider range of variables. Further advantage will be taken of different possibilities of disambiguation, such as the POS tagging provided by MADA. This would let us apply the POS n-gram approach that has successfully been applied to other languages at ForensicLab (the forensic linguistics laboratory at the Institut Universitari de Linguistica Aplicada of the Universitat Pompeu Fabra). Finally, all tests will be conducted in the future using a larger corpus of Moroccan authors and a reference corpus of writers with different national backgrounds.

Acknowledgments

Notes

[1]We use Buckwalter (2002) transliteration, except when it comes to proper names. For clitic boundaries, we use the 'tatweel' elongation character (-) for the Arabic ligated clitics (all except ر *w#²*.), and in transliteration, following Alotaiby *et al.* (2010) and others, number sign () for proclitic boundary and plus sign (+) for enclitic boundary.

[2]The authors consider س s and ـا +A as clitics, although س s is rather an inflectional morpheme indicating future time in the verb, and ـا +A a case morpheme that could also be considered as a derivational morpheme turning an adjective into an adverb. Unfortunately, adding all the percentages presented by Alotaiby *et al.* (2010) the result is over 100%.

[3]Its use was first introduced in 1911 by أحمد زكي باشا Ahmed Zaki Basha in his book الترقيم وعلاماته في اللغة العربية Altrqym wElmAth fy Allgp AlErbyp (Punctuation and its Marks in Arabic Language).

[4]According to MADA developers, 'In our tokenization tests, an Aramorph MADA build reproduced the same tokenization as a SAMA MADA build for 99.4% of the words tested'. Source: `https://lists.cs.columbia.edu/pipermail/mada-users/2012-February.txt`

References

Abbasi, A. and Chen, H. (2005a). Applying authorship analysis to arabic web content. *Intelligence and Security Informatics*, 75–93.

Abbasi, A. and Chen, H. (2005b). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5), 67–75.

Abbasi, A. and Chen, H. (2006). Visualizing authorship for identification. *Intelligence and Security Informatics*, 60–71.

Alotaiby, F., Foda, S. and Alkharashi, I. (2010). Clitics in arabic language: a statistical study. *Proceedings of Pacific Asia Conference on Language, Information and Computation*, 24, 595–602.

Ameur, A. T., H., M. and W., A.-S. (2008). Semantic-Based Segmentation of Arabic Texts. *Information Technology Journal*, 7, 1009–1015.

Awad, A. (2012). The most common punctuation errors made by the English and the TEFL majors at An-Najah National University. *An-Najah University Journal for Research (Humanities)*, 26, 1.

Bel, N., Queralt, S., Spassova, M. . and Turell, M. (2012). The use of sequences of linguistic categories in forensic written text comparison revisited. In S. Tombilin, N. MacLeod, R. Sousa-Silva and M. Coulthard, Eds., *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*, 192–209, Birmingham, UK: Aston University.

Buckwalter, T. (2002). Arabic transliteration.

Buckwalter, T. (2004). Issues in Arabic orthography and morphology analysis. In *Proceedings of the COLING 2004 Workshop on computational approaches to Arabic script-based languages*, 31–34.

Dendenne, B. (2010). *The Translation of Arabic Conjunctions into English and the Contribution of the Punctuation Marks in the Target Language. The Case of Wa, Fa and Thumma in Modern Standard Arabic.* Dissertation submitted in partial fulfilment of the requirements for a master degree in applied language studies, Mentouri University, Constantine, Algeria.

Diab, M., Hacioglu, K. and Jurafsky, D. (2004). Automatic tagging of Arabic text: From raw text to base phrase chunks. In *5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04)*, Boston, MA.

Estival, D., Gaustad, T., Pham, S. B., Hutchinson, B. and Radford, W. (2007). Tat: an author profiling tool with application to Arabic emails. In *Proceedings of the Australasian Language Technology Workshop*, 21–30.

Ghazala, H. (2004). Stylistic-semantic and grammatical functions of punctuation in English-Arabic translation. *Babel*, 50(3), 230–245.

Graff, D. (2007). *Arabic Gigaword Third Edition*. Philadelphia: Linguistic Data Consortium.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3), 251–270.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Toronto: Morgan and Claypool Publishers.

Ouamour, S. and Sayoud, H. (2012). Authorship attribution of ancient texts written by ten Arabic travellers using a SMO-SVM classifier. In *Communications and Information Technology (ICCIT), 2012 International Conference*, 44–47: IEEE.

Parkinson, D. B. (1990). *Orthographic Variation in Modern Standard Arabic: The Case of ' the Hamza*, volume Perspectives on Arabic Linguistics II: Papers from the Second Annual Symposium on Arabic Linguistics. John Benjamins Publishing Company, 72 edition.

Poulsen, J. and French, A. (2004). Discriminant function analysis (DA).

Ryding, K. C. (2005). *A reference grammar of modern standard Arabic.* New York: Cambridge University Press.

Sayoud, H. (2012). Author discrimination between the Holy Quran and Prophet's statements. *Literary and Linguistic Computing*, 27(4), 427–444.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Turell, M. T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *International Journal of Speech, Language and the Law*, 17(2), 211–250.

Warraki, N. N. a. and Hassanein, A. T. (1994). *The Connectors in Modern Standard Arabic.* Cairo: The American University in Cairo Press.

Complex Research on speech interference characteristics

Firuza Bayramova [*]

Abstract. *In the research we tried and systemized the peculiarities of Russian speech by Azerbajanians which could help to identify an unknown Azerbaijanian speaker. We analyzed perceptionally and acoustically interfered Russian speech with Azerbaijani accent and using a number of programme and mathematic methods made a list of characteristics typical of Azerbaijanian accent and their relative weight (significance).*

Keywords: Speaker identification, Russian, Azerbaijanian.

Introduction

Experts working with forensic expertise of audio records often need to analyse an interfered speech (i.e. speech with residual dialectic or foreign occurrences) of an unknown speaker with the purpose of identifying or diagnosing the speaker on the basis of his voice and speech. To solve such tasks, the expert linguist needs a description of how the given foreign accent or dialect sounds when its speaker is speaking Russian, an audio sample for it, and a possibility to check if the deviations from the Russian standard pronunciation found in the speaker's speech constitute a system of characteristics inherent in a regional variety of the language.

In the linguistic and special expert literature there is a number of works on this topic: (see Bondarko and Verbitskaya, 1987; Erofeyeva, 1997; Galyashina and Khurtilov, 1991; Galyashina and Goloshchapova, 2004 for more detail). There are also specialized software packages ("Территория" Territory, "Интерференция" Interference, "Регион" Region, "Этнос" Ethnic Group). The base for all research in this area certainly consists of, first, the comparison of language systems and, secondly, studies of the peculiarities of their interaction.

But the available data are still insufficient for a proper analysis of the interfered Russian speech. First of all, there are no works describing a system of characteristics of an interfered speech enabling the expert to match it up with this or that accent or type of accent.

---

[*]Ministry of Justice of the Russian Federation

In this research, we attempted to systematize deviations from the Russian standard pronunciation in Russian spoken by the Azerbaijani and establish the set of deviations which would enable us to say, more or less confidently, that the characteristics of an unknown speaker's accent correspond to a system of characteristics of the Azerbaijani speech. We have analysed interfered Russian speech with Azerbaijani accent, performed perceptual and acoustical analysis of deviations from the Russian norm, and, by means of software (STATISTICA 6.0, Microsoft Excel, MS Visual Fox Pro 9.0) and mathematical methods (standard deviation, average dilatation, correlation analysis etc.), have established a set of characteristics which are peculiar for the Azerbaijani accent (as a single set), and the relative significance of each of them. On the basis of the obtained selection of Azerbaijani speakers we have obtained a stable function of the characteristics inherent in this particular accent. Thereafter, we created a program for analysis of interfered speech (for the purpose of determining if an unidentified accent can be Azerbaijan). This function is stable only if the characteristics have been identified correctly.

Theoretical basis and terminology

Interfered speech is a subject of a separate linguistic research. An interfered Russian speech is a Russian speech containing characteristics of a dialect or a foreign accent. As the term itself suggests, this type of speech emerges as a result of contamination of standard speech and spheres of speech which directly influence it in the course of their interaction. One of the main criteria of interfered speech as a separate sociolinguistic phenomenon is that it is marked and recognizable, more or less, from the point of view of the ordinary linguistic consciousness of a speaker of the standard language.

Interfered speech is a rather intricately organized complex phenomenon involving different levels of language one way or another. It has many characteristics distinguishing it from standard speech. But deviations are most consistently manifested on the phonetical level. This is accounted for by the fact that phonetical skills are automated and cannot always be corrected. Standard pronunciation is also included in interfered speech, as the interfered characteristics can be present in a foreigner's Russian pronunciation or be absent altogether. Besides, some accent characteristics can be similar in different types of accent, which also shows the necessity for a detailed description of different types of accent with isolation of a complex of national and diagnostic features. That being said, the unit of study in this research are the interfered phonetical characteristics of Russian, and the unit of description is the set of identification characteristics of a given accent in the Russian speech.

Phonetical accent requires a complex analysis combining the methods of articulatory, acoustic, and perceptual phonetics. Our research is based on data obtained in the course of perceptual analysis. These data were then studied using the methods of acoustic analysis (spectrographic analysis, fundamental frequency analysis). The possibilities and advantages of perceptual and acoustic analysis make them the primary methods in voice and speech expert studies.

Reference units in the study of interfered speech are *soundclass* and *rhythmic word structure*, or RWS (rhythmic pattern, an analogue to phonetical word).

A **soundclass** is a group of elements corresponding to identical perception of non-simultaneous segments of speech with some quality differences in their acoustic and articulatory properties (Galyashina and Zlatoustova, 1999). The *interfered soundclasses*

we are considering participate in the formation of an accent in the speech of a non-native speaker. An **RWS** is a whole phonetical unit of the plane of realisation ensuring that the flow of speech is divided into words. Its realization depends on the position within an utterance and the speaker's belonging to one of the types of the pronunciation standard or to a dialect. The Russian RWS has its own peculiarity: the quality of sounds is determined by the position within the word in relation to stress, and a peculiar division into prosodic core (the stressed syllable, sometimes the stressed syllable plus the first pretonic syllable) and the surrounding syllables, both pretonic and post-tonic. The correlation of their quality in the phonetical word is what creates the RWS which is typical for modern Russian. The presence of reduced sounds and voiceless vowels is aimed at preservation of the RWS in the flow of speech. For the Russian language, the RWS is defined by such indicators as the number of syllables in a phonetical word/stressed syllable (Zlatoustova, 1975, 1994). The RWS in accented speech is subject to interfering influence of another language system.

The RWS forms the syntagma's rhythmical frame. In the flow of speech the RWS and the syntagma can be subject to different changes. It has been established that the types of RWS can be significantly deformed under the influence of its place in the intonation contour. *Intonation contour*, in its turn, is defined as the structural unity of intonation means employed to connect rhythmical structures in the utterance and divide the flow of speech into separate utterances.

Research

The research has demonstrated that all types of deviations in the interfered Russian speech are connected with the skills obtained in the course of mastering the Russian language, and shall manifest themselves to a variable degree of frequency, depending on the level of proficiency in Russian. Some analogous data were got by other researchers during their studies on the Russian speech speaking by non native speakers (see, for example, Barkhudarova, 2008; Rogoznaya, 2001).

For this research we used authentic records submitted for audio expertise. We deliberately decided not to limit ourselves with an experiment (both experimentally devised recordings and authentic real-life segments have been used). We reasoned that in this case, the research would benefit from employing the method of modelling not on the basis of preset rules, but on observed types of deviations (see Vinarskaya and Zlatoustova, 1977: 103) for a similar choice of method of modelling).

For our analysis we have chosen 50 speakers whose speech shows Azerbaijani accent. On the phonetical level, their accent manifests itself differently: from very slight to very strong. 49 of them are ethnic Azerbaijani and native speakers of the Azerbaijani language; one speaker is an Azerbaijani-born native speaker of Russian. 38 speakers are male, and 12 are female. Their age ranges from 21 to 60, their education, from secondary to higher, and they are people of different professions: some of them are engaged in technical sciences, some in humanities, and some in natural sciences, some of them are sportsmen, and some have no special education. All participants speak Russian, almost all of them are native speakers of Azerbaijani. They were born and used to live and/or live in the territory of Azerbaijan, in different cities and regions of the country: Baku, Sumgayit, Gyandzha, Agdam, Nakhchivan, Lankaran, Salyan, Tovuz, Khachmaz, Fizuli, Astara, Ordubad, Yardymli, Shemakha. Some of them moved around within the territory

of Azerbaijan, lived in Russia or Ukraine for a short time due to different circumstances, or moved to Russia. Total duration of the used spoken material from the fifty speakers was more than fifteen hours. Besides, we used control records of native speakers of the Turkish and Russian standard language reading a text and talking. For experiments with expert listeners, we also used records with Germans, Georgians, and Armenians speaking Russian, and authentic expert records with imitation of the so-called accent of people from the republics of the Caucasus.

Results

As a result of multistage research and pilot running of the program, we compiled a list of forty phonetical features, both segmental and suprasegmental, characteristic for the Azerbaijani accent in the Russian speech. In the course of the research we have studied the regularity, frequency, and strength of their occurrence in the speakers' speech, and determined their distinguishing ability. We have also developed a system of relative appraisal of accent manifestation by points.

    Table 1 shows the experimentally established list of interfered characteristics of the Russian speech with the Azerbaijani accent.

Group	Characteristic
Consonantism	**Pronunciation of a deeper or more uvular consonant [χ] instead of [x]**
	Pronunciation of a laryngeal [h] instead of the Russian velar [x]
	Pronunciation of palatalized or not velarized enough sibilants instead of [ş] and [z̧]

Table 1. Interfered Characteristics of Russian Speech with Azerbaijani Accent.

The system of characteristics inherent for the interaction of the Russian and Azerbaijani languages is shown in bold; separate features of interaction between Russian and dialect varieties of the Azerbaijani language are shown in italics; and features of the accent inherent to speakers of typologically different languages are shown in plain text.

    We have conducted a mathematical research of the described characteristics of the Azerbaijani accent in the interfered Russian speech (all mathematical calculations and software development described have been performed under the control and with personal participation of mathematician Yu.V. Zaytseva, leading expert of Autonomous Non-Commercial Organisation Laboratory for Applied Linguistics).

    The selected characteristics were evaluated by their 'weight' for the formation of a linguistic profile of the Azerbaijani accent and classified by their strength and frequency, as well as by their belonging to the Azerbaijani accent, or its regional variant, or to the characteristics of the accent which manifests itself in the Russian speech spoken by native speakers of different languages. With due consideration of these parameters, we have derived a function of the set characteristics. The value of the function informs the researcher of the degree of certainty with which he can consider the given interfered speech to be containing the characteristics of the Azerbaijani accent. The function and list of characteristics have been included in a program enabling to register the selected characteristics in the speakers' speech and obtain their sum in order to diagnose the Azerbaijani accent. Apart from the indicated list of characteristics inherent to the Azerbaijani interfered speech, the program contains a player and a sample base. By pressing

a key near the field containing the list of characteristics, you can listen to sound samples of the interfered characteristics of the accent. A field for entering a commentary to a characteristic has also been added in order to specify its content in more detail where needed. For the view of the program's working window, see Figure 1:
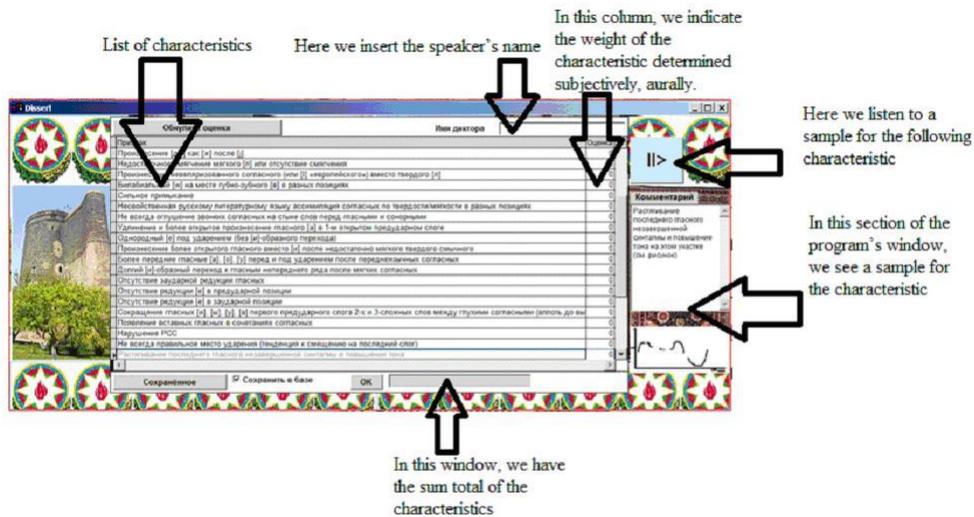
Figure 1. Interface of the Program for Research of the Russian Interfered Speech with an Unknown Accent (for Establishing the Presence or Absence of Characteristics of the Azerbaijani Accent).

Upon listening to a record, the expert can compare them to the samples from the base and evaluate them using a 5-point rating scale (the function shall be calculated using only the selected/non-selected characteristics). The values obtained as a result of calculating the function from the characteristics using the derived algorithm inform the expert of the degree of certainty with which he can consider the accent in question to be Azerbaijani (for this purpose, threshold values have been set – see Table 2).

0-5	More likely not Azerbaijani Accent
5-25	Accent Blending
25-30	May be Azerbaijani Accent
30-61,0702	Most Probably Azerbaijani Accent
61,0702 <Q≤ 3481	Almost Confidently Azerbaijani Accent

Table 2. Values of sum and their interpretation.

Approbation of the program showed that in most cases the selected complex of characteristics allows to correctly identify the characteristics of the Azerbaijani accent in the Russian interfered speech. But the key factor during a real research is the expert's readiness for diagnostic tasks of this kind. Field testing showed that experts' error in identification of these characteristics is sometimes dramatically big. For example, some listeners who had a great teaching experience in teaching Russian for non-Russian speakers could correctly define the whole set of marked characteristics of Azerbaijani accent in Russian

interfered speech and distinguish this accent from German, Georgian, Turkish and imitated accents (more than 6 approbation tasks of 8 were solved right). Other listeners, who did not have enough experience of listening to interfered speech could mix up marked accent features, incorrectly treat their significance and confuse foreign accents (less than 4 tasks of 8 were right solved).

We tried to minimize the error artificially by employing different mathematical algorithms. Firstly, the 'reliability' of the characteristics depended on their frequency in the appropriate positions per a unit of time (in fact, an allowance for the duration of the record has been introduced for sake of 'reliability' of some characteristics). In actual practice, the record can be just too short for the probability of appearance of some characteristics to reach one. Secondly, we introduced duplicated or strongly correlated characteristics, and, if the expert's evaluations of these characteristics were significantly different, the 'level of credibility' for the answers as a whole was lowered. Thirdly, we lessened the weight of the characteristics the experts made most errors with.

Nonetheless, because of the GIGO principle, it is mathematically impossible to fully exclude the error caused by the experts' errors.

Conclusion

The given list of interfered characteristics shall be useful for identification researches aimed at facilitating singling out and formulating the accent peculiarities the expert heard.

Thus, we may conclude that the characteristics identified in the interfered Russian speech with a given accent can only be considered as part of a complex. Individual features can be similar in different accents or types of accents (dialects). Obtaining reliable results of identification of a complex of national diagnostic characteristics requires a comparison with the characteristics of an accent of contrasting languages, both with a similar and a dissimilar phonetical layout. The scheme for studying interfered speech developed in this research enables us to study the peculiarities of interaction of Russian and any other language (with due consideration of its regional diversification). Apart from its scientific and theoretical significance, this information has great practical value, because this is connected with the main task of forensic expertise of an audio record: identification of a person on the basis of their speech.

References

Barkhudarova, E. L. (2008). K Probleme Sozdaniya Nacionalno Orientirovannyx Kursov Russkoj Zvuchashhej Rechi (On the Problem of Creation of Language Depended Courses of Russian Speech). In A. Arkhipov, Ed., *Phonetics and Non-Phonetics: On the Anniversary of Sandro V. Kodzasov*. Moscow: Yazyky slavyanskikh kultur.

L. V. Bondarko and L. A. Verbitskaya, Eds. (1987). *Interferenciya Zvukovyx Sistem (Interference of Sound Systems)*. Leningrad: Leningrad State University.

Erofeyeva, E. V. (1997). *Eksperimentalnoe Issledovanie Fonetiki Regionalnogo Varianta Literaturnogo Yazyka (An Experimental Study of the Phonetics of a Regional Variety of the Standard Language)*. Perm: Perm University.

Galyashina, E. I. and Goloshchapova, T. V. (2004). *Interferenciya Nacionalnogo Yazyka pri Identifikacii Russkogovoryashhego Diktora (na Materiale Yazykov Narodov Rossii): Metodicheskie Rekomendacii (Interference of the National Language in Identification of*

a Russian-speaking Speaker (on the Material of the Languages of Peoples of Russia): Methodical Recommendations). Moscow: Forensic Science Centre of the Ministry of the Interior of Russia.

Galyashina, E. I. and Khurtilov, V. O. (1991). *Diagnostika Nacionalnogo Akcenta po Fonogrammam Russkoj Rechi (po Yazykam Narodov SSSR): Metodicheskie Rekomendacii (Diagnostics of a National Accent Using Phonograms of the Russian Speech on the Materials of the Languages of the Peoples of the USSR): Methodical Recommendations).* Moscow: Forensic Science Centre of the Ministry of the Interior of Russia.

Galyashina, E. I. and Zlatoustova, L. V. (1999). Raspoznavanie Individualnyx i Gruppovyx Akustiko-Perceptivnyx Xarakteristik Govoryashhego po Zvuchashhej Rechi (Recognition of Individual and Collective Acoustic and Perceptive Features of a Speaking Person on His Speech). In *Theory and Practice of Speech Studies (ARSO)*, 60–80. Moscow: Lomonosov Moscow State University.

Rogoznaya, N. N. (2001). *Lingvisticheskij Atlas Narushenij v Russkoj Rechi Inostrancev (Linguistic Atlas of Mistakes in Russian Speech of Foreigners).* Irkutsk: Irkutsk Press.

Vinarskaya, E. N. and Zlatoustova, L. V. (1977). K voprosu o foneme (On the Phoneme). *Scientific Report of the Higher School, Philological Sciences*, 1, 103–108.

Zlatoustova, L. V. (1975). Rhytmic Structure Types in Russian Speech. In G. Fant and M. A. A. Tatham, Eds., *Auditory Analysis and Perception of Speech*, 477–485. New York: Academic Press.

Zlatoustova, L. V. (1994). Typological and Stylistic Characteristics of the Phonetic Word with Examples from some Indo-European Languages. In A. A. Polikarpov, Ed., *2ⁿᵈ International Conference on Quantitative Linguistics*, 158–160. Moscow: Lomonosov Moscow State University.

Forensic Linguistics accredited:
Four years of experiences with ISO 17020 in authorship analysis

Sabine Ehrhardt [*]

Abstract. *In 2009, the report of the US National Academy of Sciences was published by which the unsatisfactory situation in forensic sciences was impressively revealed. One major issue was how the work of forensic experts can be improved, standardised and kept monitored in order to prevent incorrect results and conduct. In this context, the Forensic Science Institute of the German Bundeskriminalamt decided in favour of an accreditation. A comprehensive quality management system has been implemented and the linguistics department works according to the norm ISO 17020 since more than four years now. This had certainly many positive effects for standardisation, transparency and credibility. But of course an accreditation is not the panacea for all the inadequacies that may prevail in forensic sciences.*

This paper focuses on how quality management was implemented in forensic linguistics, which ethical issues it addresses, how it affects scientific work routines and which improvements for case work can realistically be expected of an accreditation. A second focus will be on the problems that disciplines like forensic linguistics face when they strive for accreditation or other forms of quality assurance and professionalising practice. The results of a survey among forensic linguistics experts with a business or university background will be used to complement the experiences made with a quality management system in law enforcement.

Keywords: Forensic linguistics, authorship analysis, quality assurance, accreditation, ISO 17020.

Introduction

The report of the National Academy of Sciences (NAS report) has shown that the US forensic sciences community is fragmented and reflects unfavourable heterogeneity. As

---

[*]Bundeskriminalamt, Germany

well, it is characterised by a lack of standards and resources (NAS report: Committee on Identifying the Needs of the Forensic Sciences Community / Committee on Applied and Theoretical Statistics / National Research Council, 2009: 2-17). However the state of forensic science in other parts of the world may be, quality management is one of the major issues in this context, especially for a discipline like forensic linguistics, which is only partly established in the sense that it is not an inherent part of laboratories or forensic science institutes, and that it lacks best practices as well as standards for education and training.

The situation of forensic linguistics differs internationally, largely due to differences in judicial systems and judicial decisions concerning the admittance of linguistic evidence in court. Accordingly, the working conditions for forensic linguists may differ substantially, although similarities are most likely to be found with respect to linguists working in the academia. In Germany, the majority of forensic linguistics experts presumably belong to this group. To the author's knowledge, experts with a university background apply their science to case work just as a sideline job. In contrast, full-time practitioners are small in number. Some of them work in law enforcement since Germany is one of the few countries to have forensic linguistics expertise included into the scope of services offered by the Forensic Science Institute of the Bundeskriminalamt (BKA)[1]. Very few German experts of forensic linguistics are employed in consulting firms. As far as is known, there are just a handful of private enterprises to be in this line of business and, for at least one of these, the focus rather seems to be on handwriting analysis with forensic linguistics only to complement it.

From the perspective of clients in need of forensic linguistics expertise, it might be hard to find experts whose credentials can be easily evaluated. For one reason, Germany lacks a national board for representation and developing methods and standards in forensic linguistics. For another reason, linguistic experts are only seldom listed in registers, e.g. professional registers that a national board or association of forensic linguistics may be in charge of or official registers like local commercial directories or the Main Chamber of Commerce and Industry.

In general, the situation of fragmentation and heterogeneity as the NAS report described it can be observed in German forensic linguistics as well. It calls for different measures to overcome the current situation, particularly addressing issues like consistency, reliability of results, comparability as well as integrity of conduct. Terms that are frequently and sometimes inaccurately used in this context are *quality management*, *quality assurance*, *standardisation*, *best practice*, *code of ethics*, *accreditation* and *certification*. The expression 'quality management' was usually employed to describe processes in the business sector but now it is generically used in all kinds of sectors offering products and services to clients. It incorporates different principles of management like the customer focus and the continual improvement process. The term 'quality management' does not necessarily refer to efforts oriented towards good quality. Instead the focus is put on consistent quality. There are several components making up a quality management system with quality assurance being one of them. (The others are quality policy, objectives, planning, control and improvement according to DIN EN ISO 9000:2005 (2005: 21).) Quality assurance comprises the systematic activities to fulfil requirements concerning a predefined quality of a product or service.

Consistency as the main objective of quality management is tackled by standardisation, which refers to processes of developing and implementing standards, i.e. finding "best" solutions to specific problems and applying those methods consistently. Standards can be differentiated according to the degree of obligation ranging from voluntary standards up to legally binding (de jure) standards.

Besides ensuring consistency by standardisation, a further component of quality management is the improvement of the overall performance. Central to improvement processes is the concept of 'best practice', which refers to methods that have proved to produce superior results and that are used as benchmarks compared to other methods and procedures.

In order to confirm specific characteristics of their products and services, an organisation may choose a process of certification. This confirmation is organised to fit predefined standards like those of the International Organisation of Standardisation (i.e. certification according to ISO 9000) and it is provided by an external and official accreditation body which uses monitoring instruments such as audits, surveillances and re-accreditations to regularly inspect the adherence to the chosen standard. Thus, certification can be defined as a formal, external confirmation about the existence of an operating quality management system. An accreditation goes beyond certification in so far as it also comprises the confirmation that the organisation in question has the competence to provide the offered services or products.

In contrast to accreditation as an official form of managing quality, a code of ethics has a different status. Codes of ethics comprise a set of values and recommendations for members of an organisation or a profession. These values are the result of internal quality-managing efforts and can be considered as voluntary self-restrain rather than external regulations, although both forms may not differ much with respect to the degree of obligation that is put on the members of the organisation or profession in question[2].

For this paper, the expressions 'quality management' and 'standardisation' will be particularly relevant as these concepts play a central role in the first part, i.e. the description of the implementation of a quality management system covering the work of forensic linguists. The expressions 'best practice' and 'code of ethics' will be more important in the second and particularly in the third part, i.e. the discussion about the effects of a quality management system in forensic linguistics as part of a law enforcement authority and in the following evaluation of a survey among linguists who are working as experts under different conditions.

Accreditation and the ethical issues that are addressed

The BKA's decision to strive for an accreditation was made in 2003 on the basis of many and diverse causes[3]. The need to assure quality of results in sensitive areas like forensic sciences was generally recognised many years ago, and in accordance with this attitude there are recommendations by international organisations like the European Network of Forensic Science Institutes (ENFSI) and mandatory requirements by political institutions like the EU council. Within the field of forensic science, cases have become known in which the contribution of forensic science institutes was declined in matters of letters rogatory in international judicial proceedings[4] as well as in matters of EU programmes[5]. In both instances, the reason for declining cooperation was the participating institutes'

lack of an accredited quality management system and consequently their failure to prove that their work adheres to international standards.

In recent years, the need for forensic institutes to get accredited has become even stronger. The EU council reached a decision which states "the need to establish common standards for forensic service providers" and mandatorily requires international standards such as ISO 17025 (2005), at least for laboratories making DNA profiles and dactyloscopic data available (Council of the European Union, 2009). The EU council also published their vision of European Forensic Science 2020 (Council of the European Union, 2011). It involves the further development of forensic science infrastructure in Europe and the ensuring of a consistent administration of justice, e.g. by accrediting forensic science institutes and laboratories, establishing common best practice manuals, and conducting proficiency testing as well as collaborative exercises (Council vision for European Forensic Sciences 2020, 13/14 December 2011). ENFSI has set the standard that its member laboratories shall have accredited at least 50% of their fields of expertise where on average twelve examinations per year are performed (European Network of Forensic Sciences Institutes, 2011).

The decision to strive for an accreditation naturally implies the decision according to which norm a forensic discipline is accredited. A frequent choice for forensic laboratories is the norm ISO 17025. However, forensic linguistics as a strongly experience-based science requires additional considerations because any form of quality assurance has to comprise not just the process of producing scientific results but also the appropriate interpretation of those results. A work process that consists of both the examination of materials and the professional judgement of this examination's results is called *inspection* in the terminology of international standards. In contrast to the norm ISO 17025 which refers only to the examination process, the norm ISO 17020 also covers the "exercise of professional judgement" (2012: 7), i.e. the actual work of experts to assess, judge and interpret facts:

> This international standard covers the activities of inspection bodies whose work can include the examination of materials [...], and the determination of their conformity with requirements and the subsequent reporting of these activities to clients and, when required, to authorities. [...] Such work normally requires the exercise of professional judgement in performing inspection. (DIN EN ISO/IEC 17020:2012, 2012: 7)

In this context, the main aim of the international standard ISO 17020 is stated as "promoting confidence in bodies performing inspections" (2012: 7). In order to reach this aim diverse criteria are set up concerning the competence of an institute, its impartiality and the consistency of its work (2012: 9). By this, major ethical issues in the work of a forensic expert seem to be addressed. The issue of *competence* is approached by both resource and structural requirements. Most importantly, resource requirements include the education and training of those who perform the inspections. Forensic experts and their co-workers are supposed to be employed according to their abilities (i.e. formal proof of qualifications and degrees) and they are supposed to be regularly trained in order to preserve and further their knowledge. Structural requirements cover aspects like management, workflow, equipment, and facilities.

Impartiality refers to the presence of objectivity and consequently the avoidance of conflicts of interest and (financial) pressures, e.g. by regulating the expert's remunera-

tion accordingly or setting transparent guidelines for sponsoring of research, equipment etc. Of course, the remuneration of an expert in dependence of the outcome of the case as it might occur especially in adversarial judicial systems presents a fundamental risk to impartiality and should thus be addressed by any form of quality assurance or a code of ethics as Stygall (2009: 264) and Ainsworth (2009: 281) point out. For linguistic experts working in German law enforcement, remuneration is regulated in accordance with public sector pay and, thus, independently from the outcomes of cases. But financial pressures or conflicts of interest may nevertheless arise, e.g. by invitations and free event tickets as forms of concealed sponsoring. These risks have to be made obvious and subsequently prevented by corresponding regulations.

In order to ensure *consistency*, the inspection methods are standardised by the so called standard operation procedures (SOP). Similarly, the handling of all items that are related to the tasks at hand is regulated by a retraceable chain of custody. Consistency is also concerned when it comes to faults and mistakes. An extensive complaints and appeals process is supposed to address these aspects. Of course, consistency is not meant to be an end in itself but rather intended to increase comparability of results and transparency of procedures in the interest of clients.

As the linguistics department at the BKA is embedded in the Forensic Science Institute it benefits from a standardised infrastructure covering the entire institute and referring to generalised aspects like the training of experts in legal and law enforcement matters, the financing of equipment or research projects and the extent of documentation. The actual linguistic work is standardised by standard operation procedures that were already mentioned above. Each kind of inspection is described separately with respect to its methodological basis and its purpose. For example, there are SOPs about the inspection forms Text Comparison, Text Analysis (categorisation of an author on the basis of his/her anonymous writing), Administration of the Text Corpus and Conducting Corpus Searches. SOPs may contain statements about the scope of application, responsibilities, information about the method, principle and actual realisation, about tools and aids, documentation, quality-assuring measures, references as well as the validation of the method in question. The setting up of such a standard operation procedure obliges its users to adhere to what was fixed as well as to regularly prove that each inspection produces the expected results. This is usually done by proficiency tests or (if practicable at all) by collaborative exercises which are just one kind of control mechanisms that a quality management system has to include[6].

Concerning consistency of work, the workflow of an inspection and the subsequent report about these activities are standardised. Thus, it is clearly regulated what happens when requests of clients (i.e. police, prosecution and the courts) are received and how these requests are to be handled. Most of all, records have to be kept to such an extent that all processes and procedures can be retraced and understood. Similarly the expert's report about the conducted analysis is meant to cover all aspects involved to understand the whole process and its results. Corresponding to this aim, the structure of reports was standardised as presented in Table 1.

Before reports are sent to clients, they have to pass several checks, e.g. a formal check by co-workers, an in-content check by a second expert (peer review) and a plausibility check by persons higher in rank than the expert him-/herself.

Table 1. Standardised structure of reports.

1. Request (quotation of the original request + rephrasing by the expert)
2. Unambiguous description of the items of inspection, i.e. the texts that are to be examined
3. Methodological outline as in the SOP (terminologically adapted for non-linguists if necessary)
4. Results of the examination
   (a) Critical inspection of the items for analysis: text quality + quantity + time of origin (if necessary) → assessment of suitability for the inspection at hand
   (b) Results for each text separately (as neutral representation of facts)
5. Discussion and interpretation of the examination results
6. Conclusion, i.e. essence of *5. Discussion*

Although this structure of expert reports appears quite rigid the overall objective of these guidelines is to ensure the comprehensibility of the analysis at hand. A report is written with the aim to help the court or the police, thus, it must be understandable both in its scientific content – particularly to those without linguistic training — and its line of argumentation, which means it has to reveal which linguistic findings lead the expert to draw hisher conclusions. Furthermore, the consistent structure of reports is supposed to support comparability between reports, especially with respect to the interpretation of results. Exaggerating interpretations of findings as well as inappropriate choice of methods and incomplete representations of examination results are meant to be prevented by these structural guidelines.

Accreditation and its effects

The standardisation concerning the general conditions of case work in forensic linguistics as described above has undoubtedly positive effects. The most important of these is the influence on consistency in the experts' work which is reflected in the comparability of results and the reports about the conducted analyses whose importance Ainsworth (2009: 287) already emphasised. Likewise, the issue of adequate documentation is dealt with because a quality management system according to ISO 17020 obliges to a clearly defined procedure of keeping exact records with the effects of ensuring a traceable chain of custody for all items involved in an inspection and making all aspects of the process comprehensible from beginning to end (also cf. Ainsworth (2009: 286)). Additionally, there are positive side effects like an improved customer service due to the speeding up of routine processes and enhanced possibilities to schedule requests for example.

Nevertheless, the scope of these positive effects is limited and there are ethical issues that cannot be directly addressed by an accreditation. Ethical issues in forensic sciences arise from diverse causes. Stygall names three possible sources: the nature of the legal system, the nature of the discipline of linguistics, and the nature of scientific research in any field (Stygall, 2009: 254). The German legal system is not an adversarial system and an expert might be less liable to become biased as a consequence of working

for one party or becoming member of the attorney's team in a fashion Finegan (2009: 274) describes. But of course, a system with an impartial and independent judge who is in complete charge of jurisdiction has its own characteristics that might cause ethical issues. Butters understandably puts up the questions "[H]ow do we assure that the expert is really neutral? Who is there to point out when the neutral expert is simply wrong?" (Butters, 2009: 241). Concerning the first question, an expert's opinion might be influenced by the connecting facts of the case. Then, it is up to the defence lawyer to 1) recognise that an analysis was not impartially conducted and 2) convince the judge of the fact that a second expert needs to be consulted — with both steps possibly presenting further issues. A quality management system at least provides the framework to operate under minimized risks to impartiality; although neutrality cannot absolutely be assured since it also depends on the attitude of a forensic expert. In addition to general regulations, supplements in form of a professional code of ethics are helpful here because they also address different aspects like the image that professional linguists feel necessary to transport of their science.

The second question Butters puts up is highly relevant because it refers to the ability of judges to assess the credentials of the experts they consult. Logically, it is problematic for someone who considers it necessary to request support of a specific forensic discipline to assess this support with respect to its realisation. But this is the task that a judge has to perform – on the one hand, realising the necessity of consulting a linguist to find answers to a linguistic problem and obliging a linguist to do just that, and on the other hand, evaluating linguistic matters as to whether or not this expert is applying the science correctly, with sufficiently great knowledge and in an up-to-date fashion. When in doubt, a German judge in criminal trials has the possibility to commission a decisive expert opinion.

Ethical issues concerning the nature of linguistics can be similarly hard to address directly by an accreditation and a quality management system. Most of all, the question about what quality actually is and how this relates to the problem of the "scientificnes" of linguistics is to be answered. In order to prove the reliability and validity of methods, proficiency testing and collaborative exercises are surely steps in the right direction. However, there are approaches in forensic linguistics, e.g. purely qualitative analyses, which can hardly be captured in terms of error rates. Furthermore, proper proficiency tests and collaborative exercises require a certain quantity of forensic linguists to take part. Considering the heterogeneity of tasks, subjects and working conditions in forensic linguistics, this quantity might not be reached in sufficient rates to have all methods regularly tested. In general, this fragmentation of linguistics as a forensic discipline and, consequently, an only limited number of colleagues (or otherwise cooperating practitioners) also implies the risk of self-referentiality. If only a handful of experts set standards, these standards might just be descriptions of what is already done anyway instead of developing standards to the benefit of the forensic discipline in general.

Ethical issues also arise from the nature of science. Again, there are issues that cannot be addressed directly by an accreditation according to an international standard like ISO 17020. The norm obliges to set up general conditions that ensure an expert's competence. These general conditions comprise regulations concerning additional training as well as formal proofs of competence in form of qualifications and academic degrees. However, these general conditions are only partly suitable to influence attitudes towards

aspects of competence. It still lies in the individual responsibility of an expert to assess the own limits of knowledge and experience with the consequence to decline requests when they require a competence beyond the own realms. Furthermore, there are issues of scientific standards to be considered which might run counter to what is most helpful for the participants of an investigation or trial. Surely, experts must not refrain from appropriately giving to understand the complexity of scientific analyses with their ambiguities and limits of interpretation of linguistic findings. Similarly, the language of reports has to be adequate and in adherence with scientific practice. Technical jargon is not to be avoided or even replaced by colloquial language. Instead, a combination of both linguistic terminology and "translations" for non-linguists should be sought. All in all, this refers to what Ainsworth clearly stated: "experts owe their professional allegiance to science, not to the lawyer and client in any particular case" (Ainsworth, 2009: 284).

To sum up these last paragraphs, the accreditation according to ISO 17020 can be discussed as to whether or not it implies the risk to focus solely on workflow and organisational aspects. Issues like competence, integrity and compliance with scientific practice are not easily to be covered. It appears that an accreditation should be complemented by a code of ethics set up by forensic linguistics practitioners themselves in order to prevent that quality assurance only provides a superficial frame for procedures that in fact do not adhere to either forensic or scientific standards.

Alternative quality-assuring measures in forensic linguistics

Naturally, an institutional quality management system differs from the quality-assuring measures taken by forensic linguistics experts as privately practising professionals. In order to complement the perspective of an accreditation in a forensic science institute a small survey was carried out. Its main focus is on alternative quality-assuring measures in forensic linguistics. The questionnaire was developed with the aim to supplement the description of experiences made during the accreditation process and is restricted to these aspects. The fragmented nature of the German forensic linguistics community resulted in a limited number of participants, nevertheless their background and their working conditions reflect the heterogeneity of the work of a forensic linguistics expert, as outlined in the introduction to this paper. In total, twelve experts have been asked to take part in the questionnaire. Six of them work in the academia as their main job, four of them work in diverse branches offering forensic linguistics expertise as a sideline job, and two of them have been chosen as representatives of consulting firms. Answers of eight experts representing each of the different professional backgrounds have been received[7].

The questionnaire consisted of nine questions that can be condensed to the following four quality assurance issues:

- *Attitude of experts towards quality assurance in forensic linguistics* (Do you see a need for action concerning quality assurance in forensic linguistics? Do you think that quality assurance can be somehow problematic in less well-established forensic disciplines like forensic linguistics? Is there anything you want to comment on, encourage or criticise with respect to the efforts of standardisation and quality assurance in forensic sciences in general?)
- *Attitude of police and courts towards quality assurance* (Have you ever been asked for quality-assuring measures by those who requested linguistic expertise from

you? Have you ever been asked in court to prove that your work adheres to quality standards?)

- *Approaches to quality assurance by the experts* (How do you make sure your work is of consistent quality? Are you accredited and why [not]? Are you linking up with other forensic linguistics experts to develop methods and measures of quality assurance respectively?)
- *Standardisation of methods and its reflection in expert reports* (Concerning the linguistic methods that you use, do you adjust them according to the request you received? Do you integrate a methodological outline in your reports? Is this outline standardised or adjusted to the analyses at hand?)

The result of the survey is quite clear. Quality assurance is commonly considered to be of highest relevance. What is more, many participants of the survey urgently called for actions on both a national and an international level. They declared a definite need of at least minimum standards for methods, the appropriate evaluation of results as well as the writing of reports. Especially, the issue of inadequately compiled reports was often mentioned as the majority of experts seem to have come across reports of intolerable quality.

But the participants of the questionnaire also expressed different severe concerns. First of all, there is the concern that the possibilities for standardisation and quality assurance in forensic linguistics are limited due to the diversity of linguistic tasks that are presented to the experts by the courts and police. Furthermore, current methods of standardisation and quality management (according to international standards) are suspected to miss the point. Instead of addressing professional competence and an ethical attitude towards the work as a forensic expert, they are perceived as referring to superficial aspects of workflow and the organisation of general conditions and overemphasising consistent work routines. A further concern in connection with an accreditation is that quality management may be confined to those who can afford it as the implementing, running and sustaining of a quality management system ties up resources that individual experts simply do not command. In general, the benefits of an accreditation are recognised, but an accreditation is not seriously considered to be feasible for individually working experts.

Surprisingly, the attitude of the courts and the police towards quality assurance seems to be bordering on indifference. Only few experts declared that they are regularly asked to disclose their academic degrees. In these cases, an answer that mentions the M.A. or PhD in linguistics suffices (cf. Butters (2009: 239) who seems to have observed similar attitudes). It does not appear to be relevant either which subfield of linguistics an expert has specialised in or if this suits the analysis that has been conducted. Neither do judges try to determine whether the expert has finished additional training with respect to working in a forensic science context. Two experts said that they themselves usually point out their competence in the matter at hand.

In the questionnaire, the third complex of questions refers to what forensic linguistics experts actually do to tackle the problem of quality assurance. That this issue is in each expert's own responsibility is reflected by the heterogeneity of approaches. The quality-assuring measures listed by the participants of the survey include standardisation of workflow, methodological consistency, using of different methods to back up results, peer review, keeping oneself scientifically up-to-date, using linguistic theory as

basis of expert work, feedback of clients, appropriate time management, quality awareness, profundity, carefulness as well as relying on own publications and experiences. Most of these aspects were mentioned only once in the participating group of experts. Their variety suggests that quality assurance is not systematically approached and indeed regarded as a rather private matter. In this survey, exceptions were rare. For example, just one expert mentioned that he/she considers certification according to ISO 9000. Another expert pointed out that institutions like the Main Chamber of Commerce and Industry set up standards for regular training and insist on a proof of competence.

When it comes to standardisation of methods, the experts overall agree on its necessity. They usually have their own methodological outlines, which are supplemented by further explanations to suit the analysis at hand. However, in expert reports this issue is handled differently. Despite using a consistent methodological approach, some experts do not consider it necessary to convey it to their clients. In contrast, other experts strongly argue for an outline of methodology and the scientific grounding of the analysis that is conducted and documented in the report. Anyway, expert reports seem to be a disputed issue for the participants of the survey as many of them mentioned reports from colleagues that clearly did not adhere to what they themselves perceived as appropriate. The examples included inadequate presentation of results as well as inappropriate language marked by exaggerations and a frequent use of adverbs like "always" and "never" in the interpretation of linguistic findings.

Conclusion

The accreditation according to ISO 17020 has certainly many positive effects for the application of forensic linguistics to case work. The working conditions in a forensic science institute are suitable to implement, run and sustain a quality management system. Thus, the accreditation and its consequences on standardisation can surely be considered to be an appropriate measure to handle ethical issues of forensic sciences like those mentioned in the NAS report. However, the working conditions for the majority of forensic linguists are different from the working conditions in law enforcement. Independently and privately practising linguists cannot benefit from an already existing infrastructure; instead they have to tackle each aspect concerning their work themselves. This fact is reinforced by the fragmentation of the forensic linguistics community, which leads to fewer joint efforts and synergic effects than it could certainly be possible under circumstances of increased linking-up.

The contributions of the forensic linguistics experts who participated in the survey were very clear on the point that the development of standards for forensic linguistics is welcomed. Standards should address linguistic methodology as well as the attitude (including conduct) of forensic linguistics experts. The latter might be best approached by setting up a professional code of ethics, the former by developing best practice manuals of specific tasks that experts are frequently confronted with. It goes without saying that differences in working conditions of practising experts are to be taken into account. A general aim of all of these efforts could possibly be that forensic linguistics experts have at their disposal a range of recommendations and quality-assuring measures which suit the diversity of tasks and conditions of the application of forensic linguistics to case work.

Notes

[1]German forensic sciences in general are incorporated in law enforcement, although there is a strict division between investigation and the (scientific) analysis of traces.

[2]For example, the IAFL posted a draft of a code of ethics on its website in May 2013. The document is called "Code of Practice" and is meant to provide "principles of ethical conduct [...] intended to guide those members of the International Association of Forensic Linguistics who engage in forensic linguistic research and legal consulting and testimony" (International Association of Forensic Linguists, 2013: 1).

[3]The following description of reasons for the accreditation is based on presentations and training material of the quality-management team in the Forensic Science Institute (Bundeskriminalamt Germany). The author is especially indebted to Maria Kambosos for her valuable suggestions.

[4]This refers to a Swedish homicide case from 2003 ("Lindh") in which the letter rogatory from Sweden could not be complied. According to the Swedish institutes' norms, a cooperation with other forensic science institutes is only advisable when these laboratories can prove that they work with comparable standards. At that time, the appropriate forensic discipline was not accredited and, thus, could not prove its standards.

[5]The application of the BKA for taking part in the EU Phare Twinning Programme for Turkey has only been successful because of the support of a British forensic science institute that was already accredited.

[6]Other kinds of control mechanisms are internal audits (every 12th month), surveillances by a member of the national accreditation body together with an expert of the forensic discipline concerned (every 16th month) and the process of re-accreditation (every 5th year).

[7]The author is greatly indebted to the participating forensic linguistics experts. Their contribution to the questionnaire comprised many and diverse aspects of an expert's work. Unfortunately, not every aspect mentioned could be related here. This is not meant to imply an evaluation of the contribution in any way.

References

Ainsworth, J. (2009). A lawyer's perspective: ethical, technical and practical considerations in the use of linguistic expert witnesses. *International Journal of Speech, Language and the Law*, 16(2), 279–291.

Butters, R. R. (2009). The forensic linguist's professional credentials. *International Journal of Speech, Language and the Law*, 16(2), 237–252.

Council of the European Union, (2009). Council framework Decision 2009/905/JHA of 30 November 2009 on Accreditation of forensic service providers carrying out laboratory activities. *Official Journal L 322, 09/12/2009*, 0014–0016.

Council of the European Union, (2011). *Council conclusions on the vision for European Forensic Science 2020 including the creation of a European Forensic Science Area and the development of forensic science infrastructure in Europe. 3135th JUSTICE and HOME AFFAIRS Council meeting.* Brussels.

DIN EN ISO 9000:2005, (2005). *Quality management systems – Fundamentals and vocabulary (ISO 9000:2005). Trilingual version EN ISO 9000:2005.*

DIN EN ISO/IEC 17020:2012, (2012). *Conformity Assessment – Requirements for the operation of various types of bodies performing inspection (ISO/IEC 17020:2012); German and English version EN ISO/IEC 17020:2012.*

DIN EN ISO/IEC 17025:2005, (2005). *General Requirements for the competence of testing and calibration laboratories (ISO/IEC 17025:2005); German and English version EN ISO/IEC 17025:2005.*

European Network of Forensic Sciences Institutes, (2011). *ENFSI Standing Committee for Quality and Competence (QCC): Scope of Accreditation. Policy, Ref.Code: QCC-ACR-002.* Number 001.

Finegan, E. (2009). Expert linguists and the whole truth. *International Journal of Speech, Language and the Law*, 16(2), 266–277.

International Association of Forensic Linguists, (2013). Code of practice draft.

NAS report: Committee on Identifying the Needs of the Forensic Sciences Community / Committee on Applied and Theoretical Statistics / National Research Council, (2009). *Strengthening the Forensic Science in the United States: A Path Forward.* Washington.

Stygall, G. (2009). Guiding principles: forensic linguistics and codes of ethics in other field and professions. *International Journal of Speech, Language and the Law*, 16(2), 253–266.

The problems and practicalities of training translators and interpreters for the future envisaged by the EU Directive 2010/64/EU

Belinda Maia *

Abstract. *The theme of the IAFL conference, 'Bridging the Gap between Language and the Law', in Porto in 2012 encouraged the creation of a parallel session on aspects of multilingualism and the law, and the inclusion of a Round Table on the problems raised by the EU Directive 2010/64/EU on the right to interpretation and translation in criminal proceedings. This article describes the main points of addressed during the Round Table, which brought together representatives of the Directorate-General for Translation of the European Commission, the professional organizations EULITA and AIIC, and the TRAFUT and IMPLI projects.*
The article begins by reflecting on the historical, cultural, social and educational misunderstandings that underlie so much of what is at stake, and how all these organizations and projects are responding to the challenges ahead. It will then consider how a country like Portugal, with a language that is minor in European terms but major globally, could prepare for the future envisaged by the EU Directive, and make suggestions of ways in which the educational establishment can contribute. Much of what is suggested for Portugal is applicable in different degrees to other European countries and situations.

Keywords: EU Directive 2010/64/EU, multilingualism and the law, training interpreters and translators.

Introduction

The theme of the IAFL conference, 'Bridging the Gap(s) between Language and the Law', in Porto in 2012 encouraged the creation of a parallel session on aspects of multilingualism and the law, and the inclusion of a Round Table on the problems raised by the EU Directive 2010/64/EU on the right to interpretation and translation in criminal proceedings. The discussion brought together representatives of the Directorate-General for

---

*Universidade do Porto

Translation (DGT) of the European Commission (EC)[1], the professional organizations EULITA (European Legal Interpreters and Translators Association)[2] and AIIC (International Association of Conference Interpreters)[3], and the TRAFUT (Training for the Future)[4] and IMPLI (Improving Police and Legal Interpreting)[5] projects, as well as various people from both the professional and the academic sides of the Law and Languages.

This paper will begin by reflecting on the historical, cultural, social and educational misunderstandings that underlie so much of what is at stake. It will then proceed to describe how all these organizations and projects are responding to the challenges ahead. Finally, it will consider how a country like Portugal, with a language that is minor in European terms, but major globally, could prepare for the future envisaged by the EU Directive, and make suggestions of ways in which the educational establishment can contribute.

The challenge of Multilingualism and Multiculturalism in the global village

According to various sources, "history is written by the victors"[6] but, as Ostler (2010) shows, the complexities of the part languages play in these histories are many and varied. The victors usually try to impose their culture, religion, and legal and political systems upon the vanquished, but the process is never simple, and languages often reflect the fusion, assimilation or otherwise of the cultures involved. The awareness today of the need to respect the social, cultural and emotional importance of languages to the people who speak them is behind the European ideal of preserving the languages and cultures of individual countries, at least as far as the official languages and cultures of the countries are concerned. This ideal requires that Europeans should be able to communicate with each other, develop political, commercial and economic unity, and move between countries in a spirit of integration.

There are 24 official or working EU languages and, although some are clearly more equal than others, every effort is made to maintain the multilingual dream, at least in theory. The practical means to this end has been the encouragement of second or third language learning and the education of good translators and interpreters. However, even for these official languages, it has not always been possible to provide full language services, and a multilingual policy that includes the approximately 450 languages of all the migrant people now living and working in the EU complicates the issue considerably. It is one thing to say that the migrants should learn the language of the host country, it is quite another to enforce this. One also needs to take into account the needs of the millions of tourists that visit Europe every year.

The dream of multilingualism is expensive, time-consuming and not always efficient, but (like democracy) "it is the worst form of communication, except for all those other forms that have been tried from time to time" (Winston Churchill – adapted!). The European institutions have found that the ideal is stretched to its limits by the need to provide common political and legal systems for the European Union. The multilingualism policy that covers all EU languages has been under strain for some time, and there are plenty of arguments for limiting the number of languages requiring routine translation and interpreting services. The situation in the police system and criminal courts is under particular strain, as the national governments usually have to pay for the services. Now that immigration has added considerably to the number of languages that

are needed in the courts, the problems are multiplying and the solutions are by no means simple. The issue of multiculturalism, which cannot be separated from multilingualism, is also provoking a variety of reactions at both political and social levels.

Multiculturalism is a politically divisive issue that cannot be ignored. The everyday culture of the city dwelling populations today may be increasingly homogenized by the effect of the media, the Internet and the immediacy of communication in the global village, but this homogenization has, in its turn, produced a reaction that is particularly relevant in relation to the migrant populations. The Law must not only contend with people speaking many languages, it must also try to understand their culture, background, and levels of education. All this has to be taken into consideration if the EU Directive is to be enforceable.

The Law and Language

As volume 6/12 of the EC's 'Studies on Translation and Multilingualism – Language and Translation in International Law and EU Law' (2012) reports, international and EU lawyers/translators strive towards agreement on the legal concepts and the terms that represent them in the different languages. However, there is no guarantee that these terms, agreed by these multilingual, international groups as representing certain well-defined legal concepts, will be interpreted in the same way when used in the essentially monolingual local context of individual countries.

The understanding of legal terminology may not in itself be the main concern in criminal proceedings, but the monolingual, mono-cultural mindset of the representatives of the Law in most local contexts clearly affects their perception of translators and interpreters. The centuries-old discussion on how to make legal discourse as objective and clear as possible, leads to the tendency among legal practitioners to believe that legal language is actually objective, even though they may spend their time searching for loopholes in the law on which to base the cases for their clients. Many students of Law and legal practitioners, like the general public, have rarely had the opportunity to study languages beyond school or conversational level and are not particularly aware of cultures or legal systems outside their own, unless they are forced by circumstances to confront them. The global culture of the educated world we increasingly live in also helps to provide a veneer of uniformity that lulls us into a false sense of mutual understanding, which rarely survives close scrutiny.

The many attempts by legal systems to create a discourse as free as possible from the personal considerations of the individuals concerned, whichever side of the law they are on, and the largely monolingual culture of so many people in any particular country contributes to the idea that translation and interpreting are simply a question of mechanically exchanging one set of words with another. However, well-trained professional linguists are particularly sensitive to the linguistic nuances and cultural differences of the languages / cultures with which they work when confronted with translating and interpreting between two different legal systems.

Translators and Interpreters and the Law

The Directive on the right to interpretation and translation in criminal proceedings (European Parliament and European Council, 2010)[7] starts by drawing attention to the fact that, despite the efforts made by the European Convention for the Protection of Human

Rights and Fundamental Freedom (ECHR), of which all Member States are a party, "experience has shown that that [the ECHR] alone does not always provide a sufficient degree of trust in the criminal justice systems of other Member States"(introductory paragraph 6). Although not explicitly stated in the Directive, one can only presume from its content that this lack of trust derives, at least in part, from deficient, or non-existent, interpreting and translating services within the criminal justice systems.

On the other hand, according to the endless complaints on translators' forums and web pages about the way translators and interpreters are treated by various criminal justice systems across Europe, it would seem that there is a reason why such language services are neither adequate, properly used, nor properly catered for. The remuneration offered is much lower than that which is current in the commercial market and, understandably, it would seem that highly qualified language professionals avoid being available for the work required. This means that the criminal justice systems are forced to use unqualified or, at best, poorly qualified interpreters and translators, a situation that confirms their poor opinion of the services offered, and contributes to the establishment of the low rates that drive good professionals away.[8] So we have a Catch 22 situation here that needs to be addressed.

The problems posed by the Directive and the education of translators and interpreters

The need for interlingual communication has been around since Babel. The very banality of this fact, and the frequently forgotten fact that the majority of human beings actually live in some sort of bilingual society or situation, has contributed to the low status given to translation and interpreting. For those who moved around the known world in past centuries, there was usually some sort of lingua franca (see Ostler, 2010), and bilinguals could sometimes make a living in commercial or political situations. Communication at a more basic level relied on pidgins and sign language.

The low esteem in which translation and interpreting were and often still are held is also to be found in higher education, and these institutions must therefore accept part of the blame for preparing professionals inadequately for the many challenges posed by a multilingual society. The idea of modern language faculties as a 'light' form of higher education for predominantly female students persists, and the academics in these faculties tend to regard actual language learning as the least important part of the curriculum, a servant to the more important areas of literature, culture, or, in some cases, linguistics.

Until relatively recently, training in professional translation was the job of polytechnics devoted to producing office staff with language skills. When the obligation to provide education that might actually lead to jobs forced a crisis in humanities education in general, and modern languages departments in particular, translation gained popularity, once the possibility of providing schoolteachers was exhausted. The staff members responsible for creating the new curricula, however, often have very little idea of how to train (they still prefer 'train' to 'educate'!) translators, and even less of how to work towards the interdisciplinary needs of professional translation. The result is that the graduates from these institutions also contribute to the perception of the poor performance of 'translators'.

There are various forms of interpreting and they can require different levels of training and competence, ranging from simply assisting oral communication in informal sit-

uations through to the highly paid and sophisticated simultaneous interpreting required for international conferences. Yet, here again, even the better-educated general public thinks that 'anyone with languages' can perform on demand mechanically. Nothing could be further from the truth and becoming a good conference interpreter requires especially arduous training, and few manage to complete the serious courses offered satisfactorily.

Many people do not even understand the difference between a translator and an interpreter or the difficulties involved in becoming good at either profession. Many also think that a language student must necessarily perform as well in their foreign languages as they do in their mother tongue. These misunderstandings are not confined to the general public and the legal profession.[9]

Clearly, there is no easy solution to all these problems, as the situations involved will vary widely. European countries differ considerably as to the nature of language use, the extent to which translation and interpreting is required, and between which languages. Even the European Commission has found it more economical to use English as a 'bridge' between languages such as Slovenian and Icelandic or Portuguese, than employ a translator who is fluent in these pairs of languages to translate or interpret between them. Also, providing good language services between most European languages requires a very different approach to that which it is possible to demand between these languages and, for example, a dialect of a sub-Saharan language.

One must also take into account that the criminal situations that require translation and interpreting will also range from fraud by multinational companies to petty theft, and from traffic offences to murder. Multinational companies will be able to pay a team of qualified legal translators and interpreters to avoid conviction for fraud; the illiterate migrant may be dependent on the language skills of a marginally better educated colleague to help him prove his innocence in a case of murder.

In the case of EU languages, one could argue for a full master's level course in community interpreting and translation, and such courses already exist[10], but few countries can afford to dedicate educational resources to providing such specialized education. When there is no market for the language services beyond the occasional court case or community service work, it is unreasonable to expect the authorities to provide, or the private individual to pay for advanced qualifications. The best solution for most EU languages, therefore, is to provide specialized training as part of, or in addition to the general courses in translation and interpreting so that graduates have the qualifications to also earn their livings in the wider market for language services.

In the cases of infrequently used languages for which there is not enough demand for language services to provide a living for even a very few, individuals with good language skills can be given special training to enable them to provide an adequate service, and such courses are already offered by local councils in the larger cities of Europe (for example: the Worker's Educational Association http://www.london.wea.org.uk/community-interpreting in London). The results may not be equivalent to those expected of a highly trained interpreter, but it is the best that can be done. The relative unfairness of these situations may be unacceptable to the ideals of multilingualism and multiculturalism, but it is essential that all concerned recognize the need to be realistic.

In Portugal, the ACIDI – Alto Comissariado para a Integraçao e Diálogo Intercultural (The Comission for Intercultural Integration and Dialogue)[11] provides a telephone translation service that is admirable in the sense that it provides free 'interpreting/translation' services over the telephone to whoever requires them, usually immigrants who seek for knowledge on how to solve their social, economic and legal problems in Portugal. These services are offered by people who speak the languages required but do not necessarily have any training in what they should be doing.

Preparing for the future – international efforts

The EU Directive has no doubt caught many countries and legislations ill prepared for the eventuality of having to provide proper translation and interpreting services in all criminal proceedings. Some countries understand the problems involved and have taken measures to provide training for community interpreters and translators. However, others are either unable to understand the complexities of these services, or are unwilling to pay an appropriate price for them. These factors mean that, even if the educational establishments in these countries undertake to provide courses for such a wide variety of needs, the legal authorities must adapt to the reality of the market, and pay for the training and/or the services of those who receive proper training.

The need to provide training for good translating and interpreting services has been subject to both discussion and effective action for some years, often led by the EC's Directorates for Interpreting (DG SCIC) and Translation (DGT)[12]. The DG SCIC has provided training in conference interpreting to each new country that joins the EU and encouraged a variety of support activities. The DGT has led the development of the European Master's in Translation (EMT) Network[13] since 2006 and encourages close cooperation between universities and the language services profession. The policy for multilingualism is the driving force behind these activities, as was emphasized by the DGT representative, Catherine Vielledent-Monfort, in her presentation at the IAFL conference.

Besides the official EC initiatives, there have been several other projects led by professional organizations and academic institutions, several of which were represented at the conference. Liese Katschinka, the president of EULITA, which was established in 2009 to represent the interests of its members, presented their mission to promote 'the quality of justice, ensuring access to justice across languages and cultures and thus, ultimately, guaranteeing the fundamental principles of human rights as enshrined in the European Convention of Human Rights and Fundamental Freedoms' (EULITA's mission statement – http://www.eulita.eu/mission-statement) as well as its determination to coordinate the efforts of professional individuals, organizations and institutions dedicated to promoting quality in legal interpreting and translation across Europe.

EULITA has helped to promote projects that were presented during the conference. Christiane Driesen presented ImPLI (Improving Police and Legal Interpreting) which ran from April 2011 to September 2012 and was "a comparative study of interpreter-mediated questioning practices – especially by the police – in Belgium, the Czech Republic, France, Germany, Italy and Scotland". TRAFUT – Training for the Future was a project that ran from November 2011 to October 2012 and organized workshops in Slovenia, Spain, Finland and Belgium with a view to preparing the training of translators and interpreters for the outcome of the Directive 2010/64/EU.

Problems in the Portuguese context

One obvious failing of the system is that there are no possibilities in Portugal for either interpreters or translators to register as sworn translators with proper qualifications and status. Anyone who feels qualified can claim to be able to interpret and, when legalization of a translation is required, anyone who has done the translation can go to a notary, who may have no knowledge of the language, and swear that it is a good translation. I would hasten to add that Portugal is not alone in this respect, but it is also true that Portugal loses considerable revenue to Spain when multinationals who insist on a sworn translation have to cross the border to obtain one (Joana Forbes, personal communication).

Whatever the law may say, do, or not do, however, there can be little doubt of two things: that the representatives of the law usually have little understanding of the skills required for interpreting and translation and, consequently, that the proper payment for these services is totally inadequate for the expertise required. Several university courses include community interpreting and legal translation or related topics in their programmes, but their graduates cannot be expected to receive inadequate payment simply because courts see no difference between them and those they so often employ, with minimal or no qualifications to do the work.

Manuel Sant'iago, representing AIIC at the conference, drew attention to the laws, or lack of them, to provide for proper interpreting in legal and criminal cases in Portugal. Although Sandra Silva from the Faculty of Law (see this volume) argues that the Portuguese legal system is technically prepared to meet the demands of the Directive, Sant'iago described how the letter and the practice of the law do not always coincide.

Legal translation has gained some importance as part of the training of translators. However, it is not often easy to find people with an education in law prepared to teach translation, and too many classes are given by language teachers who are out of their depth with any but the simplest texts. On the other hand, lawyers faced with the problem of translating legal texts will often find themselves floundering as they attempt to find cultural, terminological and phrasal equivalents in the other language, as Forbes (2012) demonstrates.

The translation of legal documents is very poorly remunerated by the court system, and it is not enough to say that those who do it will develop the competences and networking contacts to then be paid good rates when working privately for law companies. This does not necessarily happen and by no means relieves the courts of the duty to treat translators with respect.

There is therefore a need for cooperation between the police, the Law and language professionals to work on multilingual matters. Forensic linguistics at all levels can contribute to this cooperation and there is a growing awareness of this.

Forensic linguistics and its contribution to a better understanding of the relationship between the Law and language in Portugal

The need to write clearly and well worries many people, as can be seen in the considerable literature on academic writing or technical communication, but the interest is particularly relevant when it comes to language in legal settings. A recent publication of articles resulting from a seminar on language and the law at the University of Coimbra (Carmo, 2013b) shows a growing understanding among judges, lawyers, and others in

the legal profession of the need for clear and carefully written legal documents, whether they are for legal information for the general public or the judicial decision on a case. Carmo (2013a: 65–74) draws attention to the connection between democracy and this requirement, and Ferreira da Silva (2013: 125–138) in the same volume describes the battle lawyers have fought to make judges evolve from the high-handed, minimal judicial decision to providing a properly argued, well-written document. There is also an article by a forensic linguist using discourse analysis to understand the stress of the position of the accused in court (Carapinha, 2013: 35–64). There are others working in forensic phonetics. However, all this work is essentially monolingual in focus.

Those of us who are involved in the educating of interpreters and translators must now ask ourselves how we can use the nascent interest in forensic linguistics in Portugal to improve the perception and treatment of our graduates. Graduates in translation and interpreting already have many of the intercultural and language skills needed to perform services for the courts. Therefore, if their original course does not already include formal preparation in this area, it should not be difficult to create a short concentrated specialization in the international norms they need to respect, followed by certification after their ability to perform properly in real-life situations has been tested. Courses in legal translation are increasingly included in the curriculum or offered as specializations to professional translators. The people who take these courses could then apply to be evaluated for inclusion in an official register of qualified legal interpreters and translators.

However, there are situations in which it is necessary to seek the help of people with the necessary language skills, but without formal qualifications as interpreters or translators. The market for such qualifications for people who need to communicate between, say, Bulgarian, Finnish and Portuguese is not sufficient to justify a university running a full master's degree in only these combinations. The alternative, someone sufficiently bilingual in these less-spoken languages, but with some other employment, should, however, be offered proper training before appearing in court or in the public services. Such training should be made obligatory, and the educational establishment, together with the professional associations, should work with the police and legal institutions to provide it.

Finally, in our multilingual, multicultural Europe there will always be those with marginal levels of language skills or even education who may suddenly be called upon to interpret or translate. However, even in these cases, it is essential that they should be helped to understand their responsibilities before they are permitted to perform, even if the crash course envisaged lasts only a few hours.

Conclusions

The Directive certainly poses several problems for all those involved in interpreting or translating for the police, the courts and other legal organizations. The work being done by EULITA and others is definitely taking us in the right direction but there is still a long way to go. There needs to be a concerted effort to create bridges between Language and the Law, as proposed by this conference. For this, a good deal of work needs to be done to raise awareness and create mutual respect between all involved. The representatives of the law need to be made more aware of the power of languages, the diversity of languages, and the cultures they represent. On the other hand, those

who prepare interpreters and translators for the needs of the courts must consider not only the need to train university graduates to provide the necessary services and / or provide life-long learning programmes for those already practising as language services providers; they also need to prepare short extra-mural courses for those who will be needed more sporadically.

All this requires those in government to understand what is necessary. The European Commission is already leading the way to providing suitable certification of interpreters and translators who can demonstrate the necessary skills with the TransCert – Trans-European Voluntary Certification for Translators project[14], and the project QUALETRA – Quality in Legal Translation[15]. Governments need to recognize these initiatives by creating the status of sworn interpreters and translators for those who attain the necessary level. Although such a move would necessarily imply a complete revision of the official remuneration at present in force, it would contribute to greater justice by encouraging the employment of the well-qualified professionals who at present try to avoid having their names on the informal list in the drawer of some court official responsible for finding interpreters and translators.

These proposed solutions are not new. The institutions and projects referred to above are leading the way, but there is a long way to go to change the mindsets of those responsible for justice, not to mention the general public's attitude to language services providers.

Notes

[1]DGT – Directorate General of Translation – http://ec.europa.eu/dgs/translation/index_en.htm

[2]EULITA – European Legal Interpreters and Translators Association – http://www.eulita.eu/

[3]AIIC – International Association of Conference Interpreters – http://aiic.net/

[4]TRAFUT – Training for the Future – http://eulita.eu/training-future

[5]IMPLI – Improving Police and Legal Interpreting – http://www.isit-paris.fr/-ImPLI-Project-.html

[6]This quote – or something similar – is ttributed to various people including Napoleon Bonaparte, Winston Churchill and George Orwell.

[7]Available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:280:0001:0007:EN: PDF

[8]The official rate payable to translators in Portugal is 0.027 cents a word; interpreters will receive 100€ for a case, no matter how long the session takes. No travel or other expenses are contemplated.

[9]Most teachers of translation will recognize the sinking feeling experienced when a colleague approaches asking if one's (20-something-year-old Portuguese native speaker) students would like the experience of translating the colleague's thesis into English, or acting as simultaneous interpreter in an international conference on engineering; no mention of remuneration.

[10]For example, the Master's Degree in Intercultural Communication, Public Services Interpreting and Translation, Universidad de Alcalá de Henares, Spain. http://www2.uah.es/traduccion/formacion/master_oficial_POP_EN.html

[11]ACIDI – Alto Comissariado para a Integraçao e Diálgo Intercultural (The Commission for Intercultural Integration and Dialogue) – http://www.acidi.gov.pt/

[12]DG SCIC – Directorate General of Interpreting – http://ec.europa.eu/dgs/scic/index_en.htm; DGT – Directorate-General for Translation http://ec.europa.eu/dgs/translation/index_en.htm

[13]European Master's in Translation (EMT) Network – http://ec.europa.eu/dgs/translation/programmes/emt/network/index_en.htm

[14]TransCert – Trans-European Voluntary Certification for Translators – http://transcert.eu/

[15]QUALETRA – Quality in Legal Translation – http://www.eulita.eu/qualetra

References

Carapinha, C. (2013). A compreensibilidade do discurso judiciário – algumas reflexões. In R. do Carmo, Ed., *Linguagem, Argumentação e Decisão Judiciária.* Coimbra: Coimbra Editores.

Carmo, R. d. (2013a). A exigência e relevância democráticas da compreensibilidade do discurso judiciário. In R. do Carmo, Ed., *Linguagem, Argumentação e Decisão Judiciária.* Coimbra: Coimbra Editores.

R. d. Carmo, Ed. (2013b). *Linguagem, Argumentação e Decisão Judiciária.* Coimbra: Coimbra Editores.

Directorate General of Translation, (2012). *Studies on Translation and Multilingualism – Language and Translation in International Law and EU Law.* Available at http://bookshop.europa.eu/is-bin/INTERSHOP.enfinity/WFS/EU-Bookshop-Site/en_GB/-/EUR/ViewPublication-Start?PublicationKey=HC3012627: European Commission.

European Parliament and European Council, (2010). Directive 2010/64/EU of the European Parliament and of the Council of 20 October 2010 on the right to interpretation and translation in criminal proceedings. *Official Journal of the European Union,* Available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:280:0001:0007:EN:PDF.

Ferreira da Silva, J. A. (2013). Decisão judiciária: processo de elaboração e fundamentação. In R. do Carmo, Ed., *Linguagem, Argumentação e Decisão Judiciária.* Coimbra: Coimbra Editores.

Forbes, J. (2012). A Tradução Jurídica no Contexto da Certificação: requisitos, estratégias e legitimidade do tradutor. Master's dissertation, Faculdade de Letras da Universidade do Porto, Porto.

Ostler, N. (2010). *The Last Lingua Franca.* London: Penguin Books.

The right to interpretation and translation in Criminal Proceedings: the situation in Portugal

Sandra Silva *

Abstract. *The 2010/64/EU Directive is the first step towards the provision of common minimum standards as to procedural safeguards within the European Union, as described on the 'Roadmap for strengthening procedural rights of suspected and accused persons in criminal proceedings'.*

Rather than explain the genesis of the directive or describe its main elements, this article focuses on the national legislation regarding the right to interpretation and translation. From this point of view, it analyses the existing regulations in order to assess if they comply with the criteria set out in the Directive and to discuss what changes have to be made to completely integrate its regulations into the internal legal order.

The main conclusion drawn from this study is that, despite some omissions, the Portuguese legislation accomplishes the minimum standards outlined in the Directive. The article ends with some recommendations to remedy the diagnosed lacks and improve the effective application of the right to interpretation and translation.

Keywords: Right to interpretation, Right to translation, 2010/64/EU Directive, National standards, Portuguese Law.

The need for interpretation and translation

We live in the era of globalization and free movement: people easily move around the globe and the world is within reach at the distance of a low-cost flight. The growing phenomenon of globalization and migrations is responsible for the continuous presence of international and foreign elements in the procedure. This new reality causes significant problems in a multilingual society such as the European Union – a geographic and linguistic puzzle with 23 official languages[1] and a wide variety of local dialects, not to mention the 'foreign' languages also frequently spoken within its doors.

*Faculdade de Direito, Universidade do Porto

The European Union presents itself as an area of freedom, security and justice, a territory with no borders to the free circulation of goods, persons, services and capital. Unfortunately, within Europe there are also no barriers to crime.

To effectively investigate and prosecute ever-increasing cross-border crimes, the European Union institutions have made a significant effort, over the last two decades, to set out various legislative instruments in the field of criminal law. The aim of these instruments is to combat severe criminality effectively, notably by promoting *judicial cooperation* between the national authorities of the Member States[2]. Since the European Council of Tampere (1999), the cornerstone of such cooperation is based on the principle of *mutual recognition* of judgments and other decisions. Naturally, the implementation of such a principle presupposes that Member States *trust* each other's criminal justice systems. The extent of mutual recognition is thus very much dependent on a number of parameters, which include the establishment of common *minimum standards* as to procedural safeguards within the European Union[3]. The 2010/64/EU Directive on the Right to Interpretation and Translation in Criminal Proceedings is the first step towards the provision of those common minimum standards. Further steps are described on a '*Roadmap for strengthening procedural rights of suspected or accused persons in criminal proceedings*', a resolution adopted by the Council on 30 November 2009[4].

Rather than explain the genesis of the Directive or describe its main elements, this article focuses on the national legislation regarding the right to interpretation and translation. From this point of view, it analyses the existing Portuguese regulations in order to assess if they comply with the minimum standards set out in the EU Directive and to discuss what changes have to be made to completely integrate its regulations into the internal legal order[5].

National Standards – do the Portuguese legal regulations accomplish the demands of the Directive 64/2010/EU?

The Portuguese Constitution (CRP)

The right to interpretation and translation is not expressly guaranteed under the Portuguese Constitution. However, it can be established as a corollary of the various fair trial rights laid out in a set of constitutional rules, e.g. the principle of equality (Art. 13), the right to a 'fair trial' (Art. 20, 4), the right of defence (Art. 32, 1) and, regarding the criminal procedure, the guarantee of an accusatory structure and the right to 'material' confrontation (Art. 32, 5). In order to have an effective and not merely formal meaning, these guarantees imply that litigants are able to understand the content of the proceedings (especially the trial), even if it does not take place in a language with which they are familiar.[6]

'A system of justice that allows a litigant to move through the courts without a complete understanding of the proceedings because of a language barrier is', as one may agree, 'an affront to the concepts of due process and equal protection.[7] This is particularly obvious in severe criminal cases, because 'no defendant should face the Kafkaesque spectre of an incomprehensible ritual which may terminate in punishment'.[8]

A defendant who does not speak or understand the language of the proceedings is clearly at a disadvantage. The law must therefore provide for the right to interpretation and translation to remedy this vulnerable situation and ensure an equal treatment (Art. 13).

Moreover, the fundamental fairness required by the 'fair trial or 'due process clause' (Art. 20, 4) – an important dimension of the guarantee to an 'effective protection of judges and courts' (Art. 20, 1) – also entails the state's obligation to provide interpretation and translation when necessary to establish adequate communication between the court and the persons involved in the proceedings. The right to a 'fair trial' regarding criminal matters demands a procedure with 'full defence guarantees' (Art. 32, 1) in a wide sense (technical and material). The right to *technical defence*, i.e. the right to have a counsel (either appointed or chosen), obviously includes the right to freely and meaningfully communicate with one's lawyer. But interpretation is also important to safeguard *material defence*, which means the defendants' right to actively and effectively participate in the proceedings, presenting evidence and challenging the evidence produced against them.

The above rights are an essential feature of the accusatory structure (Art. 32, 5) – a system where the defendant has the status of a procedural 'party' (and not that of an object) and thereby is entitled to some minimum procedural guarantees. One of those guarantees is the 'right of confrontation' (Art. 32, 5 (2)), which is interpreted not in the sense of face-to-face (physical) confrontation but as the requirement that the defendants be given an adequate and proper opportunity to challenge and question the witnesses against them in order to assess their reliability and trustworthiness[9]. No defendant can actually 'confront' witnesses against them without understanding what they are saying in court.

The Code of Criminal Procedure (CCP)

The Portuguese Code of Criminal Procedure also ensures defendants a comprehensive set of procedural rights (art. 61, 1): to be present and to participate actively in the proceedings, to be informed on the charge, to be heard by the court or the judge, to have the assistance of a counsel and to communicate with them, etc.

None of these procedural rights can be effectively safeguarded without the right to interpretation and translation, ruled in Art. 92.

The right to interpretation and translation in the Portuguese Criminal Procedure

Preliminaries

Conceptualization: interpretation v. translation

Despite the lack of distinction in the Portuguese law, interpretation and translation are not identical intellectual operations. Translations are written, as opposed to interpretations, which are oral.

A court interpreter is a 'language mediator' or 'language conduit' whose presence allows a person who does not speak or understand Portuguese to meaningfully participate in the judicial proceedings.[10] The proper role of an interpreter is to place the non-Portuguese-speaker, as closely as is linguistically possible, in the same situation as the Portuguese speaker in a legal setting.[11]

A translator uses different skills than a court interpreter. A translator converts a written document or audiotape recording from the source language (SL) into a written document in the target language (TL). However, the court interpreter performs in some

situations as a translator – e.g. when he or she *sight translates* documents presented during the hearing.[12]

The Costs of interpretation and translation

The Code of Criminal Procedure foresees that, where the right to interpretation and translation applies, it must be provided without costs to the person involved, even when this person is the accused and he or she is convicted at the end of the proceedings (Art. 92, 2 and 3).[13]

Thus, interpretation in criminal proceedings is free of charge to all persons and not only to those who could benefit from legal aid under national laws. A different solution would introduce a (constitutionally prohibited) discrimination based on nationality: Portuguese defendants would be in a better position than non-nationals (since the latter, while having the financial means to afford their own defence, would be obliged as well to pay for an interpreter).[14]

The Portuguese rule accomplishes, thereby, the provisions of the European Convention on Human Rights [= ECHR][15], the European Court on Human Rights [= ECtHR] jurisprudence (see *Lüdicke, Belkacem and Koç v. Germany* (1978) and *Öztürk v. Germany* (1984)[16]) and the 64/2010/EU Directive (Art. 4).[17]

The right to INTERPRETATION
WHO is entitled to the right to interpretation?

The Portuguese Code of Criminal Procedure guarantees the right to interpretation to any procedural participant (defendant, accused, suspect, victim, witness, expert, etc.) whose primary language is not Portuguese and who has a limited ability to read, speak, write or understand Portuguese (whom we may refer to as 'limited Portuguese proficient' or 'LPP' individuals). This right is applicable even when judges, prosecutor and lawyers understand the foreign language spoken by the limited Portuguese proficient person (Art. 92, 2). That is so because the ground for interpretation is not only to enable the court to understand and properly evaluate the testimonies produced, but also to allow the person to understand the statements of the judge, prosecutor and counsels, to hear the testimony of witnesses and to assist in his or her own defence.

Procedural participants with hearing or speech impairments, irrespective of their position in the procedure, are also entitled to interpretation. This type of interpretation presents different problems and is specially ruled in the Code of Criminal Procedure, Art. 93 (thus, accomplishing the 64/2010/EU Directive, Art. 2 (3)).

Nothing is established in the Portuguese law regarding how to assess the necessity of an interpreter. However, there is no doubt that the judicial authorities are competent to decide upon informed discretion whether to appoint an interpreter.

The ECtHR held that the judicial authorities are required to take an active approach in determining the need for interpretation and translation (*Cuscani v. United Kingdom* (2002), § 38 and 39[18]). An attorney's assurance that there is no 'language problem' is thus not sufficient. The competent judicial authority should instead conduct a direct conversation with the defendant to personally determine the extent of his language ability, i.e. how fluent the individual is in the proceeding's language.[19]

The Directive takes this demand a step further, requiring the Member States to 'ensure that a procedure or mechanism is in place to ascertain whether suspected or accused persons speak and understand the language of the criminal proceedings and whether they have the necessity of an interpreter' (Art. 2 (4)). Furthermore, a suspected and accused person must have the right to challenge a decision that finds there is no need for interpretation or translation (Art. 2 (5) and 3 (5)).

Despite the wording of the Directive, there is apparently no need for a complex autonomous verification proceeding inside the criminal procedure (with its inherent additional costs and procedural delay). A brief '*voir dire*' of the individual needing the interpreter would generally be enough. The examination should include not only questions about biographical information, but also open-ended questions such that a non-Portuguese speaker could not anticipate an answer. The following model may be useful for that purpose:

Model voir dire for determining the need for an interpreter

In general: Avoid any question that can be answered with 'yes – no' replies.

Identification questions:
'Ms. _____, please tell the court your name and address.'
'Please also tell us your birthday, how old you are, and where you were born.'

Questions using active vocabulary in vernacular English:
'How did you come to court today?'
'What kind of work do you do?'
'What was the highest grade you completed in school?'
'Where did you go to school?'
'What have you eaten today?'
'Please describe for me some of the things (or people) you see in the courtroom.'
'Please tell me a little about how comfortable you feel speaking and understanding English.'

Source: National Center for State Courts, *Court Interpretation: Model Guides for Policy and Practice in the State Courts* (quoted by Kahaner (2009:227).

When exercising their discretion, judges should always presume a *bona fide* need for an interpreter when a representation is made by an attorney that a defendant or witness has limited proficiency in Portuguese, and requests an interpreter. The 'burden of proof' regarding the ability of the defendant to understand the court's language lies with the judicial authorities, not with the defendant – says the ECtHR (*Brozieck v. Italy* (1989)).

WHAT should be interpreted?

Theoretically, an interpreter may perform three separate functions in criminal proceedings. First, he or she may translate questions posed to, and answers provided by, a non-Portuguese-speaking person during examination by judges, prosecutor and counsels – this function is often called '*witness interpreting*'. The interpreter may also translate communications between the counsel and a party during trial – this service is known as '*party interpreting*' or, since such services are most commonly needed by the defendant, '*defence interpreting*'. Finally, the interpreter (the same or another) may interpret for the defendant, or another party, statements made by the judge, opposing counsel or others during the proceedings – this function is usually referred as '*proceedings interpreting*'.

a) 'Witness interpreting'

The interpreter must, naturally, interpret the communication between the court and the defendant or other LPP individual (witness, codefendant, victim and expert), permitting judges, prosecutor and counsels to question those individuals, to understand their answers and to record their testimony as evidence.

This mode of interpreting privileges the assessment of evidence in the procedure: its purpose is to ensure that a testifying witness or defendant understands and answers the propounded questions, as well as to guarantee that judges understand the persons and the evidence that comes before them. From this perspective, interpretation is mainly an instrument to enable the court to communicate with people who do not speak the court's language, avoiding misinterpretation of testimonies, and thus allowing judges to assess evidence in a foreign language in an accurate and effective manner.

While providing this service, the interpreter uses mainly so called *consecutive interpretation*: he or she listens and speaks in a sequential manner after the speaker has completed a thought. This allows judges, prosecutor and lawyers to pay full attention to the paralinguistic elements of the discourse, including all the pauses, hedges, self-corrections, inflections, and hesitations, tone of voice, demeanor, and body language. Even if the court is unable to speak or understand the defendant's or witness' language, it may still draw inferences regarding these non-verbal elements.[20]

b) 'Party interpreting' or 'defence interpreting'

The Portuguese law recognises, apparently without restrictions, the defendants' right to benefit, without any costs, from the services of another *interpreter* of their choice to mediate the communication between them and their counsel (Art. 92, 3)[21]. These interpreters are subjected to "professional and procedural secrecy", which prohibits them from revealing the content of the communications between the defendants and their counsels (Art. 92, 4). The evidence obtained through violation of those secrecy obligations is obviously rendered inadmissible in court (Art. 92, 5).

There are at least two reasons to allow the presence of a 'party interpreter'.

First, the presence of second interpreter in the proceedings is useful to prevent and detect interpretation errors, when none of the judiciary actors knows the language of the LPP individual.

Second, and most important, the existence of a party interpreter preserves the confidence of the defendants in the justice system and in their lawyers, while guaranteeing the complete confidentiality of the defence strategy and protecting the privileged communication with the defense counsel.[22] Moreover, it prevents the court interpreter, through long association with the defendant, from becoming biased.

Finally, one could add that if court interpreters are interpreting the witnesses' testimonies, the defendants would not be able to communicate with their counsels in order to assist them in the cross-examination. However, in Portugal defendants sit in the center of the courtroom and not side by side with their legal counsel (as happens, for example, in Germany). Thus, they are only able to assist their lawyers before or after (and not also during) the witnesses' examination.

Which conversations must be interpreted?

There is no express ruling in the Portuguese legislation regarding the extent of the interpretation of client-lawyer communications, but probably no court would allow the

presence of a state appointed free interpreter in all meetings between the defendant and his lawyer. This obligation would entail excessive costs for the States and such a right could be subject to abuse (with the defence using the interpretation facilities to slow down the proceedings).

The Directive states that 'free' interpretation shall be available 'where necessary for the purpose of safeguarding the fairness of the proceedings', for communications 'in direct connection with any questioning or hearing during the proceedings or with the lodging of an appeal or other procedural applications' (Art. 2 (2)).

This text calls for two observations.

On the one hand, it is clear that the competent authorities are in a position to refuse 'free' interpretation not only for meetings between the lawyer and the defendant which solely serve dilatory purposes (to prolong the proceedings), but also for those communications that are not immediately connected with official acts of the procedure (for 'not necessary' communications the defendant should bear the costs of interpretation).

On the other hand, the reference to the purpose of safeguarding the fairness of the procedure and the openness to 'other procedural application' allows the jurisprudential expansion of this right through national courts and the European Court of Justice.[23] For example, the preparation of an 'application for bail' (recital 20 of the Directive) or of a requirement to 'instruction' (an intermediate and facultative procedural stage between the investigative phase and trial) can fall under the protection of this rule.

c) 'Proceedings interpreting'?

The Portuguese legislation does not set out other interpretation facilities than the appointing of a 'witness' interpreter and eventually of a defence interpreter. Is it enough, from the point of view of a guarantee of a fair trial, to allow defendants to communicate with the court (and previously with their lawyers), enabling them only to understand the questions directed to them?

Considerations of fairness and the potency of the accusatory system of justice forbid the state from prosecuting defendants who are in effect not present at their own trial.[24] The right to interpretation shall also 'ensure that defendants in a criminal case are put into an equal position to the persons who speak the court's language'[25], enabling them to be linguistically and cognitively present in the hearing and to actively participate in the proceedings. This will balance the vulnerable position of those who do not understand the court's language, ensuring that those persons have access to an effective defence and to a 'fair trial' under the same conditions as any other citizen, i.e. allowing them to be heard and to participate in a meaningful way.

That is not certainly the case when defendants are unable to understand either what is said in the courtroom, because they cannot follow the witnesses' cross-examination, or even what the defense counsel says on their behalf (pleadings). Therefore, it is also necessary that the defendant should be allowed to understand the entire hearing through a proceedings interpreter, who may be seated next to (or behind) the defendant and *simultaneously interpret* the statements of witnesses and everything said by judges, prosecutor and lawyers. That can be done using the 'whispering interpreting mode' or, to avoid any acoustic disturbance, with the help of electronic equipment.

Although the most important dimension of the right to interpretation refers to the defendant, identical considerations could in some extent apply to other procedural participants, specially the victim.

WHO should interpret?

a) The interpreter as an expert witness?

Portuguese law subjects interpreters to an expert's activity rules (Art. 92, 8, making applicable Arts. 153 and 162). The interpreter must therefore be 'sworn', i.e. he or she must solemnly promise the judicial authorities (judge or prosecutor) to make a faithful, accurate and impartial interpretation (Art. 91, 2).

b) Who cannot be an interpreter?

In Portugal interpreters have the same impediments as judges (Art. 47, 1). Consequently, the defendant's (or victim's) wife or husband and other relatives (parents, children, brothers and sisters) are forbidden to be interpreters in the criminal case (Art. 39, 1 (a) and (b)). Others who are forbidden to perform as interpreters are those who have formerly participated in the proceedings as prosecutors, judges, policemen, lawyers, witnesses, etc. (Art. 39, 1 (c) and (d)).

Regarding this point, Portuguese law seems to go even further than the case law of the ECtHR.

For example, in *Berisha and Haljiti v. 'the former Yoguslav Republic of Macedonia'* (2007), the ECtHR considered inadmissible the complaint of an Albanian applicant, who did not have an interpreter in a court hearing and relied on the codefendant's language assistance. In the court's opinion the fact that one of the applicants served as interpreter for the other did not invalidate proceedings, about which they had not complained at the time.

In another case, in the paradigmatic and often quoted *Cuscani v. United Kingdom* (2002) the ECtHR found a violation of Art. 6 § 1 in conjunction with 6 § 3 (e). The trial judge, instead of adjourning the hearing of the Italian defendant (because no interpreter was present), relied on the 'untested language skills' of the applicant's brother to interpret if needed, without consulting the defendant, and did so in a case that led to a four-year prison sentence and a ten-year disqualification as company director.

c) The choice of the interpreter?

Unfortunately, in Portugal there is neither an official certification for interpreters nor a mechanism to easily assess the quality of the interpreter's performance. Furthermore, the law is unclear as to which qualifications a court interpreter must possess and how those qualifications should be obtained (formal educational training, life experience, etc.); It is only necessary that the interpreter is somehow able to render accurate translations. The consequence is that in Portuguese courts almost anyone can be an interpreter: a lawyer present at trial, a bilingual court clerk or even a neighborhood Chinese grocer.

However, it is important to mention that court interpretation is a highly specialized and particularly demanding activity. 'Court proceedings not only involve interactions at a significantly higher level of difficulty than conversational language, but also require a familiarity with legal terminology and procedures and with the cultural context im-

pacting the parties in the court proceedings'.[26] The court interpreter must be able to completely understand and convey the speaker's words and presentation style in the courtroom setting, without editing, adding meaning, omitting or changing colloquial expressions or tone.[27] To be fully competent, an interpreter should therefore possess strong language skills in both Portuguese and the foreign language, 'including knowledge of legal terminology and idiomatic expressions and slang in both the source and target languages, as well as an understanding of geographic differences in meaning and dialect'. But it is also of utmost importance to 'understand the ethical and professional standards and how to apply those standards in a courtroom setting'.[28]

To assume that the mere ability to speak two languages automatically qualify an individual to interpret is, as an Jon A. Leeth noted, analogous to assuming that all people with two hands can automatically become concert pianists.[29]

d) A register of certified or qualified interpreters?

The approach of the ECtHR is that the mere appointment of an interpreter does not absolve the authorities from further responsibility. States are required to exercise a degree of control over the adequacy of the interpretation or translation (*Kamasinski v. Austria* (1989) and *Hermy v. Italy* (2006)), and judicial authorities also bear some responsibility since they are the ultimate guardians of the fairness of the proceedings (*Hermi v. Italy* (2006) and *Cuscani v. United Kingdom* (2002)).

With regard specifically to the quality of interpretation and translation, the ECtHR states that through interpreters and translators the accused or suspected persons must simply be put in position to understand the case against them and to defend themselves, in particular by putting their version of events before the court (*Hermi v. Italy* (2006)). 'Even if the Court has no information on which to assess the quality of the interpretation provided', the ECtHR claims that enough protection is guaranteed when it becomes 'apparent from the applicant's own version of the events that she understood the charges against her and the statements made by the witnesses at the trial' (*Asproftas v. Turkey* (2010)).

Regarding the choice of the interpreter, the ECtHR refuses to adjudicate on the national systems of registered interpreters as such, as it is solely called upon to refer 'on the issue whether the interpretation assistance (…) satisfied the requirements of Article 6', for which it is enough that the interpretation allows defendants to understand the evidence being given against them or to have witnesses examined on their behalf (*Kamasinski v. Austria* (1989)). In these cases, even non-official interpreters are adequate if they have a 'sufficient degree of reliability as to knowledge of the language interpreted' (*Coban v. Spain* (2003)). As for the rest, the ECtHR stated in *Sandel v. 'the former Yugoslav Republic of Macedonia'* (2010) that the court shall not unreasonably delay the proceedings while trying to find a suitable authorized interpreter in the defendant's mother tongue, when an interpreter in another language is sufficient to allow him to understand in essence the proceedings.[30]

The 2010/64/EU Directive is based on the same minimum standards[31] and also places the primary responsibility for quality on Member States, requiring them to take concrete measures to ensure that interpretation and translation is 'of a quality sufficient to safeguard the fairness of the proceedings, in particular by ensuring that the suspect or

accused persons have knowledge of the case against them and are able to exercise their right of defence' (Art. 2 (8), and Art. 3 (9)).

This obligation 'to promote the adequacy of interpretation and translation and efficient access thereto' is bolstered by a requirement that Member States 'endeavour to establish a register or registers of independent translators and interpreters who are appropriately qualified' (Art. 5 (2)). What the national regulations establish as a prerequisite to this kind of 'registration' or 'certification' is a question that I would rather leave open.[32]

The right to TRANSLATION
The scope of the Portuguese Code of Criminal Procedure, Art. 92, is not limited to interpretation of oral statements made at the trial hearing, but also covers the translation of relevant documentary material.

With regard to translation, the CCP refers solely to the necessity to convey into Portuguese documents which are written in a foreign language and not officially translated (Art. 92, 6). According to the Code, this translation facility appears to be simply a way to allow courts the assessment of documental evidence in an accurate and effective manner and not a right of the defence.

However, the Directive 2010/64/EU states that suspected or accused persons who do not understand the language of the criminal proceedings shall be provided with a 'written translation of "all" documents which are "essential" to ensure that they are able to exercise their right of defence and to safeguard the fairness of the proceedings' (Art. 3 (1)).

For this purpose, 'essential documents shall include any decision depriving a person of his liberty, any charge or indictment, and any judgment' (Art. 3 (2)), which comprises the charge, the verdict, the decision imposing preventive arrest, but also a house search warrant. Other documents in the file shall only be translated if the competent authorities consider them essential for exercise of the right of defence or a reasoned request from the defendant or his or her lawyer is made to that effect (Art. 3 (3)). If this request is denied the defendant must have the possibility to challenge the decision (Art. 3 (4)).

The reference to the ability to exercise the 'right of defence' and to the 'safeguard of the fairness of proceedings' sheds light on the nature of which 'other' documents must be translated.

Firstly, it can be argued that the evidentiary material upon which the case effectively rests is always essential to safeguard the right to a 'fair trial' and, despite the exclusion of the reference to 'essential documentary evidence' during the negotiations[33], it must be translated. For example, the written reports with the testimonies of witnesses heard in 'deposition' (Arts. 271 and 294 CPP) should always be translated, since those statements can be used as evidence in trial (Art. 356, 2, a), CPP). This conclusion indeed seems to impose itself if the right to translation is taken seriously and is linked to an effective – and not abstract – implementation of the right to be informed about the 'nature and cause of the accusation' or to 'have adequate time and facilities for the preparation of [the] defence' (art. 6 § 3 (a) and (b) ECHR).[34]

Secondly, it should be noted that, although applicable to the pre-trial phase (as the reference to the 'suspected person' clarifies), the right to translation only extends 'to those documents contained in the case file that, under national law, are already available

to the suspected or accused person, or to his lawyer'[35], because only those are meant to be necessary, at that stage, to the exercise of the defence rights.

In the context of the 64/2010/EU Directive, the translation of essential documents must, as a matter of principle, be provided in written form. In that case, the deadlines for the exercise of procedural rights (e.g. appeal) run from the translation and not from the act itself (e.g. judgment).[36]

Exceptionally, 'an oral translation or oral summary of essential documents may be provided instead of a written translation' when such translation (normally a 'sight translation') does not prejudice 'the fairness of the proceedings' (Art. 3, (7)).

This possibility was inserted in the text because various Member States insisted that admitting oral translations would be very important for daily (court) practice. In less complex cases, the accused would be better served with an oral translation 'on the spot', than with a written translation that could require several days to produce. Additionally, providing this possibility enables a considerable reduction in translation costs.

In support of their position, the concerned Member States relied on case law of the ECtHR, which considered it sufficient for the exercise of their procedural rights (e.g. appeal) to provide the accused with oral information about the content of the indictment or oral explanation of the judgment with the assistance of a lawyer, instead of providing a written translation of those acts (*Kamasinski v. Austria* (1989)[37]). This understanding has been followed by the Portuguese Constitutional Court (decision no. 547/98).

The option for an oral or a written translation must, at least, take account of the complexity of the case. It will be interesting to see how national courts and in the European Court of Justice interpret this provision.

Conclusions

The above analysis allows one to conclude that, despite some omissions, Portuguese procedural legislation accomplishes the minimum standards set out in the 64/2010/EU Directive on the right to interpretation and translation in criminal proceedings.

The Code of Criminal Procedure actually rules on the right to interpretation and translation to all persons (defendant, witnesses, experts) who cannot speak or understand the language of the proceedings because their first (or only) language is other than Portuguese (Art. 92) or they have a speech or hearing impediment (Art. 93). As to the defendant, the right to interpretation and translation appears as a corollary from an effective application of the constitutional fair trial guarantees, particularly the right of defence and the right to be assisted by a lawyer.

It should be nevertheless noted that the exercise of those rights is without any costs to those persons, irrespective their procedural role and the outcome of the proceedings.

The Portuguese legislation is furthermore truly progressive relating to client-counsel communications, since it (theoretically) provides interpretation of such conversations without limitations or restrictions. In this matter, the national rule goes far beyond what is established in the EU Directive.

However, it must be noticed that the law establishes no legal standard by which judges can properly assess the necessity of interpretation, something that is of utmost importance to prevent unknowing violations of the defendant's constitutional rights.[38]

Additionally, in Portugal there is neither an official certification or registration for interpreters nor a mechanism to easily assess the quality of the interpreter's performance. Because those aspects are totally neglected in the Portuguese legislation, some courts routinely allow untrained, non-professional bilingual individuals, without any demonstrated competence, to act as interpreters (including neighborhood grocers).

Moreover, the extent of the right to translation of documents, and the consequences of the lack of a written translation as to the exercise of procedural rights (e. g. the deadline of an appeal) is not clear.

But the major barrier to the effective application of the right to interpretation and translation is still the apparent lack of awareness of language problems and specially a lack of cultural-linguistic sensitivity among some judges, court staff and even defence lawyers.[39]

Under these circumstances, the following recommendations for effective application of the right to interpretation and translation should be attended:

- To clarify the extent of the right to a written translation of some documents and its implications;
- To provide a minimum linguistic training for judges in order to improve their awareness of linguistic difficulties and to enable them to distinguish between litigants who understand rudimentary Portuguese and those who are truly proficient in the language, as well as to meaningfully screen and assess the interpreter's skills;
- To establish official training and certification programs, at least in the most frequently spoken languages, focusing on vocabulary, legal terminology, court procedure, and professional ethics;
- To discuss with interpreters' associations, judges, lawyers and other judicial actors the possibility to enact a Model Code of Professional Responsibility for Interpreters in the Judiciary[40];
- To agree with other Member States the establishment of pools of interpreters, particularly for less-frequently-used languages, that participating States would support through shared resources and coordinated testing and administration[41];
- (In relation to this last case) to promote the use of videoconference and remote interpreting, when a court interpreter is not present and his or her physical presence is not required in order to safeguard the fairness of the proceedings (Art. 2 (6) of the Directive).[42] Although possible, the use of telephonic interpreting has limitations and therefore is not recommended, due to the fact that the lack of visual cues diminishes the capacity of the interpreter to understand the context of the spoken words in the proceeding (including the paralinguistic elements of the discourse).

Despite problems and limitations, one may share an optimistic opinion about the national interpretation and translation panorama – it may be said that in Portugal the matter is 'open to interpretation', much more than 'lost in translation'.

Notes

[1] Those languages are: Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish and Swedish.

[2]As Cras and de Matteis (2010: 153) suggest, the Framework Decision on the European Arrest Warrant (2002) is probably the most well known of these instruments.

[3]See Recital 3 of the 2010/64/EU Directive of 20 October 2010 on the Right to Interpretation and Translation in Criminal Proceedings, published in the Official Journal of the European Union L 280, 26.10.2010.

[4]The 'Roadmap' is part of the Stockholm programme (Official Journal of the European Union C 115, 4.5.2010, p. 1) and calls for the adoption of measures regarding the following rights: the right to translation and interpretation (measure A), the right to information on rights and information about the charges (measure B), the right to legal advice and legal aid (measure C), the right to communication with relatives, employers and consular authorities (measure D), special safeguards for suspected and accused persons who are vulnerable (measure E) and a Green Paper on Pre-Trial Detention (measure F). The first two steps are now accomplished with the adoption of Directives 2010/64/EU of 20 October 2010 on the Right to Interpretation and Translation in Criminal Proceedings and 2012/13/EU of 22 March 2012 on the Right to Information in Criminal Proceedings. Furthermore, on 7 October 2013 the Council adopted the Directive on the Right of Access to a Lawyer in Criminal Proceedings.

[5]In conformity with its Article 9 (1), Member States have until 27 October 2013 to adapt their national laws and regulations to the Directive 2010/64/EU provisions. Since the transitory regime provided for by the Treaty of Lisbon (Article 10 of Protocol 36) is not applicable to this situation, when the period of implementation expires, Member States which fail to transpose the Directive can be subjected to an infringement procedure by the Commission under Article 258 TFEU, including the possible imposition of executive measures and penalties by the European Court of Justice under Article 260 TFEU.

[6]See also the case-law of the ECtHR, *Kamasinki v. Austria* (1989), § 74.

[7]Conclusions of the New York City Bar Association's Committee on Legal Needs of the Poor, quoted by Cardenas (2001: 25).

[8]*United States v. Carrion*, 488, F.2d 12 (1st Cir. 1973).

[9]See also *Van Mechelen and Others v. Netherlands* (1997), § 51.

[10]See Kahaner (2009: 226).

[11]See Cardenas (2001: 27).

[12]*Sight translation* is a hybrid task by which an interpreter reads a document written in one language while rendering it orally into another language.

[13]Unfortunately, the right to interpretation and translation is not part of the 'Letter of Rights' of Article 61, therefore it is not part of the information which has to be read by the police in the present of the suspect (as happens, for example, in Spain).

[14]See Cras and de Matteis (2010: 158).

[15]Article 6 (3) (e) of the ECHR provides that everyone charged with a criminal offence has the right 'to have the free assistance of an interpreter if they cannot understand or speak the language used in court'.

[16]In both procedures, as Germany tried to obtain the reimbursement of interpreting costs from the applicants after their conviction (as then provided for by domestic law), the ECtHR made clear that the term 'free' implies a 'once and for all exemption or exoneration'. Recently, the Court also found a violation of the same provision in the cases *Işyar v. Bulgaria* (2008) and *Hovanesian v. Bulgaria* (2010), where the applicants had been charged for interpretation costs.

[17]Article 4 of the Directive 2010/64/EU states: 'Member states shall meet the costs of interpretation resulting from the application of Articles 2 and 3, irrespective the outcome of the proceedings'.

[18]In this case, the defendant suffered from a hearing impairment.

[19]In criminal matters, where fundamental liberty interests are at stake, a full comprehension of the proceedings is critical. Therefore it is not unreasonable to provide an interpreter at trial for an individual who speaks Portuguese as a second language, even when no interpreter is needed for ordinary conversation purposes. It is quite clear that the knowledge of conversional Portuguese is not enough to understand the level of language that is typically spoken in a courtroom.

[20]Kahaner (2009: 227).

[21]The interpretation of client-lawyer communications appears to be an important corollary of the effective application of the right to counsel's assistance: indeed, how could this right be ensured if the defendant and his or her lawyer were unable to understand each other?

[22]If the same interpreter is used, he or she may eventually and without noticing reveal, when interpreting what the defendant says at trial, something heard before from the accused during the interview with his/her counsel and that should remain confidential.

[23]See Cras and de Matteis (2010: 159).

[24]See Kahaner (2009: 226).

[25]Ramos (2012: 3).

[26]Kahaner (2009: 227).

[27]For example, when the defendant or a witness uses code words, as it happens frequently in drug traffic cases to confuse or conceal information, the interpreter must find a dynamic equivalent in the target language – if the witness uses the term 'soap', which in Portugal is used as a slang term for hashish, the interpreter should find a word which causes the same impact on the audience as the original, instead of translating 'soap' as 'hashish'.

[28]Kahaner (2009: 227–8).

[29]Cardenas (2001: 26).

[30]The authorities have wasted two and a half years trying to find a national Hebrew speaking interpreter (it was prohibited to recruit an interpreter from a foreign country), when it was clear from the outset that an English, Serbian or Bulgarian speaking interpreter would have been sufficient at that stage of the proceedings. This understanding was also affirmed in a case against Portugal (*Panasenko v. Portugal* (2008)). The applicant, a Ukrainian national (on trial for murder of a taxi driver), complained that his interpreter worked into Russian (not Ukrainian) and that even in Russian he was incompetent. During the trial he tried to express his complaints through the interpreter but the presiding judge told both of them not to engage in a discussion. On the basis of a recording supplied by the applicant, the interpreting was admittedly not perfect, but the ECtHR found no violation of Art. 6, § 3, (e) because 'the applicant failed to indicate how the interpreting problems had affected the fairness of the proceedings as a whole' and 'the material in the case file show[ed] that he was able to understand the oral proceedings in essence and present his version of the facts'.

[31]Art. 8 of the Directive includes a 'Non-regression clause', prohibiting any limitation or derogation from any rights and procedural safeguards that are ensured under the ECHR, the Charter of Fundamental Rights of the European Union, relevant provisions of international law and national laws which provide a higher level of protection. Recital 32 of the Directive also explains that the level of protection ensured under its provisions should never fall below the above standards, meaning that the Directive is supposed to be "Strasbourg- and Charter-proof" and should be interpreted and applied in such a way.

[32]As said before, the Portuguese judicial authorities do not demand any specific formal 'qualification' to be a court interpreter. For the most frequently spoken languages, embassies, universities and private agencies are often asked to appoint a qualified interpreter. But any other 'known in court' bilingual individual, more or less qualified, can be called to perform as interpreter. Sometimes, those individuals are simply the Chinese or Indian grocer next-door. In the USA there is a distinction between certified, professionally qualified and language skilled interpreters. Interpreters earn their certification by passing a series of rigorous written and oral exams administered by the Administrative Office of the U.S. Courts and by some States (California, New Jersey, Washington, New Mexico, New York and Massachusetts). The Administrative Office of Courts has developed such exams for Spanish, Navajo, and Haitian-Creole, although only the Spanish exam is currently administered (it is the most common spoken foreign language in USA). Apart from certified interpreters, the Administrative Office classifies two additional categories of interpreters: the *professionally qualified interpreters*, who should previously have been employed as conference or seminar interpreters with any United States agency, with the United Nations or a similar entity, and the *language skilled interpreters*, who are not certified or considered professionally qualified but can demonstrate to the satisfaction of the court their ability to effectively interpret. See Kahaner (2009: 229).

[33]This reference, contained in the original Commission proposal, was not included in the Member –State's initiative, since it met with the firm opposition of a number of national delegations who were

concerned about the financial impact of the need to proceed with translation of such (sometimes rather voluminous) material.

[34]Cras and de Matteis (2010: 159).

[35]Cras and de Matteis (2010: 159).

[36]For example, in *Panasenko v. Portugal* (2008) the ECtHR found a violation of Article 6 §§ 1 and 3 (c), since the defendant missed the appeal deadline partly because the time-limit ran from the service of the judgment in Portuguese and not from the translation.

[37]In *Kamasinki v. Austria* (1989), the ECtHR held that when a translation is necessary not every document has to be conveyed in written form. Oral (sight) translation provided by an interpreter or by the defence lawyer will be sufficient, as long as the defendant understands the relevant documents and its implications. For example, the fact that the verdict is not translated is not in itself incompatible with ECHR Article 6, provided that the defendant sufficiently understands the verdict and the reasoning thereof. The Court determined as well, in *Erdem v. Germany* (1999), that there is no general right of the accused to have the court files translated, since the various fair trial rights are attributed to the defence in general and not to the accused considered separately. 'It therefore suffices that the files are in a language that the accused or his lawyer understands'. Consequently, the applicant had no right to obtain the free written translation into Turkish of the investigation files and a 900-page judgment which, according to the defendant, was 'the accusation against him'. See more about this jurisprudence in Cape and Namoradze (2012: 83–4).

[38]For example, these unknowing violations can happen when a judicial officer fails to take into account the low proficiency of a defendant in trial, by simply believing that he must speak Portuguese because he has lived for a long time in Portugal.

[39]The following example is taken out of an American article, but the scene could easily occur in Portugal: 'The judge sat on the bench. He proceeded to read the sentence into the microphone in front of him in a barely audible, monotone voice. I interrupted, "Excuse me, Your Honor, the interpreter cannot hear you. Can you please check the microphone?" He replied, "You don't have to hear, just interpret!" In shock and disbelief, I interpreted whatever I was able to hear. I occasionally turned to look at the defendant's attorney, but he just sat there silent and motionless. As soon as the criminal sentence was read into the record, I gathered the case information I had not heard. I then shouted out pertinent dates and numbers to the Spanish-speaker man in custody as he was led away by the bailiffs' (retrieved from Cardenas (2001: 24)).

[40]As existing in the USA.

[41]Suggested also by Kahaner (2009: 230).

[42]About the use of these technological resources, see Broun and Taylor (2011).

References

Broun, S. and Taylor, J. L. (2011). *Videoconference and Remote Interpreting in Criminal Proceedings.* Cambridge/Antwerp/Portland: Intersentia.

Cape, E. and Namoradze, Z. (2012). *Effective Criminal Defence in Eastern Europe.* Moldova: Soros Foundation.

Cardenas, R. (2001). 'You don't have to hear, just interpret!': how ethnocentrism in the California Courts impedes equal access to the Courts for Spanish speakers. *Court Review*, Fall, 24–31.

Cras, S. and de Matteis, L. (2010). The Directive on the Right to Interpretation and Translation in Criminal Proceedings. *Eucrim*, 4, 153–162.

Kahaner, S. M. (2009). The administration of justice in a multilingual society – open to interpretation or lost in translation? *Judicature*, 92(5), 226–231.

Ramos, V. C. (2012). Training for effective communication through legal interpreters and translators: The defense counsel's perspective. In *TRAFUT – MADRID, 16.03.2012*: Unpublished conference presentation.

Multilingualism and certainty of law in European Union

Karolina Paluszek [*]

Abstract. *The legal system of the European Union recognizes the general principles of the legal systems of Member States, including the principle of legal certainty. The principle itself is not an uniform rule and has different aspects. The author describes particular approaches towards the principle in relation to multilingualism in the EU, paying particular attention to requirement of publication, principles of clear, understandable legislation, and protection of legitimate expectations. The description is supported by case law and examples from literature and opinions of Advocates-General. The author examines various methods of reconciliation of differences between official versions of EU legislation in order to determine whether they comply with the principle of legal certainty. The conducted analysis proves that the principle of legal certainty is not actually fully observed as it causes numerous problems both for individuals in Members States as well as for the Court of Justice of the European Union.*

Keywords: Legal certainty, multilinugual law.

Introduction: the principle of legal certainty

The legal system of the European Union, being relatively young, recognizes the general principles of the legal systems of Member States. Consequently, the principle of legal certainty, which exists in national legal orders, has been adopted by the Court of Justice of the European Union.[1] However, it has been noted that the content of this principle may vary among national legal systems and EU law (Raitio, 2003; Derlén, 2009). Raitio regards the principle as one that cannot be defined in an exact way (Raitio, 2003: 4). There are many approaches towards the scope and meaning of legal certainty. Legal scholars distinguish several aspects of legal certainty (also called subconcepts – Derlén 2009: 53), namely: the proscription of retroactivity of law (irretroactivity of law); the requirement of publication; *vacatio legis*; principles of clear, understandable legislation; and the protection of legitimate expectations.

---

[*]University of Silesia, Poland

These aspects are sometimes claimed to constitute separate principles.[2] The protection of legitimate expectations may be described as a principle according to which a reasonable person is able to predict the legal consequences of his or her behaviour and expect the authorities to act fairly and reasonably, according to the law. The principle of clear, understandable legislation and the requirement of publication play a crucial role in the analysis of the interaction between the principle of legal certainty and the multilingualism of legislation.

This analysis is an attempt to evaluate the actual use of these principles by EU bodies, in particular the Court of Justice of the European Union. The author raises the questions of how the principle of legal certainty is being respected within the multilingual European Union and whether the methods of interpretation of multilingual legislation comply with this principle.

Multilingualism of EU law

The principle concerning the equality of all official languages is expressed in the primary law of the European Union.[3]

The secondary law of the European Union becomes a part of the national legal systems of the Member States. Some legal acts are directly binding (regulations), others must be introduced in each Member State by national legal acts (Directives). All community legislation must be published in all the official languages of the EU, in order to ensure that the citizens have the possibility to familiarise themselves with their rights and obligations in their native language (if it is an official language of the EU).

This principle was adopted in the Council Regulation 1/58/EEC. The original text of the Regulation considered the languages of four Member States to be official languages of the institutions of the Community. According to Articles 4 and 5 of the original Regulation, regulations and other documents of general application shall be drafted in four official languages, and the official journal of the community shall be published in all of them. Member States and persons subject to their jurisdictions could choose one of the official languages to be the language of their communication with the EU institutions. The number of official languages has increased along with every new Member State of the EU, and necessary amendments have been made to the Council Regulation to include each new language.

Unlike other international organisations[4], the European Union did not decide to choose one or several languages to serve as its official languages. Despite the fact of using three languages as working languages in particular EU institutions (mostly English, French and German), the Union claims to recognize all 24 official languages as equally authentic. No single language version prevails over the others just because that version served as a source for the others in the process of legal drafting. This unique approach towards multilingualism is explained in the Community publications.[5]

According to Article 267 of the Treaty on Functioning of the EU[6], the Court of Justice of the European Union and the Court of First Instance are competent to make judgments on the interpretation of the Law of the European Union. The national courts of the Member States can ask the Court of Justice for a preliminary ruling when they have doubts concerning the interpretation of Community Law. The Court of Justice provides uniform interpretation in all Member States. These judgments from the Court of Justice serve as

important material in understanding the process of interpretation of multilingual law in the EU and form the basis for the following discussion.

Access to legislation in each official language

The obligation of publication EU Law in every official language is subject to the CJEU judgment in the case of *Skoma-Lux sro v Celní ředitelství Olomouc.*[7]

In the judgment, the Court of Justice expressed the right of an individual to familiarise him/herself with the legislation of the Union in any of the official languages. The lack of proper publication in an official language has important legal consequences, resulting in the legislation being unenforceable in that Member State. However, the obligation to publish the legislation in all official languages does not cause many problems in itself. The legislation is accessible through the EUR-LEX database[8], which exists in 23 language versions. The 24th language version was already being prepared when this article was being written (because of the planned accession of Croatia in 2013).

From the perspective of the principle of legal certainty, it can be seen that the purpose of this publication is to protect legitimate expectations. However, what is not so clear is the strength of the position of a single language version, accessible to the public in a particular Member State, compared to other versions in cases where there is doubt about national application and it becomes subject to the interpretation of the Court of Justice of the European Union. Doczekalska describes "a paradoxical situation, in which, on the one hand, multilingualism provides a citizen with the right to his own language, and, on the other, requires him to read all the language versions in order to apply Community law" (Doczekalska, 2009: 362). Consequently, while the lack of publication in a certain language makes the piece of legislation inapplicable to the persons in particular Member State, publication does not result in the right for an individual to refer and to observe only the version drafted in his or her mother tongue.

Importance of multilingualism in the interpretation process

In the case *Srl CILFIT and Lanificio di Gavardo SpA v Ministry of Health*[9], the Court of Justice stated that the multilingualism of EU Law must be taken into account in the process of its interpretation:

> (...) it must be borne in mind that **community legislation is drafted in several languages and that the different language versions are all equally authentic. An interpretation of a provision of community law thus involves a comparison of the different language versions**.
>
> It must also be borne in mind, even where the different language versions are entirely in accord with one another, that community law uses terminology which is peculiar to it. Furthermore, it must be emphasized that **legal concepts do not necessarily have the same meaning in community law and in the law of the various member states**.
>
> Finally, every provision of community law must be placed in its context and **interpreted in the light of the provisions of community law as a whole**, regard being had to the objectives thereof and to its state of evolution at the date on which the provision in question is to be applied.[10]

Particular language versions should be treated equally and, as such, interpretation requires all official language versions to be compared. The declared equality of all language versions results in the acceptance of a fiction that there is no single original version of Community Law, with its translations, but only originals. This has important

consequences for the process of interpretation of Community Law. Firstly, it is very hard, if not impossible, to justify the existence of translation errors, as an original cannot contain any translation errors. Accepting the existence of translation errors challenges the principle of the authentic character of all language versions, as it requires an original, free of errors and therefore more authentic than the questioned text.

Doczekalska (2009: 360) considers that the drafting of the multilingual legislation in the European Union involves processes different from a simple translation. She points out that the various authors' use the term "co-drafting" to describe the revision process conducted by the lawyer-linguists, during which the changes can be introduced into every version, regardless of the fact that one of them is being translated into other languages. During this process, the "translations" can influence the "original" (unlike a simple translation process, in which only the translated versions can be altered). As a result, it is difficult to distinguish translations from originals (Doczekalska, 2009: 361).

It should be noted, however, that the translation process is important with every introduction of a new official language (the legislation must be translated into the new language, but the source language for such translation is not officially prescribed). The new official language then gains the same legal status as the previous ones.

For the above reasons, it would be, in my opinion, unjustified to accept the meaning represented in one version over the others just for the reason that the chosen one was the source for translations into other languages. However, this lack of a reference language (impossible for political reasons) results in the uncertainty for addressees of legal norms – one cannot trust the legislation provided in one's own language.

Methods of interpretation used by the Court of Justice of the European Union to reconcile the diverging language versions

Many classifications of the methods employed by the Court of Justice in the process of reconciliation of differences between official language versions of EU law have been proposed. A brief presentation of the main approaches can be found in Derlén's work (2009: 40), which notes that the scholarly discussion concerning the interpretative agreements adopted by the Court of Justice in the face of diverging language versions displays some confusion and disagreement.

A presentation of the various methods can also be found in the Opinion of Advocate-General Stix-Hackl on the case *Simutenkov v. Ministerio de Educacion y Cultura, Real Federacion Espanola de Futbol*[11]. In order to show that there is no paradigm for "considering" all language versions in the interpretation process of EU Law, I would like to present the various methods distinguished by AG Stix-Hackl, whose task is to prepare a detailed legal opinion on a given case.[12] According to her opinion, the methods of determining the meaning of provisions of multilingual legislation are:

- determining the clearest text by:
    - a) elimination of texts that are not typical
    - b) elimination of versions which contain a translation error
- giving preference to the language versions forming a majority
- favouring a single version over the majority
- accepting the common minimum represented in all versions

Stix-Hackl also mentions one additional method: "proceeding on the basis of the original that served as a source for translations".[13] In my opinion, this method can be

useful only in cases concerning the agreements between the EU and third countries, negotiated and drafted in a certain language, not for the interpretation of legislation developed by co-drafting, the method normally employed by the EU.

Stix-Hackl also underlines the importance of considering the intention of the parties (drafters) and the objective of the provision in the interpretation (the teleological method). This method is frequently applied in the Court of Justice's case law and will be discussed in more detail later.

Examples of judgments by the Court of Justice for all methods, except the elimination of non-typical texts and acceptance of the common minimum, are given in Stix-Hackl's opinion. She disregards the importance of the common minimum, stating at the very beginning of her presentation that it is "supported neither by convincing arguments nor by the practice found in the Court's case law".[14] However, as will be shown later, the acceptance of the common minimum represented in all versions appears in some cases, presented by Stix-Hackl as examples of other methods.

The following analysis of the methods distinguished by Stix-Hackl should prove whether or not their application can be justified from the perspective of the principle of legal certainty.

The method of determination of the clearest text means, in fact, that the equality of languages is not respected and the "best" version is chosen. The application of this method does not allow the addressees of legal norms to foresee the legal consequences of their behaviour by observation of EU legislation in their mother language. Even "clear, non-ambiguous text" does not guarantee the choice of particular version in the interpretation process (because there is a possibility that other versions will be recognized as "more clear").

To disregard of versions that "contain a translation error" is, in my opinion, difficult to justify, because of the authentic character of all language versions. It had been stated before that this method can be useful for the interpretation of agreements between the EU and third countries, but, within EU, where "co-drafting" takes place, it is difficult to identify the translation errors. For example, in the judgment of the Court in case *Erich Stauder v City of Ulm*[15], referred to by Stix-Hackl, the question of differences between particular language versions is explained in the observations of the Commission, which stated that the Management Committee at one of the meetings held on 29.01.1969 decided to modify the draft of the examined provision. Two language versions were overlooked in the modification process and amended after the publication of the act. However, this appears to be a mistake in the drafting process and not a translation error. The issue raised by the Commission had been mentioned as grounds for the Court of Justice's judgment, but it does not seem to be the most important reason for its decision. It rather supports the other grounds of the judgment explained by the Court of Justice. It is therefore surprising Stix-Hackl quotes this as her only example of "elimination of versions which contain a translation error". In my opinion the judgment could serve as example for the common minimum method, which will be analysed in another part of the paper.

The disadvantage of the "favouring the majority" method is that a similar meaning represented in various language versions might result from similarities between certain legal systems or languages from the same roots (Romance languages, Slavic languages).

Because of these potential similarities and the the unequal number of languages in different groups, the choice of a version basing on the greatest number of language versions containing the same meaning cannot be regarded as sufficiently justified. Stix-Hackl observed in her opinion that giving preference to the meaning of the majority does not form an absolute rule. She cites a case[16], where the Court of Justice had chosen the meaning of the provision in question represented in a single language version over all others on the basis of the purpose of the provision.

The method of elimination of non-typical versions is similar to the "favouring the majority method". Therefore, it does not need to be analysed separately.

The method of teleological interpretation (based on the objective, purpose of the legislation or intention of the drafter), appears in some judgments indicated by Stix-Hackl and can be found in up-to-date judgements. The Court of Justice usually refers to the previous cases by pointing out that:

> it is to be borne in mind that it is settled case law that the different language versions of a text of European Union law must be given a uniform interpretation and hence, in the case of divergence between the language versions, the provision in question must be interpreted by reference to the purpose and general scheme of the rules of which it forms a part[17]

The application of teleological method requires the comparison of official versions and a choice of the meaning that complies with the purpose of the disputed provision. This method is very frequently applied by the Court of Justice, underlining its importance in achievement of uniform application of EU law in all Member States. However, its application can hardly be justified from the perspective of legal certainty. The correct interpretation of the rule requires an individual to compare all official versions.

The "common minimum" method is the only one criticized by Stix-Hackl in her opinion without reference to any judgment of the Court (she finds no case law and no convincing arguments supporting it). Stix-Hackl does not explain the method, apart from saying that "one should take the common minimum of all language versions as a starting point [for the interpretation – KP], and accept that there is merely an obligation to use endeavours."[18] According to this opinion, employment of this method would result in the acceptance of the most liberal interpretation of the disputed provision. Therefore, in this article, the common minimum method will be understood as acceptance for the interpretation that results in the lowest level of burdens and obligations for the addressees of the provision, or, in other words, the interpretation assuring the highest level of freedom.

As stated before, the elements of the "common minimum" method can be found in the judgement of the Court of Justice in case of *Erich Stauder v City of Ulm*[19], indicated by Stix-Hackl as an example of the method of elimination of the versions containing a translation error. The Court of Justice stated:

> When a single decision is addressed to all the Member States the necessity for uniform application and accordingly for uniform interpretation makes it **impossible to consider one version of the text in isolation but requires that it be interpreted on the basis of both the real intention of its author and the aim he seeks to achieve, in light in particular of the versions in all four languages**. In a case like the present one, **the most liberal interpretation must prevail, provided that it is sufficient to achieve the objectives**

> **pursued by the decision in question. It cannot, moreover, be accepted that the authors of the decision intended to impose stricter obligations in some Member States than in others.**[20]

It is clear that the Court of Justice, seeking the uniform interpretation and application of the disputed provision, accepted the most liberal interpretation that fulfilled the condition of achieving of the purposes of the legislation.

Among the latest judgements combining the common minimum method with the teleological method, the case of *The Queen, on the application of M and Others v Her Majesty's Treasury*[21] is very interesting from the perspective of legal certainty. Apart from repeating the importance of taking into account the purpose and general scheme of the whole legislative act, the Court of Justice refers to the **need of clarity of the legislation, for the reasons of assurance of legal certainty**. The Court of Justice stated that:

> It is to be borne in mind that, in construing a provision of secondary European Union law, preference should as far as possible be given to the interpretation which renders the provision consistent with the general principles of European Union law and, more specifically, with **the principle of legal certainty** (C-1/02 Borgmann [2004] ECR I-3219, paragraph 30 and case-law cited).

> That principle requires legislation, such as Regulation No 881/2002, **which imposes restrictive measures having considerable impact on the rights and freedoms of designated persons** (Kadi and Al Barakaat International Foundation v Council and Commission, paragraph 375[22]) and which, as provided in Article 10 of that regulation, in domestic law involves penalties, in this instance criminal penalties, for infringement of those measures, **to be clear and precise** so that the persons concerned, including third parties such as the social security bodies involved in the main proceedings, may **know unambiguously their rights and duties and take measures accordingly**.[23]

Finally, the chosen meaning of the provision in question is the most liberal one and falls within the scope of all analysed language versions. Therefore, the judgement also represents the method of a common minimum.

The variety of methods applied in the course of reconciliation of diverging language versions of EU law causes the feeling of uncertainty as to the consideration of all language versions in the process of interpretation. The additional problem is that, in many cases, it is not even obvious whether all the languages were really examined. The Advocates-General in their opinions and the Court in its judgements commonly fail to explain the meaning represented in each language.[24] Usually they use vague expressions like "other language versions" or "some language versions"[25], so it is difficult to state if all versions were examined, and, if not, which were subject to interpretation, and why they had been chosen over the others. The application of one or other of the methods in a particular case does not seem to depend on the character of the problem, so, even by achieving the comparison of the then all 23 versions (difficult even for a court composed of nationals from the different Member States, and almost impossible for a single person), it is hard to foresee the final interpretation of a provision in question. The aim of achieving a uniform interpretation in the whole community is often used by the Court of Justice to justify its interpretative decisions.[26] However, in doing so the Court of Justice

may disregard meanings represented in certain language versions, even the languages native to the parties involved in the particular dispute.[27]

The analysis of reconciliation methods distinguished by Stix-Hackl proves that the only method that allows factual consideration of all language versions, without eliminating any in the interpretation process, is the acceptance of a common minimum of all versions. The provisos are that the scope of a term or phrase is in question, and that a common interpretation for all official versions exists. In examined case law, the common minimum method has not been applied independently of others. The result of its application has been accepted only when the application of a liberal interpretation also enabled the achievement of the purpose of the interpreted legislation. In my opinion, the common minimum method does not contradict the intention of achieving a uniform interpretation of EU legislation in the Member States, nor the principle of equality of languages. The case law analysed confirms that the results of this method satisfy the requirements of legal certainty, particularly in cases involving infringements of obligations and penalties. However, the applicability of this method in other types of cases requires separate study.

Final remarks

Legal certainty is recognized as one of the principles of European Union's law. Accomplishment of this principle is very difficult in multilingual community, especially with regard to requirement of publication, principles of clear, understandable legislation, and protection of legitimate expectations.

The EU declares to recognize the equal authenticity of all official languages and the right of an addressee of legal rules to obtain EU legislation in his/her mother tongue, as confirmed in case of *Lux sro v Celní ředitelství Olomouc*[28]. By publishing EU law in all official languages, the Community enables the citizens to understand their rights and obligations resulting form that law. Nevertheless, the expectations followed from observation of a single language version do not find protection in the course of interpretation and application of EU law. According to case law, the interpretation requires the comparison of different official versions of the disputed act.

Consequently, citizens cannot predict the result of a legal dispute by relying on the text drafted in their mother tongue. Therefore, the publication and accessibility of a particular language version do not guarantee addressees the real possibility to recognize their rights and obligations. The aim of achieving a uniform interpretation of a disputed provision prevails over the expectations of the addressees formulated on the basis of one version.

The existence of variable methods of solving problems resulting from differences between language versions of Community Legislation does not help to achieve legal certainty for citizens of particular Member States. Not only can't they rely on one version, but they also cannot foresee which criteria will be applied in the course of reconciliation of divergences between various language versions. Moreover, there is no certainty as to which language versions will be compared.

In my opinion, the "common minimum" method of interpretation allows the realization of the principle of equality of language versions, as well as the principle of legal certainty. Individuals can trust the legislation drafted in their own language by reading

their rights and obligations. However, a more detailed analysis of the applicability of the common minimum method in different types of cases requires separate study.

Cases of the Court of Justice cited

Case C-161/06 *Skoma-Lux sro v Celní ředitelství Olomouc*, [2007] ECR I-10841.

Case 29/69, *Erich Stauder v City of Ulm*, [1969] ECR 00419.

Case C-283/81, *Srl CILFIT and Lanificio di Gavardo SpA v Ministry of Health*, [1982] ECR 03415.

Case 76/77, *Auditeur du Travail v. Bernard Dufour, S.A. Creyf's Interim and S.A. Creyf's Industrial*, [1977] ECR 02485.

Case C-265/03, *Igor Simutenkov v Ministerio de Educación y Cultura and Real Federación Española de Fútbol*, [2005] ECR I-02579.

Case C-174/05, *Stichting Zuid-Hollandse Milieufederatie, Stichting Natuur en Milieu v College voor de toelating van bestrijdingsmiddelen*, [2006] ECR I-02443.

Case C-187/07, *Criminal proceedings against Dirk Endendijk*,[2008] ECR I-02115.

Case C-340/08, *The Queen, on the application of M and Others v Her Majesty's Treasury*, [2010] ECR I-03913.

Joined Cases C-230/09 and C-231/09, *Kurt Etling, Thomas Etling* (C-230/09), *Hauptzollamt Oldenburg* (C-231/09) *v Theodor Aissen, Hermann Rohaan*, [2011] ECR I-03097.

Case C-419/10, *Wolfgang Hofmann v Freistaat Bayern*, [not yet published].

Case C-510/10, *DR, TV2 Danmark A/S v NCB – Nordisk Copyright Bureau*, [not yet published].

Case C-19/11, *Markus Geltl v Daimler AG*, [not yet published].

Notes

[1]The complex study of legal certainty in the European Union was conducted by J. Raitio (2003).

[2]J. Raitio claims the protection of legitimate expectations and principle of non-retroactivity of law to be principles frequently connected to the Legal Certainty Raitio, 2003: 4.

[3]Article 55 of the Treaty on European Union (ex article 53 TEU) states: *This Treaty, drawn up in a single original in the Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish languages, the texts in each of these languages being equally authentic, shall be deposited in the archives of the Government of the Italian Republic, which will transmit a certified copy to each of the governments of the other signatory States.*

[4]E.g.The United Nations has six official languages.

[5]According to the European Commission's booklet (2009): *EU documents are laws which have to be applied by governments, by companies, and by citizens.* **The public must be able to read and understand them***, and their elected representatives must be able to debate in Europe without being hampered by language difficulties. When decisions are taken, they need to know exactly what they are signing up to. They require – and deserve – the level playing field of a full multilingual regime.*

[6]Article 267 of the Treaty on Functioning of the EU: The Court of Justice of the European Union shall have jurisdiction to give preliminary rulings concerning:

(a) the interpretation of the Treaties;

(b) the validity and interpretation of acts of the institutions, bodies, offices or agencies of the Union;

Where such a question is raised before any court or tribunal of a Member State, that court or tribunal may, if it considers that a decision on the question is necessary to enable it to give judgment, request the Court to give a ruling thereon.

Where any such question is raised in a case pending before a court or tribunal of a Member State against whose decisions there is no judicial remedy under national law, that court or tribunal shall bring the matter before the Court.

If such a question is raised in a case pending before a court or tribunal of a Member State with regard to a person in custody, the Court of Justice of the European Union shall act with the minimum of delay.

[7]Judgment of 11.12.2007 in case *C-161/06Skoma-Lux sro v Celní ředitelství Olomouc.* [2007] ECR I-10841, in which the Court of Justice stated:

1. Article 58 of the Act concerning the conditions of accession to the European Union of the Czech Republic, the Republic of Estonia, the Republic of Cyprus, the Republic of Latvia, the Republic of Lithuania, the Republic of Hungary, the Republic of Malta, the Republic of Poland, the Republic of Slovenia and the Slovak Republic and the adjustments to the Treaties on which the European Union is founded, precludes the obligations contained in Community legislation which has not been published in the Official Journal of the European Union in the language of a new Member State, where that language is an official language of the European Union, from being imposed on individuals in that State, even though those persons could have learned of that legislation by other means.

2. In holding that a Community **regulation which is not published in the language of a Member State is unenforceable against individuals in that State**, the Court is interpreting Community law for the purposes of Article 234 EC (Case C-161/06, emphasis added).

[8]http://eur-lex.europa.eu

[9]Judgment of 6.101982. In case C-283/81, *Srl CILFIT and Lanificio di Gavardo SpA v Ministry of Health.* [1982] ECR 03415.

[10]Judgment of 6.101982. In case C-283/81, *Srl CILFIT and Lanificio di Gavardo SpA v Ministry of Health.* [1982] ECR 03415paragraphs 18-20, emphasis added.

[11]Opinion delivered on 11 January 2005 on the case C-265/03, *Igor Simutenkov v Ministerio de Educación y Cultura and Real Federación Española de Fútbol.*[2005] I-02579.

[12]The legal basis for the position of the Advocate-General is article 252 of the Treaty on the Functioning of the European Union that reads: 'The Court of Justice shall be assisted by eight Advocates-General. (...). It shall be the duty of the Advocate-General, acting with complete impartiality and independence, to make, in open court, reasoned submissions on cases which, in accordance with the Statute of the Court of Justice of the European Union, require his involvement'.

[13]The opinion was given on the interpretation of the Agreement between the European Union and a third country (Russia). The agreement was negotiated in English.

[14]AG Stix-Hackl opinion on the case C-265/03, paragraph 16.

[15]Judgment of 12.11.1969 in Case 29/69, *Erich Stauder v City of Ulm*, [1969] ECR 00419.

[16]Judgment of 15.12.1977 in Case 76/77, *Auditeur du Travail v. Bernard Dufour, S.A. Creyf's Interim and S.A. Creyf's Industrial*, [1977] 02485.

[17]E.g. the Court of Justice's judgments: of 26.04.2012 in case c-510/10, DR and TV2 Danmark A/S v NCB - Nordisk Copyright Bureau, [not yet published]; of 3.04.2008 in case C-187/07, Criminal proceedings against Dirk Endendijk.[2008] ECR I-02115; of 9.03.2006 in case C-174/05, *Stichting Zuid-Hollandse Milieufederatie and Stichting Natuur en Milieu v College voor de toelating van bestrijdingsmiddelen.*[2006] ECR I-02443; of 29.04.2010 in case C-340/08, *The Queen, on the application of M and Others v Her Majesty's Treasury*, [2010] ECR I-03913.

[18]Opinion delivered on 11 January 2005 on the case C-265/03, *Igor Simutenkov v Ministerio de Educación y Cultura and Real Federación Española de Fútbol.* [2005] ECR I-02579, para.16.

[19]Judgment of 12.11.1969 in Case 29/69, *Erich Stauder v City of Ulm*, [1969] ECR 00419.

[20]Judgment of 12.11.1969 in Case 29/69, *Erich Stauder v City of Ulm*, [1969] ECR 00419, Grounds of the judgement, paragraphs 3-5, emphasis added.

[21]Judgment of 29.04.2010 in case C-340/08, *The Queen, on the application of M and Others v Her Majesty's Treasury*, [2010] ECR I-03913

[22]The Court refers to joined Cases C-402/05 P and C-415/05 P *Kadi and Al Barakaat International Foundation v Council and Commission* [2008] ECR I-6351.

[23]Case C-340/08, *The Queen, on the application of M and Others v Her Majesty's Treasury*, [2010] ECR I-03913, paragraphs 64 and 65, emphasis added.

[24]Some judgments and opinions contain analyses of all languages (for example the Stix-Hackl opinion on the case C-265/03 Opinion delivered on 11 January 2005 on the case C-265/03, *Igor Simutenkov v Ministerio de Educación y Cultura and Real Federación Española de Fútbol*.[2005]ECR I-02579), or at least expressions indicating that such examination had been conducted (for example judgment of 5.05.2011 in joined cases C- 230/09 and C-231/09, *Hauptzollamt Koblenz v Kurt und Thomas Etling in GbR* (C-230/09) and *Hauptzollamt Oldenburg v Theodor Aissen and Hermann Rohaan* (C-231/09),[2011] ECR I-03097 , paragraph 64.

[25]See: judgment of 28.06.2012 in case C-19/11, *Markus Geltl v Daimler AG*,[not yet published] where the Court of Justice presents the differences between Italian, French and Dutch language versions on the one hand, and "the others", referring to the Danish, Greek, English and Swedish (second possible meaning), Spanish, Portuguese (third possibility) and German (paragraph 42 of the judgement). The Court of Justice underlines in paragraph 44 that it has analysed all official language versions existing at the time of adopting the interpreted directive. The Advocate-General Mengozzi refers in his opinion on the Case to the "majority of other versions" (French, Italian, English, Spanish, Dutch, Maltese, Romanian, Estonian, Finnish, Latvian, Swedish, Polish, Slovak and Portuguese) being contradictory to the German version (paragraph 63 and reference 10 of the opinion). There is no information on the examination of other versions. In the judgment of 26.04.2012 in Case C-419/10 *Wolfgang Hofmann v Freistaat Bayern*, [not yet published], the Court of Justice uses in the paragraph 69 of the judgements the expressions: "some language versions (…) and particularly German version",(…) "large number of other versions" (French and English indicated as examples). Advocate-General Bot does not refer to the differences between language versions at all.

[26]E.g. Case C-340/08 *The Queen, on the application of M and Others v Her Majesty's Treasury*, [2010] ECR I-03913; Case 29/69 *Erich Stauder v City of Ulm*, [1969] ECR 00419.

[27]E.g Case 76/77, *Auditeur du travail v Bernard Dufour, SA Creyf's Interim and SA Creyf's Industrial* [1977] ECR 02485, Case 29/69 *Erich Stauder v City of Ulm*, [1969] ECR 00419.

[28]Judgment of 11.12.2007 in case *Skoma-Lux sro v Celní ředitelství Olomouc.* [2007] ECR I-10841.

References

Commission, E. (2009).   *Multilingualism – a bridge to mutual understanding*. Luxembourg: Office for Official Publications of the European Communities, 19, retrieved on 28.02.2013 from http://ec.europa.eu/dgs/translation/publications/brochures/index_en.htm edition.

Derlén, M. (2009). *Multilingual Interpretation of European Community Law*. Alphen an den Rijn: Kluwer Law International.

Doczekalska, A. (2009). Drafting and interpretation of eu law – paradoxes of legal multilinualism. In G. Grewendorf and M. Rathert, Eds., *Formal Linguistics and Law*, 339–370. Berlin: Mounton de Gruyter.

Raitio, J. (2003). *The Principle of Legal Certainty in EC Law*. Dortrecht: Kluwer Academic Publishers.

And Justice for All:
Non-Native Speakers in the American Legal System

Leah M. Nodar [*]

Abstract. *The American Constitution promises the right to due process of law. However, some groups, including speakers of English as a second language, have historically been denied this right. The 1978 Court Interpreters Act guaranteed interpreters for anyone who 'speaks only or primarily a language other than the English language.' Decisions as to who speaks 'primarily' a foreign language are left to the presiding judge. Today, this means that non-native speakers in federal courts will generally receive highly qualified interpreters. In state courts, however, judges often rule that these speakers do not need interpreter assistance. In this paper, I analyze three dozen judicial opinions containing choices about language proficiency. These choices can be subjective, and sometimes imply that conversational English and legal English are equally difficult. This is problematic, especially since works by authors such as Lippi-Green and Nguyen have shown that judges sometimes rule that these speakers cannot be understood in educational or media environments.*

Keywords: Access to interpreters, non-native speakers, language evaluation, state court.

American Interpretation Laws

One of the main topics of the recent IAFL conference was Directive 2010/64/EU and the host of changes it will entail for legal interpretation and translation within the EU. The upcoming challenges are multi-faceted, but one area that should not be ignored is who decides whether an interpreter is needed in the courtroom, and how that decision is made.

Oddly, the EU can use America's example as both a guiding light and a cautionary tale. The American court system is bifurcated into a federal system and a state system. The federal system is generally considered a model of interpretation excellence. The state system is not.

---

[*]University of Mississippi

The laws involving interpreters are based around the Fifth and Sixth Amendments to the American Constitution, guaranteeing an impartial trial and due process of law, respectively. Interpreters are guaranteed because without them non-native English speakers cannot understand or respond to the charges against them, the most fundamental aspect of due process (Epstein and Walker, 2010: 517–139, 552–570).

In federal courts, these rights are most specifically enshrined through the FCI (2010: 28 U.S.C. § 1827 and 1828). This act states in part that a qualified interpreter's presence is guaranteed in 'all proceedings…conducted in, or pursuant to the lawful authority and jurisdiction of a United States district court' for 'people who speak only or primarily a language other than the English language.' Further legislation has established criteria for certifying qualified court interpreters, leading to the extremely rigorous Federal Court Interpretation Exam (FCIE). The languages tested by the FCIE have varied over the years, but currently the most common language tested is Spanish, where the pass rate for legal interpreters is 4% (Bussade, 2010). The Act was designed with a specific subsection of cases in mind – federal criminal defense. Other cases, such as civil court cases, are not mentioned and non-native speakers continue to have less access to qualified interpreters in these cases than in criminal cases.

The push for access to interpreters at the state level has been more recent, even though the relevant laws are older. The Civil Rights Act of 1964 prohibits discrimination based on, among other things, national origin (Questions, 2011). The idea that this includes linguistic features such as language and accent has slowly gained strength, and in 2000 Executive Order 13,166 required that any recipient of federal funding – which includes most state and local courts – review and improve access for Limited English Proficiency (LEP) speakers (Executive Order 13,166, Title 3 C.F.R., 2001 comp signed 11 August 2000). Although this was a sweeping order covering a wide variety of funding recipients, the website set up to coordinate the response of these recipients (http://www.lep.gov) spells out the specific demands upon the legal system, saying that 'nearly every encounter an LEP person has with a court is of great importance or consequence to the LEP person…[The Department of Justice] emphasizes the need for courts to provide language services free of cost to LEP persons' (Questions, 2011). Since then, many state legislatures have passed statutes concerning court interpretation, most of which state that a judge may appoint an interpreter whenever necessary. Typical is Mississippi's law, saying 'a court interpreter shall be appointed when the judge determines, after an examination of a party or witness, that: (a) the party cannot understand and speak English well enough to participate fully in the proceedings and to assist counsel; or (b) the witness cannot speak English so as to be understood directly by counsel, court and jury. […] The court should examine a party or witness on the record to determine whether an interpreter is needed if: (a) A party or counsel requests such an examination; (b) It appears to the court that the party or witness may not understand and speak English well enough to participate fully in the proceedings; or (c) If the party or witness requests an interpreter' (Miss. Code Ann. § 9-21-79, 1972, 2010). State laws are interpreted as applying to all cases, including civil cases, though they are not always carried out this way in daily practice. All of these state laws assume that the judge is capable of effectively assessing language ability.

Such laws may suggest to the layman that an interpreter is automatically received whenever a non-native speaker makes a request. However, many outside factors can influence the decision to engage or not engage an interpreter.

One is budgetary. Though many areas have access to interpreters of common languages such as Spanish, a qualified interpreter for a less common language may be hard to come by. The small town of Oxford, Mississippi, recently had a state-level trial-court case involving a Tagalog speaker; unsurprisingly, the court did not have a system in place to deal with the situation. Finding a qualified Tagalog interpreter would require paying not only interpreter fees, but also lodging and transportation(Bussade, 2010). Interpreters for extremely uncommon languages might need to be flown in from New York City or Los Angeles, and such spending in such a small town could create serious difficulties in taking care of other justice needs. For extremely small courts, even common languages can pose a problem. Spanish-speaking attorney Domingo Soto commended the integrity of the judges of the tiny town of Andalusia, AL, for consistently hiring him to drive the two hours from larger Mobile, AL, and translate, despite the cost to them; in some of the other small courts in the region, due process is ignored (Soto, 31 March 2011).

A second problem, sometimes related to the first, is time. The difficulties of securing an interpreter can affect another Constitutional right: the Sixth Amendment right to a speedy trial. This was another aspect of the problem for the small-town judge dealing with a Tagalog speaker – securing and transporting a qualied Tagalog interpreter could take an inordinately long time. In this case, the judge chose to allow an untrained family member to interpret (Bussade, 2010). Unfortunately, the quality of such interpretation 'is quite poor indeed' (Berk-Seligson, 2002: 9). One employee of a small Mississippi Justice Court said that, as far as she knew, no interpreters had ever been hired by the court; non-native speakers were expected to bring their own (Lafayette County Justice Court, 2011).

Even with Spanish interpretation, demand can outstrip supply. Judge Michael Mc-Maken, a state district court judge in the larger city of Mobile, Alabama, speaks of the delay and difficulty in getting an interpreter, saying that the interpretation laws were designed for much larger and higher-level courts where single cases last for days or even weeks, and not for the hectic pace of district courts. A federal court that covers a state might have a few thousand cases a year; a municipal court may have tens of thousands, and scheduling problems grow proportionately. To try and balance the competing needs for understanding and speed, Judge McMaken also allows some family interpreting, assigns Spanish-speaking defendants as much as possible to Spanish-speaking attorneys, and has even learned some Spanish himself (2011).

The most qualified interpreters (i.e., those who have passed the FCIE or are otherwise certified) are most easily engaged for federal criminal cases. Federal criminal defense attorney Carlos Williams said that asking a judge for an interpreter is 'a formality,' and said they are provided as a matter of course whenever requested (2010). As noted, however, this request can cause difficulties in smaller courts. Moreover, in these small courts interpreters are not only harder to find, but may not be as qualified. Bilingual attorney Soto, for example, refers to a case he saw where the interpreter seemed to catch only every other word to explain why judges may prefer to continue in broken English rather than rely on possibly poor interpretation (15 Feb 2011).

These problems are common in small court systems. They are not, however, considered acceptable by the Department of Justice. United States Assistant Attorney General Thomas Perez addressed some areas where non-native speakers' needs are not being fully met in a letter to the courts:

> Some courts only provide competent interpreter assistance in limited categories of cases [i.e., only in criminal cases, not civil]…Many courts…authorize one or more of the persons involved in a case to be charged with the cost of the interpreter…[in order] to discourage parties from requesting or using a competent interpreter. […] Some states provide language assistance only for courtroom proceedings…DOJ continues to interpret Title VI and the Title VI regulations to prohibit, in most circumstances, the practices described above.(Perez, 2010: 2-3)

Subjective Factors in Language Evaluation

With this understanding of the legal framework of court interpretation and some of the factors that affect it in everyday use, we can move to a discussion of the kinds of linguistic factors that affect a judge's ability to evaluate language. It may seem that a court-ordered judgment of a non-native speaker's English ability should be a simple thing. Many factors, including phonology, prosody, grammar, and lexicon, influence accent, intelligibility, or both. Surely, then, one could judge ability by summing up such factors – lower intelligibility would be caused by less English-like use of speech features, and more English-like usage would cause a more English-like and less foreign-sounding accent with greater intelligibility. However, there is an underlying assumption here that does not stand up to scrutiny: the idea that the native speaker judging the speech is objective.

Research in the field of accent perception has shown that listeners are influenced not just by what they hear, but also by what they expect to hear. This phenomenon has been dubbed Reverse Linguistic Stereotyping (RLS) – when assumptions about a speaker's group affect judgment of that person's speech (Kang and Rubin, 2009: 441–456). In one groundbreaking study of RLS, a female Caucasian native English speaker was audiotaped giving a lecture, as if for a college course. Students were played this audiotape and shown a photograph of either a Caucasian woman or an East Asian woman. Students shown the East Asian photo reported hearing an East Asian accent in the recording and performed significantly worse on tests of the lecture material (Rubin, 1992: 511–531). Repetitions and variations on this experiment have shown similar results, including one with a recording of a male voice with a slight Dutch accent which was marked as having an East-Asian accent by students shown a photo of an East Asian man and as having a Standard North-American accent by students shown a photo of a Caucasian man (Rubin *et al.*, 1999: 1–12).

Other studies have shown other ways in which judgments of accent can be unrelated to reality. Niedzielski (1999: 62–85) played a recording of a native Detroit speaker to Detroit students, telling some the speaker was from Michigan and others that she was from Canada. The speaker, like most in the Detroit area, had an accent with certain vowels (particularly the /aw/ diphthong) affected by Canadian Raising, making them higher and more fronted than in the Standard variety of English. This feature is stigmatized by Detroit residents, who consider it a feature of Canadian, not Detroit, speech. Students given the recording with the Michigan label marked the vowels incorrectly, labeling them as Standard English vowels. However, students given the same recording and told it was

of a Canadian speaker marked the vowels correctly, hearing and noting the effects of Canadian Raising on the vowels.

These two studies show expectations influencing perception in opposite ways. When a speaker is expected to speak with a foreign accent, a foreign accent is heard; when a speaker is expected to speak with a standard accent, a standard accent is heard. This by itself suggests that subjective perception of accent is not reliable. Beyond that, two other major outside influences on language ability assessment deserve recognition – familiarity and motivation.

'Familiarity' here can mean familiarity with one particular variety of non-native English, which helps comprehension of that variety. It can also mean familiarity with non-native speakers of English in general. If raters of English-language ability are familiar with non-native speakers (for example, if they have non-native friends) they will tend to give higher ratings of comprehensibility and intelligibility in non-native English speech (Kang, 2008: 181–205). Rubin (nd) puts this into the context of one common college experience when he argues that more contact with non-native teaching assistants ('sticking with' their classes) leads to improved listening skills for students over the long term, and thus to better understanding of other non-native teaching assistants. Gass and Varonis (1984: 65–87) found that the single most important aspect of familiarity for facilitating conversation between native and non-native speakers was familiarity with topic; familiarity with the particular speaker, familiarity with the speaker's variety of English, and familiarity with non-native speech in general all also made communication easier. This understanding, however, is not necessarily two-way. A judge who deals frequently with non-native speakers may grow more familiar with the patterns of non-native speech and find it easier to comprehend, but that doesn't mean that the speaker is equally familiar with the patterns of English.

Another important non-speech factor in language assessment involves personal attitude towards the 'communicative burden.' Lippi-Green argues that

> [w]hen speakers are confronted with an accent which is foreign to them, the first decision they make is whether or not they are going to accept their responsibility in the act of communication…[M]embers of the dominant language group feel perfectly empowered to reject their role, and to demand that a person with an accent carry the majority of responsibility in the communicative act. […] Accent…can sometimes be an impediment to communication [.] In many cases, however, breakdown of communication is due not so much to accent as it is to negative social evaluation of the accent in question, and a rejection of the communicative burden. (Lippi-Green, 1997: 70–71)

This idea has been borne out by studies such as that of Lindemann (2002: 419–441), where native English speaking students were evaluated on their perceptions of Korean students and then set to do a task with native Korean speakers. Students with positive perceptions of Korean students used collaborative conversation strategies; or, to use Lippi-Green's terminology, they took up their share of the communicative burden. These students completed the task successfully and also regarded themselves as successful. Students with negative perceptions of Korean students sometimes rejected their share of the communicative burden, using conversation strategies that Lindemann labels as problematizing (denigrating the contributions of their partners) and avoidance (using passive-listener strategies, such as leaving out information or failing to ask questions).

Those who used avoidance strategies failed at the task, and all of these students with negative perceptions, regardless of actual success or failure, regarded themselves as unsuccessful. This occurred despite the fact that the Koreans students were the same each time (that is, each Korean student completed the task twice). Motivation – the simple desire to communicate – leads to more successful and more satisfactory communication.

Judgments of language ability are not based purely on objective factors. Expectation, familiarity, and motivation all play a role in affecting assessment of foreign-accented speech. Judging language ability accurately is therefore very difficult, especially for the untrained.

Language Evaluation in American Courtrooms

To look at how judges in American courts evaluate language, I read hundreds of judicial opinions and found 36 that specifically mentioned how the court dealt with language proficiency. These opinions were found using the LexisNexis Academic's 'US & State Legal Cases' search engine to find opinions containing phrases such as 'English second language,' 'non-native speaker,' 'Limited English Proficient,' and 'no interpreter.' In order to make sure results were relevant in light of current legislation, all opinions used were from February 2009 to February 2011.

Of the 36 cases found, twenty-two were federal United States district court (U.S.D.C.) cases, two were U.S. circuit court (federal appeals) cases, ten were state appeals cases, and two were state Supreme Court cases. Most of the judges in these courts were not themselves evaluating a person's language proficiency, but were instead evaluating the evaluation of a lower court. These kinds of opinions are useful because they show not only how the trial court made an opinion, but also whether an appellate judge considers the decision acceptable. There are a few different standards an appeals court might use, but in these cases the appellate judges usually decided whether or not a trial court decision was 'clearly erroneous.' They do not retry the evidence themselves, but rather look to see if the lower court made an obvious error or abused its discretion. If the appellate court finds that a compelling argument can be made for either side of an issue, then the decision is not 'clearly erroneous' and is affirmed.

The higher-level court cases are also important because they may establish precedent. They result in a decision that judges of lower courts within their jurisdiction must abide by in later cases, also known as being binding on those courts. The United States Supreme Court is the highest court in the nation, meaning that its opinions are binding for all other courts in the U.S. Other courts have smaller regional areas where they establish precedent. For example, a state appeals court's decision may be binding for all the trial courts within a state. Of the cases included in this analysis, six establish precedent.

Not every case ends in an opinion. The type of court that tends to have the most difficulty properly using interpreters, state trial courts, is also a type of court that does not lend itself to opinions. For one, there is no lower jurisdiction to explain precedent to; for another, the cases are rarely controversial; for a third, as noted above, the case loads in such courts are far greater than in higher courts, leaving little time for writing. Despite the fact that it would focus my data on other courts, I chose to look at judicial opinions for two reasons. One was simple feasibility – I could not perform a nationwide study of trial court cases with Limited English Proficient speakers. The other was importance. Trial court judges can vary greatly, and a single decision made in a small court may not

make a major difference in the area. Higher court opinions make greater waves. Many lower court judges will read higher court opinions to understand the reading of the law, making sure that they themselves have a correct reading so that fewer of their cases will be overturned. Collections of opinions make up case law (as opposed to statutory law), which is often what lawyers build their cases upon.

Each of these cases involved a non-native English speaker and the question of whether that person spoke English well enough to legally take a particular action. Specific actions varied, and thus the standard used by judges varied as well. The two most relevant and largest groups were cases involving the ability to legally stand trial in English, which requires that the speaker be 'competent,' and the ability to legally waive a right, which requires enough English to make the waiver 'knowing, intelligent and voluntary.' These two kinds of cases were at the core of 24 opinions.

Of the 36 cases, in only 4 did the judge decide in favor of the non-native speaker. Two cases, one state appeals and one federal district case, help show why these cases so frequently go against the non-native speaker. When trial court cases are decided against a defendant, that person will often try to get the conviction overturned on any grounds they can think of, including claiming insufficient English ability.

Thus, for example, in *State v. Nieves* (Court of Appeals of Ohio, 8th District, 2010; binding on Cuyahoga County), the appellate court ruled against a non-native speaker who said he did not understand enough English to make a guilty plea, since it found that the trial court judge had apparently offered him an interpreter and he had refused, saying that 'he learned English after moving from Puerto Rico to New York when he was five years old; he took classes in English since he was seven years old; he went to school through the twelfth grade; he was employed locally in the criminal justice system as a corrections officer; and, he has been in an English-speaking environment for 28 years.'

In *United States v. Nguyen* (U.S.D.C. for the District of Massachusetts, 2009), the non-native speaker had not only told the trial court judge that he did not need an interpreter, but was provided with one anyway, who sat nearby in case he wanted something translated. His ineffective assistance of counsel claim was rejected.

These claims, which the judge in Nguyen described as 'disingenuous,' are not uncommon. An interpreter I interviewed said that circumspect attorneys will bring her to meetings and have her translate for non-native speakers who have been in the country for over twenty years and are completely fluent in order to try and stave off such litigation (Bussade, 2010). It is not unusual for people who have been ruled against in court to attempt to have the ruling overturned on any grounds they can think of, and in this analysis that manifests itself as stating a lack of understanding regardless of actual proficiency.

Another point worth making is that many of the criminal case opinions reference clear evidence that the non-native speaker was very likely guilty. In some of the cases involving a non-native speaker giving consent to search, for example, the search turned up damning evidence. The speaker would have a strong motive to exaggerate language troubles in order to get the evidence suppressed.

That said, there are some areas where judges lean on very subjective language evidence when they rule against non-native speakers. Statements of officers, detectives, and other officials that a non-native speaker spoke English well are used to counter state-

ments by the speaker suggesting lack of understanding. However, for much the same reasons that a speaker might want to exaggerate language problems, the police have a strong motivation in these cases to overestimate a speaker's ability, e.g. if a search turned up useful evidence, they would need to make the case that the search was legally consented to by the non-native speaker. To put it another way, one attorney I interviewed said that police tend to decide that a suspect speaks fluent English 'if he knows how to order a beer' (Soto, 2011).

In *Ivanova v. Astrue* (U.S.D.C. for the Northern District of Texas, 2010) and in the precedents cited by some judges, it seems that the level of English proficiency required to receive interpreter assistance must be at or very near zero. Ivanova is the only case here where a judge specifically evaluated someone's speech and found she did not speak enough English to take an action – the judges in the other three cases decided in favor of the non-native speaker shied away from such specifics and spoke in broader terms – and the person in question spoke only 'a few words' of English. In the precedent case *Gonzalez*, cited in *United States v. Putrous* (U.S.D.C. for Eastern District of Michigan, 2010), the non-native speaker's English was so poor that he 'initially did not realize he was attending his own trial.' *State v. Sun Yong Kish* (Court of Appeals of Minnesota, 2009) is a notable exception here, since the precedent cited, Farrah, was a case where a defendant's 'trouble understanding' and 'problems in communication' overrode an officer's statement that he felt he and the non-native speaker had understood each other.

In several cases, including *Suda v. Stevenson* (U.S.D.C. for the District of South Carolina, 2009), *United States v. Silva-Arzeta* (U.S. Court of Appeals, 10th Circuit, 2010), *State v. Lunacolorado* (Court of Appeals of Oregon, 2010), and *State v. Mohamed* (Court of Appeals of North Carolina, 2010), the judges state that such linguistic showings as answering yes/no questions, having telephone conversations (with no indication of complexity), and responding to simple orders indicate a level of English proficiency that includes the ability to comprehend what it means to waive a right. Similarly, in *Isanan v. Johnson* (U.S.D.C. for the Western District of Virginia, 2009) the judge states that because the non-native speaker was able to get a GED while in jail, '[c]ommon sense dictates [that he had] adequate proficiency in English' to do legal research. These cases show a blind eye to the enormous difference in complexity between conversational and legal English. Such rulings are examined in a concurring opinion of impressive linguistic depth in *Lunacolorado*. The judge writes:

> There is a vast difference [...] between being able to carry on a conversation in English and being able to understand and waive constitutional rights. [...] The language used in courts and legal proceedings is much more complex than conversational English. [...]. Moore and Mamiya[1] cite studies finding that 'the difficulty of court language [is] at the 14th grade level for Spanish' and 'court language is at the 12th-grade level plus technical legal language.' [...] Even if defendant was able to carry on a conversation in English, there is no evidence in this record that defendant's understanding of English is at the 12th-grade level that Moore and Mamiya suggest is necessary for defendant to understand the officers' questioning and for defendant's subsequent waiver of his constitutional rights. However, there is...no requirement under Oregon law that a defendant understand English at the 12th-grade level. (*State v. Lunacolorado*, Court of Appeals of Oregon, 2010)

This opinion sums up a troubling aspect of these cases: knowledge of relatively simple English is assumed to imply understanding of legal-level English. In the clear majority of the cases seen here, if the non-native speaker spoke *some* English, then the defendant spoke *enough* English. The concept of familiarity, as discussed above, may be playing a role here in leading judges to underestimate the difficulty for others of a variety of speech they themselves are familiar with. This mingling of conversational and legal English is in stark contrast to the situations shown in the works of Lippi-Green (1997: 259–60, 160) and Nguyen (1993: 1325–1361), whose research involves non-native speakers in Equal Opportunity Employment cases seeking approval of their English ability. In these cases, the courts rarely or never conflate conversational and business English. Speaking conversational English does not lead to an assumption of enough English skills for the workforce, even though (depending on the job) legal English is likely to be much more difficult than English in the workplace and is less likely to be similar to the English learned at school or in daily life. The lack of objective guidelines (Lippi-Green discusses 31 cases where intelligibility 'was a matter of opinion only'), means it is possible that non-native speakers might be found poor enough at English to bar them from jobs, but good enough to not need an interpreter in court!

The most recent court evaluations may show evidence that these problems are being addressed. The ruling in *Narine v. Holder* (U.S. Court of Appeals, 4th Circuit, 2009) referenced only a 'non-native English speaker,' without specific notes of very low proficiency. Even more unusual, in *Ling v. State* (Supreme Court of Georgia, 2010) the court found that conflicting evidence was enough to issue a call for a new trial, even though the dissenting judge listed evidence similar to that in many of the other cases here when he argued that the non-native speaker had been competent to stand trial with no interpreter. These two cases may show signs of a shift away from previous standards, where conflicting evidence generally led to a decision of competency, and towards a new standard requiring more evidence before a non-native speaker is considered capable of taking a legal action without an interpreter present.

Methodology Issues and Further Research

My research was constrained by its use of judicial opinions, and especially by finding them through LexisNexis. There was no way to search for every possible phrase a judge could use to reference language ability. Using 'key phrases' to find opinions led to a sizable group of relevant cases, but many other relevant opinions that used different terminology were left out. Even some of the cases given here had further appeals that did not come up in the searches. Most notably, in *Ramos-Martinez v. United States* (U.S.D.C. for the District of Puerto Rico, 2009), the judge stated that he knew and trusted the 'seasoned' previous judge and lawyer, and therefore the non-native speaker must have spoken fluent English. This appalling judgment was taken to task, reversed, and remanded for further proceedings in a later appeal which did not show up in my original searches.

Furthermore, while these opinions are useful in a broader view of the law, a better understanding of how language evaluations are made in the courtroom would be gained by traveling to state district courts around the nation and recording such evaluations as they occur.

Finally, this study was hampered by my lack of legal knowledge. In the same way that I have argued that judges are not linguistically nuanced, I am not legally nuanced, and there may be aspects to this problem that I am simply not seeing.

The judge's statement in Lunacolorado that non-native speakers need a 12th-grade level of English to understand legal language suggests an area that deserves further consideration: many native speakers of English do not have this level of schooling and may therefore be similarly lost during legal proceedings. Native English speakers may speak dialects completely different from the professional dialect used in the courtroom, and the subjective evaluation of understanding may apply to them as well. Further work on the broader effect of language issues on defendants is merited.

Conclusions

If speakers are waiving rights without knowing why or going through trials without understanding their content, justice is not being served. However, a standardized system of testing or evaluation would face serious problems. In a system where getting an interpreter can be a major time and budget concern, attempting to have trained linguists evaluate every non-native speaker would be impractical. And although the use of objective English tests such as the TSE would help in the employment cases discussed in Nguyen and Lippi-Green, they would be less helpful in the legal context discussed here, where a non-native speaker with a strong motive for showing lack of understanding could purposefully fail. An experienced judge is surely the most straightforward means of separating the fraudulent cases from the meaningful. However, this means that judges must step up their linguistic understanding.

The simplest answer is to educate judges on the issues, for example by presenting at judiciary conferences. This is the approach used by Dr Bussade (2010), who suggests that before deciding if a person needs an interpreter the judge should ask questions requiring more than a yes/no answer – asking 'Where were you born?' rather than 'You were born in Brazil, correct?', for example, elicits an answer that gives a better indication of English skill (even better is 'Tell me about your hometown').

It is also important for judges to realize that legal language is more complex and less commonly taught than conversational English, and that knowledge of the latter does not prove knowledge of the former. Judges are trained in law, and so used to the complexities of legalese that they may not always realize just how bizarre the language can sound to the layman, much less to the foreign layman. Even in cases where judges do make an excellent effort to simplify legal matters, as in *Suda*, the non-native speaker may not understand; simple yes or no answers to complex questions do not by their nature 'make[ ] it [...] clear' that a non-native speaker understands English. Asking the speakers to rephrase their rights in their own words would leave much less doubt.

A basic shift in perspective, pushed along by the Department of Justice and the executive branch, seems to be in its early stages. In *Ling v. State*, a case that caused a great deal of controversy, the Supreme Court of Georgia says that a list of evidence similar to that of many other trials is 'conflicted' and does not clearly show English language ability. Greater pressure has been placed on the lower courts to show that a non-native speaker can fully understand and participate in the legal process before denying an interpreter. I welcome this development. Judges are not trained in language evaluation, and attempting to bring trained language evaluators to every case would be nigh-on im-

possible. Thus, a little humility about the high likelihood of subjective factors affecting judgment is necessary. If some evidence gives the impression that a non-native speaker has a high level of English ability and other evidence suggests the opposite, greater weight should be given to the latter; currently, it seems the former is worth more.

A recent development is the rise of the federal Telephone Interpreting Program (TIP), where non-native speakers are given small microphones and earpieces for live interpretation and have their own speech interpreted through a speaker system to the court. These systems are currently more likely to be in place in federal courts, and not in state courts (Soto, 31 March 2011). Many interpreters at the IAFL conference spoke out against telephone interpreting, and it certainly has its flaws. Technical problems remain serious, and such interpretation is not enough for every situation; in trials where many people are speaking, for example, a telephone interpreter with no visual cues may become confused as to who is speaking when (Cruz *et al.*, 2009). However, if more state courts can insert these systems, I believe that they could do great good. The choice currently is not usually between a family interpreter or a certified, competent live interpreter; it is between a family interpreter or none at all. Adding a third option via the TIP has been an important step forward in access to interpreters, and increasing the reach of the program to smaller courts would increase non-native speakers' access to due process.

Redrawing the line between 'some' English ability and 'enough' English ability may lead to fluent speakers receiving interpreters they do not need, but that outcome is surely better than speakers who do need interpreters failing to receive one. The level of English ability required to deal with legal issues should be considered higher than that shown in conversational English. Greater use of qualified interpreters both in person and through the Telephone Interpreting Program will lead to greater access to justice for non-native English speakers.

Notes

[1]The judge refers to Mamiya as if he were one of the editors of Immigrants in Courts. He is not, but Moore and Mamiya collaborated on the included article "Interpreters in Court Proceedings" (29—45), which contains the reference to 12th-grade level English being required for the courtroom.

Cases cited

All cases were accessed through the LexisNexis database.

Search term 'English second language'

Akhavein v. Argent Mortg. Co. No. 5:09-cv-00634 RMW. U.S.D.C. for the Northern District of California. November 2009. 2009 U.S. Dist. LEXIS 102666.

Fessehazion v. Hudson Group. 08 Civ. 10665 (BSJ) (RLE). U.S.D.C. for the Southern District of New York. August 2009. 2009 U.S. Dist. LEXIS 74443.

Gonzales v. Martel. NO. CV 08-4087-AHS (MAN). U.S.D.C. for the Central District of California. December 2009. 2009 U.S. Dist. LEXIS 126547.

Isanan v. Johnson. Case No.7:09CV00411. U.S.D.C. for the Western District of Virginia. October 2009. 2009 U.S. Dist. LEXIS 101174.

Livsey v. Adventist LaGrange Mem'l. Case No. 09 CV 3129. U.S.D.C. for the Northern District of Illinois. July 2010. 2010 U.S. Dist. LEXIS 69687

People v. Charles. 957/2007. Supreme Court of New York. June 2009. 2009 N.Y. Misc. LEXIS 1835.

People v. Stewart. A121972. Court of Appeals of California, 1st Appellate District. January 2010. 2010 Cal. App. Unpub. LEXIS 321.

State v. Nieves. No. 92797. Court of Appeals of Ohio, 8th Appellate District. February 2010. 2010 Ohio App. LEXIS 448.

State v. Sanjeep Shrestha. A09-1424. Court of Appeals of Minnesota. May 2010. 2010 Minn. App. Unpub. LEXIS 460.

Suda v. Stevenson. Civil Action No. 3:09-1077-MBS-JRM. U.S.D.C. for the District of South Carolina. November 2009. 2009 U.S. Dist. LEXIS 124125.

Tumbaco Chavez v. Belleque. Civil No. 08-287-BR. U.S.D.C. for the District of Oregon. January 2011. 2011 U.S. Dist. LEXIS 8152.

United States v. Diaz. CASE NO.: 1:09cr9-SPM. U.S.D.C. for the Northern District of Florida. October 2009. 2009 U.S. Dist. LEXIS 106600.

United States v. Resendez-Sanchez. No. CR S-05-0323 GEB DAD P. U.S.D.C. for the Eastern District of California July 2010. 2010 U.S. Dist. LEXIS 72164.

Yazzie v. Astrue. CIV 08-8059-PCT-MHB. U.S.D.C. for the District of Arizona. September 2009. 2009 U.S. Dist. LEXIS 90295.

Search term 'non-native English'

State v. Cruz. Appeal No. 2009AP2106. Court of Appeals of Wisconsin, District One. December 2010. 2010 Wisc. App. LEXIS 1035.

Narine v. Holder. No. 08-1299. U.S. Court of Appeals, 4th Circuit. March 2009. 2009 U.S. App. LEXIS 5336.

United States v. Sylla, et al. 08-CR-906(KAM). U.S.D.C. for the Eastern District of New York. February 2010. 2010 U.S. Dist. LEXIS 13056.

Search term 'Limited English Proficient'

Ling v. State. S10G0460. Supreme Court of Georgia. November 2010. 2010 Ga. LEXIS 890.

Lopez v. Bay Shore Union Free Sch. Dist. 09-CV-502 (ADS) (ARL). U.S.D.C. for the Eastern District of New York. November 2009. 2009 U.S. Dist. LEXIS 104894.

United States v. Silva-Arzeta. No. 07-5140. U.S. Court of Appeals, 10th Circuit. April 2010. 2010 U.S. App. LEXIS 8660.

Search term 'no interpreter'

Avichail v. St. John's Mercy Health Sys. Case No. 4:09CV00069 ERW. U.S.D.C. for the Eastern District of Missouri. January 2011. 2011 U.S. Dist. LEXIS 1493.

Bui v. Astrue. Civil No. 08-1279-CL. U.S.D.C. for the District of Oregon. July 2009. 2009 U.S. Dist. LEXIS 80928.

In re I.R.T. NO. COA10-790. Court of Appeals of North Carolina. December 2010. 2010 N.C. App. LEXIS 2322.

Ivanova v. Astrue. Civil Action No. 3:09-CV-2349-K (BH). U.S.D.C. for the Northern District of Texas. May 2010. 2010 U.S. Dist. LEXIS 54588.

Kakal v. State. No. 0-779/09-1422. Court of Appeals of Iowa. February 2011. 2011 Iowa App. LEXIS 90.

Maldonado v. United States. No. C 08-4074-MWB, (No. CR 05-4084-MWB). U.S.D.C. for the Northern District of Iowa. January 2010. 2010 U.S. Dist. LEXIS 3676.

People v. Guevara. B216485. Court of Appeals of California, 2nd Appellate District. July 2010. 2010 Cal. App. Unpub. LEXIS 5955.

Ramos-Martinez v. United States. Civil No. 08-1388(JAF), Crim. No. 01-638. U.S.D.C. for the District of Puerto Rico. May 2009. 2009 U.S. Dist. LEXIS 38774.

State v. Lunacolorado. A139805. Court of Appeals of Oregon. November 2010. 2010 Ore. App. LEXIS 1431.

State v. Mohamed. NO. COA09-943. Court of Appeals of North Carolina. July 2010. 2010 N.C. App. LEXIS 1270.

State v. Sun Yong Kish. A08-1342. Court of Appeals of Minnesota. August 2009. 2009 Minn. App. Unpub. LEXIS 884.

United States v. Garcia. CR 05-30128. U.S.D.C. for the District of South Dakota, Central Division. October 2009. 2009 U.S. dist. LEXIS 99919.

United States v. Nguyen. CRIMINAL NO. 05-10277-PBS. U.S.D.C. for the District of Massachusetts. March 2009. 2009 U.S. Dist. LEXIS 27604.

United States v. Putrous. Criminal No. 03-80743, Civil No. 08-13047. U.S.D.C. for Eastern District of Michigan. September 2010. 2010 U.S. Dist. LEXIS 94507.

United States v. Thepmontry. CR 03-30103-01. U.S.D.C. for the District of South Dakota, Central Division. December 2009. 2009 U.S. Dist. LEXIS 116755.

Zeciri v. Luna. Case No. 3:08-cv-001-JWS-JDR. U.S.D.C. for the District of Alaska. September 2010. 2010 U.S. Dist. LEXIS 105318.

References

Berk-Seligson, S. (2002). *The Bilingual Courtroom: Court Interpreters in the Judicial Process.* Chicago:: University of Chicago Press, 2nd ed. edition.

Bussade, J. (2010). Personal Communication. Personal interview.

Cruz, R., Dann, R., Estill, J., Festinger, N., Feuerle, L., Framer, I., Harpstrite, P., Kaufman, K., Kelly, N., Lindquist, P., Oakes, G., Palacio, M. and Zarenda, N. (2009). NAJIT Position Paper: Telephone Interpreting in Legal Settings.

Epstein, L. and Walker, T. (2010). *Constitutional Law for a Changing America: Rights, Liberties, and Justice.* Washington, DC: CQ Press.

Executive Order 13,166, Title 3 C.F.R., (2001 comp., signed 11 August 2000). Improving Access to Services for Persons with Limited English Proficiency.

FCI, (2010). Federal Court Interpreters Act of 1978, 28 U.S.C. § 1827 and 1828.

Gass, S. and Varonis, E. M. (1984). The Effect of Familiarity on the Comprehensibility of Non-Native Speech. *Language Learning*, 34(1), 65–87.

Kang, O. (2008). Ratings of L2 Oral Performance in English: Relative Impact of Rater Characteristics and Acoustic Measures of Accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 6, 181–205.

Kang, O. and Rubin, D. (2009). Reverse Linguistic Stereotyping: Measuring the Effect of Listener Expectations on Speech Evaluation. *Journal of Language and Social Psychology*, 28(4), 441–456.

Lafayette County Justice Court, (2011). Personal Communication. Telephone interview.

Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States. *Language in Society*, 31(3), 419–441.

Lippi-Green, R. (1997). *English with an Accent: Language, ideology, and discrimination in the United States.* New York: Routledge, Inc.

McMaken, M. (2011). Personal Communication. Telephone interview.

Miss. Code Ann. § 9-21-79, 1972, (2010). Miss. Code Ann. § 9-21-79, 1972.

Nguyen, B. B.-D. (1993). Accent Discrimination and the Test of Spoken English: A Call for an Objective Assessment of the Comprehensibility of Non-Native Speakers. *California Law Review*, 81, 1325–1361.

Niedzielski, N. (1999). The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology*, 18(1), 62–85.

Perez, T. (2010). Language Access Guidance Letter to State Courts.

Questions, (2011). Questions and Answers.

Rubin, D. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, 33(4), 511–531.

Rubin, D. (n/d). Help! My Professor (or Doctor or Boss) Doesn't Talk English.

Rubin, D., Ainsworth, S., Cho, E., Turk, D. and Winn, L. (1999). Are greek letter social organizations a factor in undergraduates' perceptions of international instructors? *International Journal of Intercultural Relations*, 23(1), 1–12.

Soto, D. (2011). Personal Communication. Telephone interview.

Williams, C. (2010). Personal Communication. Personal interview.

Fear as a key element in deceptive and threatening narratives

Miriam Jiménez Bernal [*]

Abstract. *Deception detection is one of the trending topics in Forensic Linguistics nowadays. Some researchers have already exposed the difficulties of establishing a clear relationship between veracity or deception and linguistic elements (Adams, 2002), while others have attempted to describe to what extent linguists can inform the law enforcement officers on these matters (Eggington, 2008).*

After having collected several Nigerian scam emails and real threat and/or extortion letters, and having analysed the different narratives they contain, we found that fear emerged as a common element, even though its purposes and linguistic materializations are very different. In order to satisfy their ultimate goal – that is, getting money in a fraudulent way –, senders try to persuade the recipients that paying is the best choice, either in a subtle way – appealing to their greed or altruism – or in a violent way – terrorizing the recipient.

The aim of this article is to present briefly how "fear" contributes to the construction of deception and threat through the abovementioned narratives. We will take into account not only the linguistic elements used to cause real fears in their recipients, but also the fake and real fears their senders may have, adding examples to our findings. We will pay specific attention to the influence of gender stereotypes in certain narratives where the senders' gendered identity was – or could have been – fake.

Keywords: Deception, threat letters, extortion letters, Nigerian scam emails, fear.

Introduction

We can no longer doubt that new technologies bring new problems and new opportunities. Maybe we will never receive a threat or extortion letter, especially if we are not wealthy enough to bring the attention of criminals to us, but, if we have an e-mail address, we will probably receive a spam email.

[*]Universidad Autónoma de Madrid

Spam emails can be used to sell pills or luxury watches, among other things, but we may also be the recipients of messages offering us a wonderful job (or love) opportunity or the chance to be rescuers of an exotic and rich heir, a lady in trouble (Nigerian scam emails). Senders will try to convince us to send back some money or personal data using their persuasion and by means of a narrative.

Senders create a fake identity for these texts, and these gendered identities contribute to the veracity of the story in several ways. In this paper, we are going to explore the relationship between the gendered identities and the expressions of fear.

Aim of the study

The main aim of this paper is to show the relation between two different types of narratives, the ones contained in Nigerian scam emails and the ones contained in threat and extortion letters, applying a gender perspective to try to figure out to what extent gender is related to the ways of persuasion we are studying.

We focus on three related objectives in our current research: 1) connecting persuasion and fraud through the use that senders do of narratives; 2) explaining how deception can be a relevant part of persuasion in certain texts (Nigerian scam emails and threat or extortion letters); and 3) finding out the essential role of fear in these narratives, by means of its linguistic expressions.

Narratives, deception and fear: a theoretical frame

Narratives are stories used to justify a belief or behaviour, to support a specific action on the basis of a previous experience that cannot be contested. They have been studied lately, especially from the Critical Discourse Analysis perspective, and some researchers such as Ochs (2000) point out the fact that we relate them with literary forms.

In this case, narratives are aimed at 1) justifying a personal contact with an unknown person that might be perceived by the recipient as an intromission, and 2) creating an identity and telling an identity-coherent story that will finally contribute to the main objective of the letters or emails, that is, getting money in a fraudulent way. The relevance of these narratives is obvious, since they are meant to evoke the specific feelings and ideas that will lead the recipient to the issuance of money or personal data.

Legitimization is the process that allows these narratives, or any other discourse, to have enough authority to be considered believable. From the three interdependent strategies described by Martín Rojo and van Dijk (1997), the one related to the construction of a narrative that will build a specific interpretation framework is the most useful one for us.

Regarding fraud, it is defined as criminal deception, that is, as involving criminal deception for personal gain (Eggington, 2008). In order to define deception, we will follow the one suggested by The Oxford English Dictionary: deception is the act of deceiving or cheating, and deceiving means to cause to believe what is false, to mislead as to a matter of fact, lead into error, impose upon, delude and so on. From a linguistic perspective, as Eggington (2008) states, it involves the manipulation of language to achieve a desired end, where that end misrepresents the facts.

In the Nigerian scam emails and the threat and extortion letters we have analysed, senders try to persuade the recipients through the use of the abovementioned narratives, either in a subtle way, by telling a story, or in a violent way, threatening the recipient.

In both cases, we observe how identity based and message based deception (Hancock, 2007) limits are not as well defined as we may have thought, complementing each other to make the narrative believable.

Persuasion usually involves some kind of manipulation, and manipulation in these narratives is related to Grice's cooperative principle (1975). Senders are giving the recipient some information, and given the veracity maxim ("do not say what you believe to be false"), the recipient will tend to think that the story they are telling is true, and that there is a huge reward for their courage (either if it is related to greed and compassion or to personal safety). As Eggington (2008: 256) states, "their foundation is the potential victim's assumption of truth". In other words, the speech act of persuasion has three elements: the argument must be deductively valid, there must be a commitment on the respondent's side to the premises of the argument, and there is "a special proposition that is designated as the conclusion of the argument" (Walton, 2007: 54).

It has been said that deception is obvious when presented facts do not match the "objective, verifiable reality" (Eggington, 2008: 261), and that is true, but we will have to develop a further analysis to detect deception at a deeper level, since some of these narratives include links to pretended reliable sources, such as the BBC.

In order to study veracity in language, Olsson (2008) establishes different categories in narratives: time, place, sequence, descriptions and superfluity, and tense or aspect. These categories are very relevant for our study, but we will especially focus on linguistic elements related to tense, sequence and descriptions. Coherence among these elements and the identity-related features shown in the narratives, as we will observe, is essential to their veracity.

Coherence is relevant, then, in order to make these narratives as believable as possible, and the gendered identity of the sender is an essential part of it. The story can have no flaws, so the recipient finds it plausible, and the main character in this story has to have the right attributions and characteristics, as we will see in our analysis.

Once we have established the relation between narratives, persuasion and deception, we need to focus on fear and the reason why it is present not only in threat or extortion letters, but also in Nigerian scam emails.

Fear is an emotion, a strong feeling that evokes physiological responses. Parrot (2001) defines fear as two different phenomena: the first one is fear caused by horror provoked by a situation or scene, and it is related to physical answers as alarm, shock or fright. The second one is fear derived from a nervousness process, and it is related to anxiety, uneasiness and worry. We will consider it here as a continuum, from a light fear to terror, bearing in mind the idea of fear as a construction, most of the times to protect against other fears (Glassner, 1999).

Thus, fear can be studied as a phenomenon that starts with the individual interpreting a situation or event as potentially dangerous or threatening, that is, as an anticipation of some kind of damage, either in a physical or psychological aspect of the self. Vulnerability, or self perceived vulnerability, seems to be related to the perception of threats, and, as Parrot (2001: 43) states, "the fearful person describes himself or herself as relatively weak or low in potency" – a fact that is well proven in the studied texts supposedly written by females.

As we have seen, fear is not always motivated by a real cause. Sometimes, fear is just a mere response to our ideas, our mood or a situation that creates an atmosphere that shakes us. We always experience some kind of anxiety when we face a financial transaction on the Internet, not to mention receiving a message from an unknown contact that tries to get money or personal data from us. Thus, creating an atmosphere where trust is promoted is very important.

The narratives contained in threat or extortion letters and Nigerian scam emails try to evoke feelings such as greed or compassion, the altruistic side of people (Whitty and Joinson, 2009), appealing to the most unconscious part of the self of the recipient. But there is also something important we must take into account: the fact that the interpretation frameworks they create are so personal, and so related to fairytales and novels, can make recipients feel ashamed.

We must not forget the relevance of the face: the recipients that pay a huge amount of money because their lives are threatened maintain their faces safe, because nobody would blame them for succumbing when such an important threat was pending over them. Nonetheless, when the recipients that pay a huge amount just because of their greed or for being naïve, they cannot preserve their image, and the shame of being called naïve or greedy can prevent them from reporting their experience to the Police. And shame can be considered another type of manifestation of fear.

Threats use fear to manipulate recipient's thoughts and feelings. Thus, fear was expected in threat and extortion letters, and when we started analyzing the texts using Tropes[1], we found violent expressions and terms related to fear. What we did not expect was to find out that expressions of fear are very common also in Nigerian scam emails, and this tool, Tropes, showed us this relevant finding.

Now, if we want to include the gender perspective in this point, we must talk about style: the variable element in human behaviour (McMenamin, 2002: 125). Style reflects the requirements of the genre (and certainly those of the text's purpose) as well as the unconscious linguistic choices the writer has acquired during his/her lifetime (McMenamin, 2010: 488).

Style then is influenced by several factors, such as the text genre or the sociolinguistic background of the writer. In this case, we do not know the actual sociolinguistic factors that influence the senders, but they create a character with certain identity characteristics: a specific gender, a specific age, a specific ethnicity. All these factors have to be coherent in order to make the narrative believable, as we will see in our results.

Data

Our analysis is based on two corpora. The first one consists of Nigerian scam emails, that is, those mails we found everyday in our electronic inbox, from a sender that is unknown to us, with subjects such as *Business* or *Dear Friend*. This corpus contains 30 emails, extracted from a bigger corpus of around 300 emails received between 2010 and 2012. We decided to include 15 emails with a pretended male sender and 15 emails with a pretended female sender, in order to have the same number of emails from each gendered identity. The languages used are Spanish, English and French.

We also found two emails that seemed not to follow the general trend. One of them has some business content and it is signed by a woman, but together with her son. The second one contains a narrative that reminds us very much of the female senders narra-

tives, but in this case, it is signed by a man. Even if we find them interesting for further research, especially regarding their implications for gender issues, we will not use examples from these texts, since they are not representative.

The second corpus consists of 12 threat and extortion letters, from real cases investigated and already solved by the Police. All of them were written by males, even though there is one case in which the author pretends to be "two worried mothers".

Extortion and threat letters are documents usually sent by an unknown sender and received by companies and some individuals (generally, businessmen). Through these texts, the sender tries to force the recipient to give them some money, usually a huge amount, frightening them with some threats that can involve personal, familiar or economic damages.

We can distinguish between the letters sent to companies and the letters sent to individuals. On the one hand, the ones sent to companies usually mean an economic damage, and are supposed to be sent by groups, although none of our corpus cases pretended to have been sent by a terrorist group.

On the other hand, the ones sent to individuals are either supposedly written by terrorist groups or groups (of at least two people) that know the recipient and have a personal interest in damaging him. The purpose, anyway, is to frighten the recipient so he (since the addressees are always males) gives them a huge amount of money.

In both cases, addressees are to follow some instructions in order to deliver the money in a safe way.

Analysis

Methodology

In order to make an analysis as complete and accurate as possible, we have used an interdisciplinary methodology. We used (Critical) Discourse Analysis, as well as some concepts from Women's studies and feminist theories, such as that of gender attributions, considered as those characteristics traditionally associated with masculine or feminine behaviours.

Firstly, we divided the letters and emails into several parts, in order to be able to identify the ones in which fear appeared either in an explicit or implicit way. Our first division, based on the CARs model proved not to be very useful, so we decided to follow a more complex model based on episodes (Picornell, 2010) rather than on pre-defined parts.

The new division gave us the chance to distinguish between the different episodes in the narrative, allowing us to compare the right pieces of information in both the threat and extortion letters and the Nigerian scam emails.

Secondly, we tried to identify the expressions of fear in terms of topics and, thirdly, we focused on morphosyntactic features related to 1) expressions of fear or worry in the sender, or 2) threats to the recipient.

As we mentioned before, our analysis takes into account the gender perspective, not only by separating male and female identities, but also by applying concepts such as that of gender attributions[2]. Our study is mainly qualitative, although we hope to be able to produce some quantitative results in a not too distant future.

Results and discussion

First of all, we would like to explain that we will talk about female or male senders, depending on the gendered identity adopted by the sender, although these adjectives will sometimes correspond to an unknown sender (probably a male). Since all the threat and extortion letters were written by males and we cannot know who the real author of these Nigerian scam emails is, although we presume they are also males, we consider Judith Butler's notion of gender performativity[3] (1999) a rather interesting concept to be used in further analysis.

In this section, we will present the results in two areas: topics and morphosyntactic features. We will also include some examples to support our findings.

Common topics

Following Parrot's wide definition of fear, the topics we found more frequently that seem to be related to this feeling are the ones detailed in Figure 1. We have only included topics that we considered to be relevant, with approximately 20 appearances in Nigerian scam emails and 8 in threat or extortion letters, either in an explicit or implicit way, even though the examples in Figure 1 are just explicit ones.

Figure 1. Most common topics related to the fear continuum.

Nigerian scam emails		Threat and extortion letters	
Death	>30 hits	Death/physical damage	>8 hits
Cancer	>20	Criminals/terrorist	
Urgent/urgently	>20	groups	>8

These results support the idea of fear, understood as a continuum, being present in these texts for persuading the recipients, even though persuasion is developed in two very different ways: 1) appealing to feelings such as greed or compassion,

(1)
```
I am constrained to contact you because of the abuse I am
receiving from my step mother
```

(2)
```
Recently, my Doctor told me that i would not last for the next 5
months due to cancer & stroke illness
```

and 2) appealing to the survival instinct of the recipient.

(3)
```
además del daño físico (muerte por cualquier causa)
besides physical damage (death by any cause)
```

(4)

```
amenazado de muerte
death threatened
```

In threat and extortion letters, terms related to fear and threat are not always as obvious and explicit as in Nigerian scam emails stories. Most of the times, we must infer them from what is said – or not said – in the conditional sentences and exhortations. Nonetheless, the use of insults ("puta", "cabrón") and references to terrorism and criminal groups are frequent.

Vocabulary and morphosyntactic features

Let us focus now on morphosyntactic features related to either expressions of fear or expressions used to provoke fear in the recipient. Expressions of fear or worry on the sender's side are more common in texts with a female sender. They use adjectives and singular pronouns. Female senders pretend to be worried or frightened.

(5)

```
Recently, my Doctor told me that i would not last for the next
5 months due to cancer & stroke illness.  [...] Having known
my condition i decided to seek for a competent and reliable
God fearing person or church to entrust this fund to utilize
this fund the way i narrated above as my late husband desired.
[...] I don't want a situation where this money will be used in
an ungodly way.  This is why I am taking this decision.  I am
not afraid of death hence i know where i am going.  I don't
need any telephone communication in this regard because of my
health hence the presence of my husband's relatives are around
me always.  i don't want them to know about this development.
```

In this extract, we can see that death is mentioned (one of the most common fears, even though it is seeking for compassion, since the sender declares that she is not afraid, due to her Christian beliefs). The fear here seems to be the one of the relatives finding out that they are not getting the money when she is dead, since she believes them to be unworthy.

(6)

```
My father was killed by government of Charles Taylor, he accuse
my father of coup attempt.  [...] I am constrained to contact
you because of the abuse I am receiving from my step mother.
She planned to take away all my late father's treasury and
properties from me since the unexpected death of my beloved
Father.
```

Here, we can observe how the fear atmosphere is created: there is a previous death, there is wicked stepmother abusing our heroine, and, in the ending lines, urgency is mentioned several times.

(7)

```
Somos dos madres quienes escribimos esta carta, muy preocupadas
por dos hijos nuestros – uno de ellos amenazado de muerte –
[...]Dos madres amenazadas.  Dos hijos en peligro de muerte.
¡Toda una familia unida contra el terror!
```

```
We are two mothers writing this letter, very worried about two
sons of ours – one of them death threatened – [...]Two threatened
mothers.  Two sons at risk of death.  A whole family united
against terror!
```

In these lines, we find words like "worry", "death threat" or "terror", trying to support the story of two worried mothers that want to cooperate with the company, instead of willing to commit some fraud or extortion.

Sometimes, e-mail senders try to avoid the possible anxiety caused by their narratives through the use of formulaic expressions such as

(8)

```
God will grant you the willingness and interest to diggest this
humble narrations though it might be so surprise and strange to
believe my story but i knew by the reason of the almighty you
will humbly understood and accept to proceed with my proposal
though we have not met or seen eachother before.
```

This use is equally valid for male and for female senders, as we have found it in both kinds of emails. However, these expressions are not used in threatening letters, because their purpose is, precisely, to cause fear in the recipient.

Expressions to provoke fear in the recipient are more common in male texts. Threats are usually expressed in the conditional form:

(9)

```
if you didn't come up with the certificate we shall confiscate
the funds into World Bank account then charge you for money
laundry.
```

(10)

```
De no realizar el deposito o no ponerse en contacto con nosotros
en las proximas 48 horas , su X , usted y su familia.  Se
convertiran en un objetivo potencial del grupo.
```

```
If you do not make the deposit or contact us during the next
48 hours, your X, you and your family will become a potential
target for the group.
```

Also, we would like to remark that instructions appear in all these texts, so we might think it is related to the commission of a crime or offence, and not to a particular feature of these genres.

If we focus on male senders, we observe that they usually try to frighten the recipient by using a group as a shield, and the group has a certain authority, because groups seem to be much more threatening than individuals. Plurals are used throughout the text below, as well as some imperatives and exhortations, together with a threat of charging the recipient for "money laundry", in the conditional form. So, if being smart enough to get some quick and easy money is not enough, maybe a threat of being charged works.

(11)
```
We, office of the international police association (IPA)
[...]Now, the diplomat is under detention in the office of (IPA)
security, and we cannot release her until we carry out our
proper investigation on how this huge amount of money managed to
be yours before we will release her with the box.  So, in this
regards you are to reassure and prove to us that the money you
are about to receive is legal by sending us the Award Ownership
Certificate showing that the money is not illegal.  [...]You are
advised to forward immediately the Award Ownership Certificate
[...]Furthermore, we are giving you only but 5 working business
days to forward the requested Award Ownership Certificate.
Please note that we shall get back to you after the 3 working
business days, that if you didn't come up with the certificate
we shall confiscate the funds into World Bank account then
charge you for money laundry[...].
```

The implicit fears found in the narratives, either in threat and extortion letters or Nigerian scam emails, have been inferred and extracted from instructions and excuses made by the senders. One of the main fears, as we already mentioned, is the one of not being taken seriously:

(12)
```
(solo este aviso, no se lo tomen a broma o sino muerte [...]los
ultimos a los que le enviamos una carta parecida y nos tomaron
de broma ¡que pena!, ya no lo pueden contar (sic)
```

```
Just this advice, don't take it as a joke or death [...]The last
ones we sent a similar letter to and took it as a joke what a
pity!, cannot talk anymore.
```

The most relevant fear, though, is the one of being caught:

(13)

```
sirvanse de esta carta para envolver la plata, no una copia sino
esta misma hoja (sic)
```

```
Use this letter to wrap the money, not a copy but this very
letter.
```

Apparently, in extortion and threat letters, the sender tries to fake a multiple identity, since groups seem to be more frightening than individuals, either when they are males or when they are females. As we explained before, fear was expected to be present in threat, since threat is used to create fear in the recipient.

In Nigerian scam emails, senders are almost always individuals, but there is a difference between male and female senders. While male senders seem to be supported by a group (their company or organization, or the status of the group of workers they belong to, as in the case of lawyers), female senders are just individuals. Presumably, individuals are more fragile and vulnerable than groups and we could assume that this loneliness is intentionally remarked by senders to provoke compassion.

Gender has a crucial role in all those narratives, since the characters (most of them made up for the occasion) have a gender (or gendered) identity: they are all male or female and have attributes that are traditionally associated with males or females.

The female characters in the narratives are either vulnerable young ladies, victims of terrible relatives (reminding us of the fairytales heroines), or mothers, including the old women that are not mothers but would have wanted to. When they are young ladies, they are usually orphans and the implicit promise in their narratives is not only about getting a lot of money (this is most of the time pretty explicit), but also about the possibility of having a relationship with an exotic and wealthy young lady. It is easy to seem frightened when one is alone, and frightening when one is in a group.

When they are "almost" mothers, they usually impersonate the virtuous woman, respectful of God's law, suffering from a terrible illness but strong enough to seek for a good destiny to her inheritance (because they are always widows). And if we talk about actual mothers, they also impersonate moral purity and worry about their descendants, as shown in the corresponding extortion letter.

Fear (of God, of losing freedom…) is an essential part of these characters: it positions them in a situation in which the assistance and help from the recipient is absolutely needed, most of the times in the shape of money. The only example we found about a woman writing with a business purpose, was the one of a lady supported by her title …and her son.

The male characters, on the contrary, are supported by their own expertise and knowledge. They are experts (lawyers, accountants, managers or policemen) and they are courageous and clever. They offer money through not very legal transactions and businesses, and give the recipient an opportunity to be as courageous and clever as them.

Conclusions

To cause real fears, male senders use threats in both kinds of texts. Fear raises from a situation such as a personal threat ("te vamos a matar" – we will kill you – ; "we will charge you for money laundry") or an economic threat, for companies ("vamos a publicar información sobre vuestros clientes" – we will make public some information about your clients). In order to create the perfect atmosphere for frightening the recipient, senders pretend to be a group (or at least a couple), since a group seems to be more effective to intimidate the reader, especially if it is a terrorist group or a group related with the law.

Fears that are evoked in recipients are the fear of losing one's freedom, the fear of losing the most beloved ones, the fear of losing the status of the company and the trust of the clients, the fear of losing one's own life.

To appeal to altruism and compassion, female senders fake fears and express them explicitly in both types of texts. Fears that are expressed by senders here are the fear of losing one's own life and the fear of leaving one's money in inappropriate hands, mainly. Both of them are easily understood – especially the first one, since fear of death is one of the most common fears.

Maybe the most relevant finding is that, unknowingly and unwillingly, senders also express their real fears in an implicit way in both texts, by using the same kind of linguistic strategies and features.

Fear helps senders in the creation of fake identities and fake narratives. When these narratives and identities are consistent and coherent, they influence the interpretation framework of the recipients, sometimes causing the desired response.

Future research

Given the fact that compiling spam emails is easy and quick, and since we believe that the study of their narratives and the comparison with other narratives from texts related to the forensic area is very interesting, we will continue working with these texts. Our intention is, firstly, to include other feelings and topics, such as death, illness, or business, and secondly, to support our qualitative analysis with a quantitative analysis.

Notes

[1]Tropes is a tool designed for Semantic Classification, among other uses. It is free and available at http://www.semantic-knowledge.com/download.htm

[2]The notion of gender attributions involves, among other issues, those characteristics and activities that are traditionally and commonly associated to either women or men according to their sex, i.e., tenderness applied to women or strength applied to men, for instance (Beltrán *et al.*, 2001).

[3]According to Judith Butler (1999), gender, as a social construction, cannot be considered as an intrinsic characteristic of individuals but rather as a performance and, so, it can be subverted.

References

Adams, S. H. (2002). *Communication under stress: indicators of veracity and deception in written narratives.* PhD thesis, Virginia State University, Virginia.

E. Beltrán, V. Maquieira, S. Álvarez and C. Sánchez, Eds. (2001). *Feminismos. Debates teóricos contemporáneos.* Madrid: Alianza Editorial.

Butler, J. (1999). *Gender trouble: feminism and the subversion of identity.* London: Routledge.

Eggington, W. (2008). Deception and fraud. In M. T. Turell and J. Gibbons, Eds., *Dimensions of Forensic Linguistics*. Amsterdam: John Benjamins.

Glassner, B. (1999). The construction of fear. *Qualitative Sociology*, 22(4), 301–309.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. Morgan, Eds., *Studies in Syntax and Semantics III: Speech Acts*, 183–198. New York: Academic Press.

Hancock, J. T. (2007). Digital deception. Why, when and how people lie online. In A. Joinson, K. McKenna, T. Postmes and U. Reips, Eds., *The Oxford handbook of Internet psychology*. Oxford: Oxford University Press.

Martín Rojo, L. and van Dijk, T. (1997). 'There was a problem and it was solved!': Legitimating the expulsion of 'illegal' migrants in Spanish parliamentary discourse. *Discourse Society*, 8(523).

McMenamin, G. (2002). *Forensic linguistics: Advances in Forensic Stylistics*. London: CRC Press.

McMenamin, G. (2010). Forensic stylistics: Theory and practice of forensic stylistics. In M. Coulthard and A. Johnson, Eds., *The Routledge Handbook of Forensic Linguistics*. London: Routledge.

Ochs, E. (2000). Narrativa. In T. van Dijk, Ed., *El discurso como estructura y proceso*. Barcelona: Gedisa.

Olsson, J. (2008). *Forensic linguistics*. New York: Continuum International Publishing Group.

Parrot, W. (2001). *Emotions in social psychology: Essential readings*. Philadelphia: Psychology Press.

Picornell, I. . (2010). The rake's progress: linguistic strategies for deception. In S. Tomblin, N. McLeod, R. Sousa-Silva and M. Coulthard, Eds., *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*. Birmingham: Centre of Forensic Linguistics, Aston University.

Walton, D. (2007). *Media argumentation: dialectic, persuasion, and rhetoric*. Cambridge: Cambridge University Press.

Whitty, M. and Joinson, A. (2009). *Truth, lies and trust on the Internet*. New York: Routledge.

Investigating legal language peculiarities across different types of Italian legal texts:
an NLP-based approach

Giulia Venturi *

Abstract. *In this paper, the author carried out the linguistic profiling of a corpus of different types of Italian legal texts exemplifying different sub-varieties of Italian legal language by relying on a wide range of different linguistic features (lexical, morpho-syntactic and syntactic) automatically extracted from the output of a multi-level automatic linguistic analysis of texts. The devised comparative approach allowed investigating the linguistic variation* i) *between the considered corpus of legal texts and a corpus of newspaper articles representative of Italian ordinary language and* ii) *among the considered types of legal texts (legislative acts, administrative acts, the Italian Constitution and legal cases). Achieved results can provide the starting point to identify areas of lexical, morpho-syntactic and/or syntactic complexity within a legal text in order to assess its readability as well to perform a number of different computational forensic linguistics tasks.*

Keywords: Linguistic profiling, Natural Language Processing, legal language analysis, legal genres, readability assessment, syntactic complexity.

Introduction

Over the last few years, the use of Natural Language Processing (NLP) tools and techniques has spread within computational forensic linguistics studies. In spite of the fact that they address different purposes, such as authorship attribution (Sousa-Silva *et al.*, 2010) or automatic deception detection (Fornaciari and Poesio, 2013), these studies share a common approach: they succeed in their specific goal by exploiting the distribution of a number of different linguistic characteristics automatically extracted from the linguistically analysed text. More generally, this is the basis of the so-called "linguistic profiling" within which "the occurrences in a text of a large number of linguistic features, either individual items or combinations of items, are counted" (van Halteren, 2004: 202).

---
*Instituto di Linguistica Computazionale "Antonio Zampolli", CNR

In this paper, the *linguistic profiling* of a corpus of different types of Italian legal texts exemplifying different sub-varieties of Italian legal language is carried out by relying on NLP-enabled linguistic features automatically extracted from text. It stems from studies on register analysis, such as those carried out by Biber and colleagues (1993; 1998; 2009), and in particular from studies focussed on the linguistic characteristics specific to different types of legal texts, i.e. legal texts pursuing different communicative purposes (Bathia, 1993). Unlike previous studies, feature extraction was carried out on the output of a multi-level automatic linguistic analysis: a wide typology of selected features was looked for within each level of the automatic linguistic analysis (lexical, morphosyntactic and syntactic). The eventual goal is to demonstrate how automatic linguistic analysis can provide a useful starting point for the linguistic profiling of legal texts.

The author followed a comparative approach with two specific goals. The first goal consists in finding significative linguistic variation between the corpus of legal texts under consideration and a corpus of newspaper articles representative of Italian ordinary language. This is quite crucial, since legal language *differs* from ordinary language, but it is not dramatically *independent* from every day speech (Mortara Garavelli, 2001; Rovere, 2005). The second goal is the investigation of the linguistic characteristics that make various types of legal texts different and distinguishable. The interest in pursuing this goal relies on the widely acknowledged fact that the "term 'language of the law' encompasses several usefully distinguishable genres" and that "these genre distinctions are also reflected in the lexico-grammatical, semantico-pragmatic, and discoursal resources that are typically and conventionally employed to achieve successful communication in various legal settings" (Bathia, 1987: 227). For this purpose, we compared the different types of texts included in the corpus of legal texts described here.

Methodology

The approach to linguistic profiling devised here stems from Dell'Orletta *et al.* (2013) and includes three main ingredients:

- a collection of legal texts exemplifying different sub-varieties of Italian legal language and of texts representative of Italian general language used as reference corpus,
- a collection of NLP tools performing the automatic linguistic analysis of texts,
- a method of linguistic profiling which allows the comparison of different corpora according to the distribution of a set of linguistic features.

Corpora

For the specific concerns of this study, we built a corpus of legal texts exemplifying different sub-varieties of Italian legal language, i.e. legislative acts, administrative acts, the Italian Constitution and legal cases. Following the legal text classification suggested by Bathia (1987), they belong to two main classes of documents used in two different legal settings. While the legislative and administrative acts, as well as the Italian Constitution, are types of documents used in a "legislative setting", the legal cases are typically used in a "juridical setting". According to the classification of the Italian legal texts proposed by Mortara Garavelli (2001: 19–34), the Italian Constitution belongs to the class of "normative" texts (used in a "legislative setting"). However, it was analysed here separately from the legislative texts in order to highlight its linguistic specificity with respect to other normative texts.

Table 1. The corpora of legal texts.

Legal text sub-genre	Enacting/resolving authority	N. word tokens
Legislative acts	Italian State	744,064
	Piedmont Region	112,474
	European Union	453,328
Sub-total		**1,309,866**
Italian Constitution		**10,487**
Administrative acts	Italian State	107,240
	Piedmont Region	182,213
	European Union	17,951
Sub-total		**307,404**
Legal cases	Administrative Court	87,653
	Court of Civil Cassation	184,905
	European Convention on Human Rights Court	543,582
	Constitutional Court	53,377
	Ordinary tribunals	53,775
Sub-total		**923,292**
TOTAL		**2,551,049**

Table 1 shows the internal partition of this Italian legal text corpus. As can be seen, both legislative and administrative corpora include acts enacted by the Italian State, the Piedmont region and the European Union. The two corpora are respectively made up of legislative acts, such as national and regional laws, European directives, legislative decrees, and administrative acts, such as ministerial circulars and decisions. The collection of legal cases contains texts resolved by various Italian administrative courts, the Italian Constitutional court, the Italian Court of Civil Cassation, the European Convention on Human Rights Court, and by various Italian ordinary tribunals. They all concern the principle of state *liability*, i.e. the general principle established by the European Court of Justice according to which Member States should pay compensation to individuals who suffered a loss by reason of the State failing to comply with the European law (Lazari, 2005). Finally, the Italian Constitution was analysed in its 1947 original version.

According to the comparative approach devised here, we chose a corpus of newspaper articles taken as representative of general Italian language as a baseline for comparison. The choice of the journalistic prose as reference genre is inspired by the work of Rovere (2005) who investigated the morpho-syntactic and syntactic differences of a corpus of different types of Italian legal texts against a corpus of Italian newspapers, focussing on the relationship between the semantic and syntactic valencies of some verbs.

However, unlike the case of Rovere, two newswire corpora were considered here: a collection of articles taken from "La Repubblica" daily newspaper and from "Due Parole"[1], a newspaper written by Italian linguistic experts in text simplification using a controlled language (Piemontese, 1996). According to their linguistic peculiarities, as empirically demonstrated by Dell'Orletta and Montemagni (2012) and Dell'Orletta *et al.* (2011b), the two corpora can be seen as two opposite poles of the same textual genre: "Due Parole" represents a newspaper explicitly written using plain language while "La Repubblica" articles represent the opposite extreme being written using everyday language. They were both taken as reference corpora here since they allow extensive investigation of the linguistic peculiarities of legal texts.

Natural Language Processing tools

All the corpora were automatically analysed by a collection of statistical NLP tools jointly developed by the Institute of Computational Linguistics "Antonio Zampolli" in Pisa (ILC-CNR) and the University of Pisa. They were morpho-syntactically tagged by the Part-Of-Speech tagger described in Dell'Orletta (2009) and syntactically annotated by the DeSR parser, the dependency parser described in Attardi (2006).

The tools are able to make evident the implicit linguistic information contained in texts by *annotating* them at increasingly complex levels of analysis. Namely, they split the whole text into sentences, segment each sentence into orthographic units (tokens), assign all possible morphological analyses to each token, assign the appropriate morpho-syntactic interpretation in the specific context and identify existing syntactic dependency relations between tokens (e.g. subject, object, etc.). For example, the following sentence is annotated as seen in Table 2[2]:

(1) Gli Stati membri provvedono affinché il gestore sia obbligato a trasmettere all'autorità competente una notifica entro i seguenti termini. ('Member States shall require the operator to send the competent authority a notification within the following time-limits'.)

In Table 2, it can be noted that each word form (in the column headed FORM), univocally marked by a numerical identifier (column ID), is associated with its corresponding lemma (column LEMMA), its coarse- (column CPOSTAG) and fine-grained (column POSTAG) part-of-speech and its morphological treats (column FEATS). Moreover, the annotation makes explicit the head of the dependency syntactic relation in which each word is involved (column HEAD) and the type of dependency relation (column DEPREL). For example, Table 2 shows that the word *notifica* ('notification') is the object (obj) of the verb *trasmettere* ('send').

Table 2. Example of an annotated sentence in CoNLL format.

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL
1	Gli	Il	R	RD	num=p|gen=m	2	det
2	Stati	Stati	S	SP	-	4	subj
3	membri	Membro	S	S	num=p|gen=m	2	mod
4	provvedono	provvedere	V	V	num=p|per=3|mod=i|ten=p	0	ROOT
5	affinché	Affinché	C	CS	-	4	mod
6	il	Il	R	RD	num=s|gen=m	7	det
7	gestore	Gestore	S	S	num=s|gen=m	9	subj_pass
8	sia	Essere	V	VA	num=s|per=3|mod=c|ten=p	9	aux
9	obbligato	Obbligare	V	V	num=s|mod=p|gen=m	5	sub
10	a	A	E	E	-	9	arg
11	trasmettere	trasmettere	V	V	mod=f	10	prep
12	all'	A	E	EA	num=s|gen=n	11	comp_ind
13	autorità	Autorità	S	S	num=n|gen=f	12	prep
14	competente	competente	A	A	num=s|gen=n	13	mod
15	una	Una	R	RI	num=s|gen=f	16	det
16	notifica	Notifica	S	S	num=s|gen=f	11	obj
17	entro	Entro	E	E	-	11	comp_temp
18	i	Il	R	RD	num=p|gen=m	20	det
19	seguenti	Seguente	A	A	num=p|gen=n	20	mod
20	termini	Termine	S	S	num=p|gen=m	17	prep
21	.	.	F	FS	-	4	punc

It should be mentioned here that even if the NLP tools exploited in this study represent the state of the art for the Italian language (see the results of the Evalita 2009 evaluation campaign[3], however their performances are affected by the particular type of texts at hand. Since (Gildea, 2001), it is widely acknowledged that state-of-the-art linguistic

annotation tools suffer from a dramatic drop of accuracy when tested on domains outside of the data from which they were trained or developed. As recently testified by the "Domain Adaptation Track" (Dell'Orletta *et al.*, 2012a) and by "The SPLeT-2012 Shared Task on Dependency Parsing of Legal Text" (Dell'Orletta *et al.*, 2012b), the legal domain does not represent an exception. In order to cope with this problem, for the specific concerns of this study, the statistical NLP tools were specialised by combining two training sets: the ISST-TANL treebank consisting of newspaper articles (Dell'Orletta *et al.*, 2012b) taken as representative of general language usage, and the TEMIS corpus (Venturi, 2012), a syntactically annotated corpus of Italian legislative and administrative texts. This allowed maintaining the state-of-the-art performance of the NLP tools exploited here.

Comparative approach and linguistic features

The comparative approach followed in this study stems from the literature on register variation and is mostly inspired by the work of Biber who claims that "we need a baseline for comparison to know whether the use of a linguistic feature in a register is rare or common" (Biber *et al.*, 1998). For the specific concerns of this study, the corpora illustrated in Section have been compared at two different levels. Firstly, the distribution of some linguistic features was comparatively observed within the whole collection of legal texts and the newswire corpora. This perspective of analysis has allowed the highlighting of how and to what extent legal language differs from ordinary language. Secondly, the different types of legal texts were compared at different levels of specificity in order to investigate the linguistic variation between documents used in a "legislative" and "juridical setting" and between documents enacted or resolved by different authorities. This second level of analysis allowed us to show the *multiform and complex* nature specific to the legal language (Cortelazzo, 1997) highlighting the peculiarities of different legal language sub-varieties.

Starting from the output of the automatic linguistic analysis, all the corpora were searched for with respect to four classes of linguistic features: raw text, lexical, morpho-syntactic and syntactic. As illustrated in 3, this four-fold partition closely follows the different levels of linguistic annotation and it was prompted by Biber's idea that "linguistic features from all levels function together as underlying dimensions of variation and [. . .] there are systematic and important linguistic differences among registers with respect to these dimensions" (Biber, 1993: 220–221).

Two different criteria have guided the choice of these linguistic features. Firstly, some of them are contained in the Italian "Guide to drafting administrative acts" devoted to suggesting how to draft acts in a plain language[4]. Gathering the suggestions put by the "Directive on the simplification of the language of administrative acts" of the Ministry for Civil Service Reform and by the "Rules and suggestion to drafting legislative acts" adopted by the Italian local authorities, the "Guide" is today the most up-to-date collection describing the lexical, morpho-syntactic and syntactic characteristics that a legal document is expected to conform to in order to be written in a *plain, simple and comprehensible* language.

Secondly, this work would like to follow the methodology devised by Dell'Orletta and Montemagni (2012) who demonstrated the high discriminative power of the set of linguistic features considered here to monitor diastratic, diamesic and diaphasic varieties of Italian language. Their results were also confirmed when these features were adopted to linguistically profile Italian educational materials, such as textbooks for primary and

Table 3. The considered linguistic features.

Type of feature	Level of linguistic annotation	Feature
Raw text	Sentence splitting	Sentence length calculated as the average number of words
Lexical	Lemmatization and morpho-syntactic annotation	Percentage of all unique words (types) included in the "Basic Italian Vocabulary" (De Mauro, 2000) calculated on a per-lemma basis
		Distribution of the occurring Basic Italian Vocabulary words into the usage classification classes of *fundamental vocabulary* (very frequent words), *high usage vocabulary* (frequent words) and *high availability vocabulary* (relatively lower frequency words referring to everyday objects or actions and thus well known to speakers)
Morpho-syntactic	Morpho-syntactic annotation	Part-Of-Speech distribution
		Noun/verb ratio
Syntactic	Dependency annotation	Features based on the structure of a syntactic tree: - average depth of the whole tree, calculated in terms of the longest path from the root of the dependency tree to some leaf, - average length of the longest dependency links measured in terms of the words occurring between the syntactic head and the dependent
		Features concerning the use of subordination: - distribution of main vs subordinate clauses, - subordinate/main clauses ratio, - average depth of 'chains' of embedded subordinate clauses and their distribution by depth
		Features concerning nominal modification: - average depth of embedded prepositional complement 'chains' governed by a nominal head and their distribution by depth

high school and written productions of learners of Italian (Dell'Orletta *et al.*, 2011a), and to assess the readability of newspaper texts (Dell'Orletta *et al.*, 2011b) and of different text genres (Dell'Orletta *et al.*, 2012c).

Legal texts vs newspaper articles: linguistic peculiarities

The linguistic characteristics specific to the corpus of legal texts taken as a whole are reported and discussed in what follows. They provide first insights into the peculiarities of legal language compared to ordinary language.

Raw and lexical text features

According to one of the first suggestions contained in the "Guide to drafting administrative acts", a legal text should contain *short* sentences. As Table 4 shows, the comparison of the legal and newswire corpora revealed that on the contrary the whole legal corpus (*Legal* in the Table) contains sentences longer than those occurring both in "La Repubblica" (*Rep*) and in "Due Parole" (*2Par*).

Table 4. Average sentence length in legal and newswire corpora.

	Legal	Rep	2Par
average sentence length	28.91	26.54	19.20

A similar recommendation towards a text written in a plain language holds for the lexical profile that a legal text should exhibit. According to the "Guide", a legal text should

mainly contain words belonging to the "Basic Italian Vocabulary" since they are more frequently used and consequently more comprehensible. As Table 5 reports, the legal corpus contains a rather lower percentage of "Basic Italian Vocabulary" (BIV) lemmas (calculated in terms of types, as reported above in Table 3) with respect to newswire corpora.

Table 5. % of lemmas (types) belonging to the "Basic Italian Vocabulary" in legal and newswire corpora.

	Legal	Rep	2Par
lemmas belonging to BIV	10.94	67.09	74.58
lemmas NOT belonging to BIV	89.06	32.91	25.42

When we further compare the corpora with respect to the percentage distribution of the occurring "Basic Italian Vocabulary" lemmas into the usage classification classes (i.e. *fundamental*, *high usage* and *high availability*), some interesting results can be observed. As Table 6 shows, the legal corpus not only contains a lower percentage of fundamental vocabulary, but the difference between the percentage of this class and the *high usage* and *high availability* vocabulary is lower. In the legal corpus the difference between the distribution of fundamental and high usage vocabulary is of 4.24 percentage points, while in "La Repubblica" corpus the same difference is of 56.51 percentage points and in "Due Parole" it is of 66.07 points. Such a difference shows the particular tendency in the legal corpus towards the use of *high usage* and *high availability* vocabulary.

Table 6. % of lemmas (types) distributed in the three usage classes in legal and newswire corpora.

	Legal	Rep	2Par
fundamental vocabulary	44.29	75.46	79.99
high usage vocabulary	40.05	18.95	13.92
high availability vocabulary	15.66	5.28	5.28

Morpho-syntactic features

At the morpho-syntactic annotation level we observe a different distribution of Parts-of-Speech (PoS) categories occurring in the legal and newswire corpora. The percentage distribution of the considered PoS are reported in Table 7. It can be noted that the legal corpus has a higher percentage of **prepositions**, **numbers**, **nouns** and **adjectives** while it contains a significantly lower percentage of **verbs** and **adverbs**.

Interestingly, these results are in line with the literature on register variation. According to Biber (1993), both a high co-occurrence of **nouns**, **prepositions** and **adjectives** and different noun/verb ratio values represent two significant dimensions of variation among textual genres. Concerning this latter characteristic, Biber (1993) found that highly informative genres such as academic prose are characterised by a higher noun/verb ratio with respect to texts representative of fiction prose or with respect to speech. By relying on the output of a collection of NLP tools, a similar tendency was observed by Montemagni (2013) for the Italian language. She demonstrated that a corpus of fiction texts and of speech transcriptions have a lower noun/verb ratio than to a corpus of newspaper articles.

Table 7. Morpho-syntactic features in legal and newswire corpora.

	Legal	Rep	2Par
adverbs	2.26	4.83	3.52
prepositions	19.92	16.41	15.28
numbers	5.59	2.39	2.73
verbs	9.38	12.89	13.66
nouns	29.95	27.19	29.30
adjectives	7.69	6.40	5.92
noun/verb ratio	3.19	2.11	2.14

The different distribution of nouns, prepositions and adjectives in the legal and newswire corpora suggests that we are dealing with two different linguistic varieties. The higher noun/verb ratio provides further evidence that legal texts are even more informative than newspaper articles. It is a straight consequence of the higher percentage of **nouns** and, above all, of the quite low percentage occurrence of **verbs** occurring in the legal texts.

The significant higher occurrence of **numbers** in the legal texts is a further peculiarity of this genre. It is due to several different reasons, e.g. the ample use of numbers corresponding to textual partition numbering (e.g. article, paragraph), the occurrence of dates, the citations between legal cases or legislative acts also expressed through numerical identifiers, etc.

Syntactic features

As reported above in Table 3, in this study we considered three types of syntactic features concerning: the structure of a syntactic tree, the use of subordination and the nominal modification.

Features based on the structure of a syntactic tree

Two features have been taken into account: the average depth of the syntactic tree and the average length of the longest dependency links.

Consider the two following sample sentences extracted from the legal corpus:

(2) I proprietari, possessori o detentori a qualsiasi titolo dei beni indicati al comma 1, hanno l'obbligo di sottoporre alla Regione i progetti delle opere di qualunque genere che intendano eseguire, al fine di ottenere la preventiva autorizzazione. ('The owners, possessers or holders on whatever basis of the goods mentioned in paragraph 1, have the obligation to submit to the Region the projects of the works of any kinds they plan to carry out, in order to obtain prior authorization.')

(3) Chiunque immette sul mercato i preparati pericolosi di cui al presente decreto, in violazione delle disposizioni in tema d'imballaggio e di etichettatura di cui agli articoli 8, 9 e 10, nonché in violazione delle disposizioni sulla classificazione di cui all'articolo 3, è punito con l'ammenda da euro centoquattro a euro cinquemilacentosessantacinque. ('Anyone who places on the market dangerous preparations provided for in this decree, in violation of the provisions on packaging and labelling referred to in articles 8, 9 and 10, and in violation of the provisions on the classification referred to in article 3, is punished by a fine of one hundred and four euro to euro five thousand one hundred sixty-five.')

In (2), the syntactic tree has a maximum depth=8. As Figure 1 shows, it is calculated as the sequence of eight consecutive dependency links[5], i.e. direct object (*obj*), argument (*arg*), preposition (*prep*), direct object (*obj*), complement (*comp*), preposition (*prep*) and relative modifier (*mod_rel*), which starts from the root of the syntactic tree *hanno* ('have') and ends to the leaf *eseguire* ('carry out').

In line with the literature on measuring dependency distance, the length of the longest dependency link is measured here in terms of "intervening words" (Hudson, 1995: 16) between a dependent and its parent. In (3), the longest dependency link is 47 tokens long[6]: it is the subject relation between the dependent *chiunque* ('anyone') and its governing head corresponding to the syntactic root of the sentence (*punito*, 'punished').

As can be seen in Table 8, legal sentences are characterised by deeper syntactic trees and much longer dependency links.

Table 8. Features based on the structure of syntactic trees in legal and newswire corpora.

	Legal	Rep	2Par
average syntactic tree depth	6.95	6.51	5.29
average length of the longest dependency links	12.88	10.28	7.91

These results suggest that within legal texts there is a more complex syntactic structure with respect to newspaper articles. Not only can a deep syntactic tree be indicative of increased sentence complexity as stated by e.g. Frazier (1985) and Gibson (1998), but the same holds for long dependency links. In a dependency representation scenario, Hudson (1995: 15-16) claims that the "dependency distance", i.e. "the distance between words and their parents, measured in terms of intervening words", "might be relevant to how hard a sentence is to process". Accordingly, the greater the dependency distance, the more complex is the sentence. This is in line with the findings of studies carried out in the cognitive and psycholinguistic field. In particular, it has been ascertained that "there is a finite span of immediate memory and that [...] this span is about seven items in length" (Miller, 1956: 9). It follows that it is perceptually costly to carry on analysing sentences with long dependencies.

Features concerning the use of subordination

Concerning the use of subordination, the "Guide" recommends a limit on the subordinate clauses. This follows from the literature which relates syntactic complexity to the occurrence of embedding structures and in particular to the presence of subordinate clauses. According to Beaman (1984: 45) "syntactically complex authors [...] use longer sentences and more subordinate clauses". Typically, the use of parataxis is preferable to a hypotactic structure since a coordinated construction is in principle more easy-to-read and comprehensible than a subordinate one (Piemontese, 1996) which, on the contrary, is cognitively more complex (Givón, 1991).

In contrast with the recommendations of the "Guide", as Table 9 shows, the results of the whole legal corpus are characterised by a higher percentage of subordinate clauses organised in long 'chains', and consequently by a higher subordinate/main clauses ratio.

Moreover, the legal corpus contains a higher percentual distribution of deeply embedded subordinate clauses. For example, 'chains' of 3 embedded subordinate clauses constitute 3.51% of the total amount of 'chains' of subordinate clauses occurring in the whole legal corpus while they have a coverage of only 2.89% in "La Repubblica" corpus and of 1.32% in "Due Parole". This quite differs from the distributions found in "Due Parole", less sharp but still statistically significant when the "La Repubblica" ratio is considered.

Table 9. Use of subordination in legal and newswire corpora.

		Legal	Rep	2Par
distribution of main vs subordinate clauses	main clauses	64.84	67.33	73.55
	subordinate clauses	35.16	32.36	26.14
subordinate/main clauses ratio		0.54	0.48	0.36
average depth of 'chains' of embedded subordinate clauses		1.25	1.17	1.01

In spite of the fact that subordination is typically taken as an index of structural complexity, as Mortara Garavelli (2003: 6-8) observed for the Italian legal texts, the use of hypotactic structures can be justified when they allow making plain the hierarchical order of pieces of discourse information otherwise hardly comprehensible. On the contrary, *horizontal* coordinated constructions may cause a conceptual burden not smaller than that of a *vertical* hypotactic structures. While coordinate connectives flatten the hierarchical organization of the discourse, an embedded subordinate construction allows keeping the ordered distribution of pieces of information, e.g. the cause/effect order, exceptions, conditions, etc. Therefore, the higher use of subordination in the legal texts might suggest that the logical structuring of legal discourse is typically expressed through hypotactic structures organised in a hierarchy.

Features concerning nominal modification

The interest in investigating these features stems from Mortara Garavelli's statement that legal sentences are characterised by embedding constructions of nominal modifiers that are typically prepositional complements (Mortara Garavelli, 2001: 171-175). According to her view, such syntactic behaviour causes *structural complexity* of sentences which can affect the *transparency and understandability* of legal documents.

Consider the following sample sentence extracted from the legal corpus:

(5) Il Consiglio è giunto ad un **accordo** _sui_ contributi dei singoli Stati membri _all'_adempimento _dell'_impegno globale _di_ riduzione _delle_ emissioni _della_ Comunità nelle conclusioni del Consiglio del 16 giugno 1998. ('The Council **agreed** *upon the contributions of each Member State to the overall Community reduction commitment* in the Council conclusions of 16 June 1998.')

In (5), the noun *accordo* (the verb 'agreed' in the English translation) is modified by a sequence of 6 embedded prepositional dependency links[7].

The results reported in Table 10 provide empirical validations of Mortara Garavelli's theoretical claims: the legal corpus is characterised by significantly longer complement 'chains'.

The legal corpus is also characterised by a higher percentual distribution of deeply embedded sequences of prepositional complements. In particular, legal texts are char-

Table 10. Nominal modification in legal and newswire corpora.

	Legal	Rep	2Par
average depth of embedded complement 'chains'	1.84	1.35	1.24

acterised by a lower percentage of sequences including one prepositional complement (53.41%) with respect to "La Repubblica" (73.32%) and "Due Parole" (79.40%), and by longer sequences including up to 6 complements. For example, 'chains' of 3 complements constitute 11.80% of the total amount of prepositional complement 'chains' occurring in the legal corpus while they have a coverage of only 4.64% in "La Repubblica" corpus and 2.71% in "Due Parole"; in addition the legal texts are characterised by 5.23% of sequences including 4 prepositional complements while in "La Repubblica" they constitute 0.99% and 0.48% in "Due Parole".

Different types of legal texts in comparison

The in depth analysis of the linguistic peculiarities of the different types of legal documents is illustrated in what follows. It aims at highlighting, on the one hand, the linguistic variation between documents used in a "legislative" and "juridical setting", and, on the other hand, the documents enacted or resolved by different authorities.

Different sub-genre of legal texts

The linguistic profiling of the legal documents used in a "legislative setting" (i.e. legislative and administrative acts as well as the Italian Constitution) and in a "juridical setting" (i.e. legal cases) has shown that they differ significantly in many respects at the level of raw and lexical features.

As Table 11 shows, the legal cases (*Cases*) with the longest sentences and the lowest percentage of lemmas (types) belonging to the "Basic Italian Vocabulary" (BIV) result to be the legal sub-genre most different from newspapers. On the contrary, the Italian Constitution (*Const*) is the most similar to them. It contains the highest percentage of words belonging to BIV as well as of the *fundamental* vocabulary and, interestingly, the shortest sentences – even shorter that "Due Parole" corpus (19.20). This is due to the fact that the Constitution witnesses the linguistic efforts of the founding fathers towards a simple and plain legislative drafting in principle comprehensible to a wide public of readers (De Mauro, 2006).

Table 11. Raw and lexical features in sub-genres of legal texts.

	Leg	Admin	Const	Cases
average sentence length	24.99	30.13	16.59	37.02
lemmas belonging to BIV	16.34	22.14	54.87	15.40
lemmas NOT belonging to BIV	83.66	77.86	45.13	84.60
fundamental vocabulary	47.06	49.62	62.65	47.87
high usage vocabulary	39.56	38.80	30.86	38.85
high availability vocabulary	13.38	11.58	6.50	13.29

If we consider the distribution of the morpho-syntactic characteristics, a number of variations can be observed. In particular, we can see that the administrative (*Admin*) and legislative (*Leg*) documents have a higher percentage of nouns and a lower percentage of

verbs than the legal cases. This affects the different values of the noun/verb ratio which is higher in *Admin* and *Leg* than in *Cases*. Following Biber's (1993) outcomes, these results suggest that we are dealing with different language varieties possibly pursuing different communicative purposes. We might put forward here the hypothesis that this provides empirical evidence of the acknowledged difference between the communicative purpose of the documents used in a "legislative setting" and in a "juridical setting". The first ones are in their nature *performative* but they also have distinguishable characteristics of *informative* texts since "every attempt is made to write not only clearly, precisely and unambiguously but also all-inclusively" (Bathia, 1987: 230). Conversely, legal cases serve several different communicative purposes (Bathia, 1993). In particular, the three-fold internal structure of Italian legal cases, which is overtly established by article 118 of the Italian Civil Procedure Code, corresponds to three different communicative purposes (Santulli, 2008): *narrative* (corresponding to the legal case section where the facts which are relevant for the case are reported), *argumentative* (the function of the section where the judge reports the motivations of the final decision) and *performative* (the function of the last section, i.e. the final decision).

This difference may affect the different noun/verb ratio values reported here.

Table 12. Morpho-syntactic features in sub-genres of legal texts.

	Leg	Admin	Const	Cases
prepositions	20.64	21.24	18.63	18.48
adjectives	8.16	8.21	8.40	6.82
nouns	30.27	31.17	30.16	29.09
verbs	8.59	8.47	11.50	10.78
noun/verb ratio	3.53	3.68	2.62	2.70

Moving to the analysis of the syntactic features, it results that the corpus of legal cases contains the deepest syntactic trees and the longest dependency links (see Table 13), i.e. two of the features mostly indicative of syntactic complexity (see Section 'Features based on the structure of a syntactic tree'). Conversely, the 'easiest' structures occur in the Constitution which shows values lower also with respect to "Due Parole" overtly written in a plain language. Among the documents used in a "legislative setting", the administrative acts result to be the most complex ones.

Table 13. Syntactic features in sub-genres of legal texts.

		Leg	Admin	Const	Cases
average syntactic tree depth		6.15	7.67	4.73	8.38
average length of the longest dependency links		12.28	12.61	6.75	14.43
distribution of main vs subordinate clauses	main clauses	73.39	68.70	86.07	52.43
	subordinate clauses	26.31	31.30	13.93	47.57
subordinate/main clauses ratio		0.36	0.46	0.16	0.91
average depth of 'chains' of embedded subordinate clauses		1.18	1.24	1.03	1.35

The four legal sub-corpora differ greatly with respect to the percentage distribution of the subordinate clauses. In particular, the legal cases contain the highest percentage of

subordinate clauses organised in deep 'chains': sequences of e.g. 2 embedded subordinate clauses constitute 20.58% of the total amount of 'chains' of subordinate clauses in this legal sub-genre while they are 15.75% in *Admin*, 12.54% in *Leg* and only 3.17% in *Const*. Consequently, legal cases have the highest subordinate/main clauses ratio. Interestingly, if we focus on the documents used in a "legislative setting", we can see that the Constitution shows values lower also with respect to "Due Parole" while the administrative acts contain the highest percentage of subordinate clauses organised in longer sequences with respect to the legislative texts.

The *Leg* corpus stands out for the greater use of nominal modification. In particular, it contains the highest percentual distribution of deep embedded sequences of prepositional complements. 'Chains' of e.g. 3 complements constitute 12.07% of the total amount of prepositional complement 'chains' while they are 11.10% in *Cases*, 9.77% in *Admin* and 5.67% in *Const*.

Legal texts enacted or resolved by different authorities

The "legislative setting"

Significant linguistic variation can be observed by comparing legal documents used in the same setting but released by different authorities. Starting from the analysis of the documents used in the "legislative setting", acts enacted by the European Commission, the Italian State or by the Piedmont Region differ significantly at the level of raw and lexical features.

The national (*AdminState* in all tables) and regional (*AdminReg*) administrative acts not only have the longest sentences but they also contain the lowest percentage of lemmas (types) belonging to the "Basic Italian Vocabulary" (BIV) and of fundamental vocabulary (see Table 14). The European administrative texts (*AdminEU*) represent the opposite pole with the shortest sentences and the highest percentage of BIV. A similar variation between national (*LegState*) and European (*LegEU*) legislative texts can be observed, even if regional legislative acts (*LegReg*) do not follow this trend.

Table 14. Raw and lexical features in legislative and administrative texts.

	LegState	LegReg	LegEU	AdminState	AdminReg	AdminEU
average sentence length	27.16	20.04	23.42	33.81	29.09	23.43
lemmas belonging to BIV	19.92	28.72	26.38	30.80	24.85	49.84
lemmas NOT belonging to BIV	80.08	71.28	73.62	69.20	75.15	50.16
fundamental vocabulary	53.18	48.12	51.35	52.67	52.41	60.50
high usage vocabulary	39.25	37.77	37.87	38.31	36.88	32.25
high availability vocabulary	12.63	9.05	10.78	9.01	10.70	7.25

Moving to the analysis of the morpho-syntactic features, both administrative and legislative acts enacted by the Italian State and the Piedmont Region result to be characterised by a higher percentage of prepositions and nouns with respect to the European documents and by a lower percentage of verbs (see Table 15). This affects the different noun/verb ratio (see Table 15).

Table 15. Morpho-syntactic features in legislative and administrative texts.

	LegState	LegReg	LegEU	AdminState	AdminReg	AdminEU
prepositions	21.48	20.68	19.27	21.08	21.45	20.15
nouns	30.56	31.35	29.52	30.42	31.72	29.95
verbs	8.13	6.42	9.86	8.40	8.37	9.95
noun/verb ratio	3.76	4.89	2.99	3.62	3.79	3.01

National and regional administrative acts resulted to be more syntactically complex than the European acts: they have deeper syntactic trees and longer dependency links (see Table 16). Similar to the national and regional legislative acts, they contain longer sequences of embedded prepositional complements: in *LegEU* and *AdminEU* 'chains' of e.g. 3 complements constitute 11.18% and 7.98% respectively of the total amount while they a coverage of 12.54% in *LegState*, 12.24% in *LegReg*, 9.71% in *AdminState* and 9.99% in *AdminReg*.

A distinguishing characteristic of the European documents is the higher occurrence of subordinated constructions with respect to national and regional documents. As Table 16 shows, both legislative and administrative European legal documents are characterised by a higher percentage of subordinate clauses and by a higher subordinate/main clauses ratio. As discussed in Section 'Features concerning the use of subordination', these results should be combined with the study of how pieces of discourse information are logically organized throughout a document. However, this syntactic behaviour confirms a tendency already observed in this study: the linguistic similarity of the European legal language to ordinary language is greater than the language variety used in the documents enacted by the Italian State and the Piedmont Region.

Table 16. Syntactic features in legislative and administrative texts.

		LegState	LegReg	LegEU	AdminState	AdminReg	AdminEU
average syntactic tree depth		6.24	5.46	6.24	8.19	7.55	6.50
average length of the longest dependency links		14.82	8.52	9.83	14.22	12.14	9.79
distribution of main vs subordinate clauses	main clauses	74.48	86.80	70.03	68.14	69.35	66.62
	subordinate clauses	25.52	13.20	29.97	31.86	30.65	33.38
subordinate/main clauses ratio		0.34	0.15	0.43	0.47	0.44	0.50

The "juridical setting"

A number of different linguistic peculiarities can be noted in the five sub-corpora of legal cases. As Table 17 shows, the documents issued by the Constitutional Court (*CasesConst* in all tables) contain the deepest syntactic trees and the longest dependency links while the legal cases resolved by Ordinary Tribunals (*CasesOrd*) and Administrative Courts (*CasesAdm*) have an opposite behaviour. Moreover, the *CasesConst* corpus contains the highest percentual distribution of deeply embedded sequences of prepositional complements. 'Chains' of 3 complements e.g. constitute 9.84% of the total amount while they

have a coverage of 9.27% in *CasesCass*, 8.14% in *CasesECHR*, 7.98% in *CasesOrd* and of 7.56% in *CasesAdm.*

Moving to the analysis of the use of subordination, it resulted that the documents issued by the Court of Civil Cassation (*CasesCass*) and by the European Convention on Human Rights Court (*CasesECHR*) are characterised by the highest occurrence of subordinated constructions. Even if, according to the literature on the topic, these results can be interpreted in various manner (see Section 'Features concerning the use of subordination'), they confirm the linguistic differences between Ordinary Tribunals and Administrative Courts cases, on the one side, and the other types of legal cases, on the other side.

Table 17. Syntactic features in legal cases.

		CasesAdm	CasesCass	CasesECHR	CasesConst	CasesOrd
average syntactic tree depth		6.73	8.48	8.76	8.76	7.24
average length of the longest dependency links		12.37	15.84	14.33	18.05	12.42
distribution of main vs subordinate clauses	main clauses	60.60	52.67	50.28	56.68	59.03
	subordinate clauses	39.40	47.33	49.72	43.32	40.97
subordinate/main clauses ratio		0.65	0.90	0.99	0.76	0.69

Interestingly, the corpus of Constitutional Court cases is characterised by the highest percentage of lemmas belonging to the "Basic Italian Vocabulary" (BIV) and by the highest percentage of *fundamental* vocabulary (see 18). Apparently inconsistent with the literature on text complexity, this demonstrates, on the contrary, how an exhaustive linguistic investigation should consider the complex interaction of different kinds of linguistic features. This result tells us that in *CasesConst* a basic vocabulary is used in complex syntactic constructions occurring in long sentences while in *CasesAdm* a more 'complex' vocabulary is used in less complex syntactic structures occurring in short sentences.

Table 18. Raw and lexical features in legal cases.

	CasesAdm	CasesCass	CasesECHR	CasesConst	CasesOrd
average sentence length	31.22	40.79	36.87	46.37	31.54
% of lemmas belonging to BIV	28.91	25.08	20.64	38.56	32.02
% of lemmas NOT belonging to BIV	71.09	74.92	79.36	61.44	67.98
% of *fundamental* vocabulary	54.80	56.48	50.62	59.09	57.67
% of *high usage* vocabulary	36.27	34.53	37.28	33.53	34.05
% of *high availability* vocabulary	8.93	8.99	12.09	7.39	8.28

Conclusion

In this paper, the author presented an NLP-based study aimed at performing the linguistic profiling of a corpus of different types of Italian legal texts exemplifying different sub-varieties of Italian legal language. She analysed the distribution of a wide range

of different linguistic features automatically extracted from text in order to investigate the linguistic variation between *i)* the considered corpus of legal texts and a corpus of newspaper articles representative of Italian ordinary language and *ii)* between different types of legal texts that have been compared at different levels of specificity.

The followed comparative approach has allowed the investigation of lexical, morpho-syntactic and syntactic characteristics which make the corpus of legal texts different from newspaper articles. Interestingly, the legal language resulted closer to the ordinary language used in "La Repubblica" corpus than in "Due Parole" corpus. It is particularly the case when we took into consideration the low percentage of lemmas belonging to the "Basic Italian Vocabulary" or features typically taken as indices of syntactic complexity such as syntactic tree depth and length of dependency links. This shows that it is very often the case that legal texts do not conform to the suggestions put by the "Guide to drafting administrative acts", the most up-to-date guide describing the lexical, morpho-syntactic and syntactic characteristics that a legal document is expected to have in order to be written in a *plain, simple and comprehensible language.*

The present study has also highlighted systematic differences between the considered sub-varieties of the legal language. The 'genre-internal' perspective of analysis adopted here has shown that significant linguistic variations exist not only among documents used in different settings but also among documents enacted or resolved by different authorities. It has been demonstrated that the Italian Constitution articles were written using a legal language variety very close to the language of "Due Parole" corpus that was specifically written using a plain and controlled language. This empirically witnesses the linguistic efforts of the founding fathers towards a simple and plain legislative drafting. We have also discussed which are the linguistic characteristics making the legal texts enacted by the European Commission more similar to the ordinary Italian than the acts enacted by the Italian State and the Piedmont Region. Finally, it has been shown that among the legal cases the ones resolved by the Constitutional Court have the greater number of features typically taken as indices of syntactic complexity.

If on the one hand the approach to the linguistic profiling proposed here has been devoted to showing how computational linguistic analysis techniques can help to shed light on some main peculiarities of the Italian legal language, providing quantitative validations of theoretical claims from the literature, on the other hand, different types of applications could benefit from the results of this study. They can be used as a starting point to identify areas of lexical, morpho-syntactic and/or syntactic complexity within a legal text and/or a single sentence in order to assess their readability. Similarly, the investigation of the wide range of linguistic characteristics carried out in this study might be exploited to perform a number of different computational forensic linguistics tasks – first and foremost authorship attribution.

Acknowledgments

The linguistic profiling methodology followed here represents an essential line of research of the ItaliaNLP Lab[8] at the Institute of Computational Linguistics "Antonio

Zampolli" (ILC-CNR). The author greatly acknowledges the help and contributions in particular of Simonetta Montemagni and Felice Dell'Orletta whose invaluable comments and ideas helped to carry out the work presented here.

Notes

[1] http://www.dueparole.it/

[2] The annotation format adheres to the standard CoNLL-2007 tabular format used in the "Shared Task on Dependency Parsing" (Nivre *et al.*, 2007).

[3] Evalita, 2009

[4] The Italian version of the "Guide" is available at http://www.pacto.it/content/view/416/48/

[5] In Figure 1, each consecutive dependency link is highlighted with a frame.

[6] Note that also punctuation is included.

[7] The prepositions heading the prepositional complement are underlied here. Note that the complement *dei singoli Stati Membri* ('of each State Member') has not been included in the embedded sequence since it modifies the noun *contributi* ('contributions').

[8] http://www.italianlp.it/

References

Attardi, G. (2006). Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06)*, 166–170, New York City, New York.

Bathia, V. K. (1987). Language of the law. *Language Teaching*, 20(4), 227–234.

Bathia, V. K. (1993). *Analysing genre. Language Use in Professional Settings.* London and New York: Longman.

Beaman, K. (1984). Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In D. Tannen and R. Freedle, Eds., *Coherence in Spoken and Written Discorse*, 45–80.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics Journal*, 19(2), 219–241.

Biber, D. and Conrad, S. (2009). *Register, genre, and style.* Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus linguistics. Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Cortelazzo, M. (1997). Lingua e diritto in Italia. il punto di vista dei linguisti. In L. Schena, Ed., *La lingua del diritto. Difficoltà traduttive. Applicazioni didattiche. Atti del primo Convegno Internazionale*, Milano: Roma, Cisu (Centro d'Informazione e Stampa Universitaria).

De Mauro, T. (2006). Introduzione. Il linguaggio della Costituzione. In *Costituzione della Repubblica Italiana (1947).* Torino: UTET.

Dell'Orletta, F. (2009). Ensemble system for Part-of-Speech tagging. In *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia.

Dell'Orletta, F., Marchi, S., Montemagni, S., Plank, B. and Venturi, G. (2012a). The SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts. In *Proceedings of the 4th Workshop on "Semantic Processing of Legal Texts"*, 42–51.

Dell'Orletta, F., Marchi, S., Montemagni, S., Venturi, G., Agnoloni, T. and Francesconi, E. (2012b). Domain Adaptation for Dependency Parsing at Evalita 2011. In *Working Notes of EVALITA 2011.*

Dell'Orletta, F. and Montemagni, S. (2012). Tecnologie linguistico-computazionali per la valutazione delle competenze linguistiche in ambito scolastico. In S. Ferreri, Ed., *Atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI)*, 343–359, Roma.

Dell'Orletta, F., Montemagni, S., Vecchi, E. M. and Venturi, G. (2011a). Tecnologie linguistico-computazionali per il monitoraggio delle competenze linguistiche di apprendenti l'italiano come L2. In G. C. Bruno, I. Caruso, M. Sanna and I. Vellecco, Eds., *Percorsi migranti: uomini, diritto, lavoro, linguaggi*, 319–336. Milano: McGraw–Hill.

Dell'Orletta, F., Montemagni, S. and Venturi, G. (2011b). READ-IT: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, 73–83.

Dell'Orletta, F., Montemagni, S. and Venturi, G. (2012c). Genre-oriented Readability Assessment: a Case Study. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education (SLP-TED)*, 91–98.

Dell'Orletta, F., Montemagni, S. and Venturi, G. (2013). Linguistic Profiling of Texts Across Textual Genre and Readability Level. An Exploratory Study on Italian Fictional Prose. In *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-2013)*, 189–197, Hissar.

Fornaciari, T. and Poesio, M. (2013). Automatic deception detection in Italian court cases. *Artificial Intelligence and Law*.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen and A. M. Zwicky, Eds., *Natural Language Parsing*, 129–189. Cambridge: Cambridge University Press.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1–76.

Gildea, D. (2001). Corpus variation and parser performance. In *Proceedings of Empirical Methods in Natural Language Processing*, 167–202.

Givón, T. (1991). Markedness in grammar: distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language*, 15(2), 335–370.

Hudson, R. (1995). Measuring syntactic difficulty. Unpublished paper.

Lazari, A. (2005). *Modelli e Paradigmi della Responsabilità dello Stato*. Torino: Giappichelli.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on pur capacity for processing information. *Psycological Review*, 63, 81–97.

Montemagni, S. (2013). Tecnologie linguistico-computazionali e il monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata*.

Mortara Garavelli, B. (2001). *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani.* Torino: Einaudi.

Mortara Garavelli, B. (2003). Strutture testuali e stereotipi nel linguaggio forense. In P. M. Biagini, Ed., *La lingua, la legge, la professione forense. Atti del convegno Accademia della Crusca*, 3–19„ Milano: Giuffrè.

Nivre, J., Hall, J., Kubler, S., McDonald, R., Nilsson, S., Riedel, S. and Yuret, D. (2007). The conll 2007 shared task on dependency parsing. In *Proceedings of the EMNLP–CoNLL*, 915–932.

Piemontese, M. E. (1996). *Capire e farsi capire. Teorie e tecniche della scrittura controllata.* Napoli: Tecnodid.

Rovere, G. (2005). *Capitoli di linguistica giuridica. Ricerche su corpora elettronici.* Alessandria: Edizioni dell'Orso.

Santulli, F. (2008). La sentenza come genere testuale: narrazione, argomentazione, per-
    formatività. In G. Garzone and F. Santulli, Eds., *Linguaggio giuridico e mondo contem-
    poraneo*, 207–238. Milano: Giuffrè.

Sousa-Silva, R., Sarmento, L., Grant, T., Oliveira, E. and Maia, B. (2010). Comparing
    Sentence-Level Features for Authorship Attribution in Portuguese. In *Proceedings of
    the 9th International Conference on Computational Processing of the Portuguese Lan-
    guage*, 51–54.

van Halteren, H. (2004). Linguistic Profiling for author recognition and verification. In
    Proceedings of the Association for Computational Linguistics, 200–207.

Venturi, G. (2012). Design and Development of TEMIS: a Syntactically and Semantically
    Annotated Corpus of Italian Legislative Texts. In *Proceedings of the 4th Workshop on
    "Semantic Processing of Legal Texts"*, 1–12.

The Linguistic Functions of 'Knowingly' and 'Intelligently' in Police Cautions

Margaret van Naerssen *

Abstract. *This is a report on one of several analyses of a videotaped police interview. This analysis focused on the delivery of the police caution (in the USA, Miranda warnings) which embodies certain constitutional rights. The goals of this paper are (a) to show how a linguistic analysis of the delivery of a police caution might help in determining whether a persons civil rights have been respected when being questioned by the police, and (b) to improve the acceptability of linguistic evidence.*

Keywords: Police caution, non-native speakers, comprehensibility.

Introduction

Forensic linguists are likely to have to address comprehensibility issues when working on cases involving communications between law enforcement officers and a suspect or witness. One common issue involves police warnings/ cautions about a person's right *not* to say anything that might be self-incriminating.

The goal of this paper is to show how a linguistic analysis of the interactions in the delivery of a police caution might help in determining whether a person's civil rights have been respected when being questioned by the police.

This paper begins with a brief mention of international civil rights, followed by a sampling of research on comprehensibility in regard to the US-based police caution known as the Miranda. The focus of the paper then is narrowed to a specific case in the USA.

As this paper is tied to a specific case the author worked on, in some places 1st person singular pronouns are used for a natural flow of ideas.

---

*Immaculata University, Pennsylvania, USA

Background

International Perspective

The case in this paper involves the rights under Miranda in the US legal system. However, for the international community of forensic linguists it is appropriate and useful to view these rights in international law: the 1966 International Covenant on Civil and Political Rights (1966). Eades, writing about interpreting needs, reminds us of the international protection of the rights of accused second language speakers in the legal process (2010: 64). Likewise, for this paper, Article 14 of this covenant also reflects other rights of those in detention. Below is the relevant excerpt on Article 14.

> **International Covenant on Civil and Political Rights, Article 14, Part 3**
> Tenets on the rights of individuals under criminal charges, including (d)…and to defend himself in person or through legal assistance of his own choosing;
> - to be informed, if he does not have legal assistance, of this right; and
> - to have legal assistance assigned to him, in any case where the interests of justice so require, and
> - without payment by him in any such case if he does not have sufficient means to pay for it…

Many nations have signed this Covenant; however the implementation of these rights varies around the world. For some countries signing the Covenant has been aspirational, setting ideals to work towards. In some cases it might be for political image. Still in others much effort has be applied towards protecting these rights. However, in such countries, legal practitioners may vary in their understanding and perspectives. There may also be political shifts affecting the implementation of the protection. Further details are beyond the scope of this paper. (Also see Universal Declaration of Human Rights)

The basic concepts appear in various forms in national legal documents around the world. For example, in Malaysia the key document is the Malaysian Criminal Procedures Code, Section 13. In the USA the police caution is grounded in the 5th Amendment of the US Constitution.

Police Caution in the USA

The Fifth Amendment in the US Constitution is intended to protect a person against self-incrimination. The amendment declares that 'No person …shall be compelled in any criminal case to be a witness against himself.' In *Miranda v Arizona*, 1966, the ruling of the US Supreme Court means that before interrogating an individual in the inherently coercive setting of custody, a government agent must warn the individual that

- he has a right to remain silent
- that anything he says may, and will be, used against him in the court
- he has the right to consult with an attorney and this is a continuing right
- that if he is indigent, an attorney will be appointed to represent him

Three necessary conditions apply for a person to waive those rights before an agent proceeds with a police interrogation. A waiver must be done 'voluntarily,' 'knowingly,' and 'intelligently'. However, the Supreme Court did not specify the wording, thus, it varies around the country.

Comprehensibility: 'Knowingly' and 'Intelligently'

'Knowingly' and 'intelligently' are especially difficult criteria to examine as they are mental conditions. Much has been written about the Miranda in the literature of both the legal community and in the field of reading comprehension. Questions have also been raised by forensic linguists. The literature on Miranda is only sampled here.

Eugene Briere's (1978) article gave direction for those working in the USA with non-native speakers. He used second language testing and readability formulas to show the difficulty a non-native English speaker might have with Miranda. He also recognized the limitations of readability formula. Shuy (1997: 182–185) explores ten questions about Miranda from a linguistics perspective. Of particular relevance here is his question 'What does "to understand" mean?' Included in this question is whether suspects always really understand when they answer 'yes.' Shuy calls for 'better measures of understanding than the feeble, self-reported measure.' More recently, for forensic linguists in the USA, Solan and Tiersma (2005: 78–79, 81, 87, 90) have pointed to the important work of Grisso (2003) on comprehensibility of Miranda by juveniles and vulnerable adults. Ainsworth (1993) also looks at vulnerable populations and Miranda. Comprehensive research by Rogers and colleagues looks at readability / comprehensibility by native speakers (Rogers *et al.*, 2007, 2011).

Comprehensibility continues to be a concept for which it is difficult to find linguistic evidence. In one case Shuy (1997: 185) compared the language patterns a suspect exhibited during the Miranda with the language the suspect used later during tape recordings between the suspect and his attorney.

Finally, a very thorough legal examination of comprehensibility and cultural/ linguistic issues has been compiled by Floralynn Einesman (2010). This particular article can leave one wondering how likely it would be to get a handle on the cross-cultural / linguistic issues in a way that might be persuasive to a court.

Case Overview

After reviewing the literature, I wondered what more could be done linguistically to try to assess 'knowingly' and 'intelligently'. Then as I began examining video recordings in a Miranda case from California, I was reminded how interactive the procedure is. Importantly, the interviewing agent is also part of that process.

I have chosen to reflect on the process of working on that case on the chance the process might be of use to those new to working with such cases or to those who might want to adapt some strategies in their own legal systems. Also, through this reflection I want to emphasize that even with experience, a linguist may still need to 'live with' the evidence. Frequently we need to let ideas float around some before coming to an understanding of what can be done with the language evidence and what the limitations are.

Introduction to Case

Most of the cases I work with involve persons who are non-native English speakers in socio-legal contexts outside of their past life experiences. However, in this particular case the Miranda proceedings were conducted in Spanish by a native Spanish speaking agent. Also, the Miranda rights text was written in Spanish. Unlike my other Miranda-related cases, the suspect's English proficiency was not at issue. Thus, some aspects of

the case are parallel to that of a native English-speaking agent communicating with a native English-speaking suspect.

Very importantly, however, the socio-legal context was new to him as he had never been in the USA, nor had he ever been arrested. He brought socio-cultural expectations from his own life experiences. The case is summarized in the box below.

Case Summary: US v Mr. C

- Mr. C is a 22-year-old Mexican male, a Spanish monolingual, from a low-income Tijuana barrio. He had had occasional jobs as a dishwasher and DJ-ing, reading lists of songs at a local radio station.
- Mr. C was offered a job across the border in an auto-repair shop: `Take this car and meet us at X restaurant to meet your new employer.`
- At a border stop, just across the Mexican-US border drugs were found in car. He was arrested for drug trafficking.
- He was processed at a border detention center and Mirandized.
- He was then transferred to a detention center in San Diego where he was Mirandized again and interviewed by a bilingual Homeland Security agent.
- The proceedings were conducted all in Spanish and were video recorded.

Key legal question: How much did he know about the drug trafficking?
Alternatively: Was he unknowingly set up by others?

Language and Relevant Third-party Evidence

As linguists we understand that we cannot get into the head of another person to know exactly what the person is/was thinking. Nevertheless, language can be one window into the mind, if sufficient language evidence is available. Then linguistic tools might be applied.

In this case there was extensive video recording. It covered not only the Miranda process, but also the lead-up to it and closing interactions after the Miranda was delivered. A bilingual transcription of the interview was also available.

After determining that my Spanish comprehension was strong enough to work with the language evidence in the video recording, I considered initial strategies. In this case for acceptable evidence, readability scores of the Miranda parts could be related to the actual delivery of the Miranda. Also, from the field of sociolinguistics Conversation Analysis could probably provide tools for examining the interactions.

I also received additional evidence in the form of a forensic evaluation report by bilingual neuro-psychologist, Dr. Y. This included a test score of the suspect's reading ability in Spanish and evaluation reporting on his cognitive processing abilities. In addition, there was a transcript of his testimony in a court hearing. I did not look at this evidence until after I completed my analyses of the police interview.

Legal and Linguistic Questions

As any researcher, I still had to identify and conceptualize the relevant questions to ask, starting from this legal question: *How likely is it that Mr. C understood his Miranda rights and that he 'knowingly' and 'intelligently' waived them in the interview?* Mr. Cs attorney posed the next question to consider:

Are Mr. Cs signatures/ initials on parts of the Miranda statements and his state-
ments of 'sí' (yes) and head nods in response to 'Do you understand?' questions,
sufficient evidence of his ability to have 'knowingly' and 'intelligently' waived
his rights?

Language is not used in isolation. An understanding of the socio-cultural context would
be critical, the police interview. I could draw on tools from sociolinguistics, specifically
Conversation Analysis. As I became more familiar with the interactions in the interview,
I began asking what other kind of linguistic evidence would I need? What would it look
like? Would I be able to see any patterns in the interactions. Where were the breakdowns
in communication.

As the interviewing agent dominated the communication, I soon realized that I
needed to analyze his contributions to the interactions and not just the minimal re-
sponses of Mr. C. After all, wasn't it the interviewing agent's responsibility to assure
that the Miranda was effectively delivered? This then led to the following these specific
linguistic questions about the interview.

1. *What linguistic evidence is there that the Interviewing Agent made appropriate ef-
   forts to check that Mr. C understood his Miranda Rights and the Waiver? (Compre-
   hension Checks)*
2. *What linguistic evidence is there that the Interviewing Agent (April 00) made ap-
   propriate efforts to clarify apparent points of confusion/ misunderstanding of the
   Statements of Rights/Waiver? (Assists)*

   Additional questions were also developed for the third party evidence:
3. *How does Dr. Y's forensic psychology report contribute to an understanding of Mr.
   C's reading ability in terms of comprehending the Miranda?*

   Expanding on the reading process, I also asked,
4. *What do reading and psycholinguistic research and practice say about reading pro-
   cesses that is relevant to Mr. C's reading of the Miranda statements?*

   For this I looked to work on reading comprehension (reading-aloud, Plain Lan-
   guage/Plain English, and readability testing, and Schema Theory, especially
   Rumelhart's foundational work (in Carrell and Eisterhold, 1983.

   Finally, recognizing that a law enforcement interview/ interrogation of a suspect
   is inherently stressful, and that this can affect the effectiveness of a person's
   communications, I asked:
5. *What linguistic evidence is there of potential stressful conditions?*

Only Questions 1–3 are covered in this paper. In the official expert report Questions
4 and 5 were explored and related to the findings for Questions 1–3.

Interview Context

A brief description of the interview context is provided here. The interview took place
in a federal detention center in Southern California. In addition to Mr. C, two agents
participated in the interview. The Interviewing Agent (Agent) was the Spanish-English
bilingual Homeland Security agent. The observing agent appeared to be a monolingual

English speaker and was an agent of the US Immigration and Customs Enforcement Department (ICE). They sat on one side of a table; Mr. C sat across the table from them.

The delivery of the Miranda began after the Agent had asked Mr. C some identifying questions. The Agent handed Mr. C a Spanish copy of the Miranda to read aloud. The Agent would prompt him to begin each Miranda statement by starting to read the first few words or by telling him to go ahead. After each statement, the Agent would check for understanding with the standard comprehension check: *Do you understand?* Mr. C was also then asked to initial each statement indicating he had understood it. The assumption made by the Agent was that being able to read aloud the statements was evidence of comprehension, and that Mr. C's initials confirmed this.

An example is given below. The Spanish used in the example and elsewhere, e.g., in tables in the Appendices, was taken from the bilingual transcripts from the federal case, and checked against the video. It would not have been appropriate to make changes in official records. Also, some language may reflect the idiolect of the bilingual federal agent and the spelling by the official bilingual transcriber. There were a few minor differences between the video and the transcript, but they were not relevant to the analyses reported here.

A reviewer of the manuscript for this paper appropriately raised a possible concern about the possibility that the 'errors' may have affected comprehensibility by the Suspect. It was not possible to determine this. These 'errors' remain in the tables in the Appendices. Any editorial attempts to make corrections could cause problems in the totals in the analyses. 'Errors' are underlined in the example below (but not in the tables).

```
Agent:     Sigue.
           Go ahead.
Suspect: Usted tiene el derecho de permanecer callado.
           You have the right to remain silent.
Agent:     ?Si entiende eso?
           Do you understand that?
Agent:     Ponga sus iniciales ahí, que entiende ese.
           Put your initials there if you understand that (pointing to the place
           on Mr. C's copy where he should put his initials).
```

In the two boxes below are the English and Spanish versions of the Miranda used in this case, taken from the transcripts. Readers may wish to guess which Miranda statements (parts) might be the easiest and the most difficult to read with understanding.

Procedures

To try to maintain my objectivity as much as possible, I realized the order in which the analyses were done would be important. I did not want the results from readability scores to subconsciously influence my examination of the interview interactions.

The procedures used are first presented in a summary form, followed by a more detailed description of some steps.

Summary of Steps for the Conversation and Readability Analyses
- Viewed video of interview in Spanish for overview
- Skimmed bilingual transcript once, especially English

English version of Miranda --

1. You have the right to remain silent.
2. Anything you say may be used against you by the Court or in any other proceedings.
3. You have the right to consult with an attorney before you make any statements or answer any questions.
4. You have the right to have an attorney present with you during the interrogation.
5. If you cannot afford an attorney, one will be provided to you before we ask you any questions if you wish.
6. If you decide to answer our questions now, you retain the right to stop the interrogation at any time or to stop the interrogation for purposes of consulting with an attorney.

Waiver Statement: I have been read and explained this statement of my rights, and I have understood these rights completely. I waive them freely and voluntarily without being threatened, intimidated and without promise of compensation.

Spanish Version of Miranda (Spanish edited)

1. Usted tiene derecho a permanecer callado.
2. Cualquier cosa que usted diga puede ser usada en su contra por el tribunal o en cualquier otro procedimiento.
3. Usted tiene derecho a consultar un abogado antes de hacer cualquier declaración o contestar a cualquier pregunta.
4. Usted tiene derecho a tener un abogado presente con usted durante el interrogatorio.
5. Si no puede pagar un abogado se le proporcionará uno antes de que le hagamos cualquier pregunta si usted lo desea
6. Si decide responder a nuestras preguntas ahora, usted retiene el derecho de detener el interrogatorio en cualquier momento o de detener el interrogatorio con el propósito de consultar a un abogado.

Me han leído y explicado esta declaración de mis derechos y he entendido completamente estos derechos. Renuncio a ellos libre y voluntariamente sin ser amenazado, intimidado/a y sin promesa de compensación.

- Viewed interview, making notes of communication breakdowns on Spanish transcript
- Examined communication breakdowns for Agent's functions of Assists & Comprehension Checks & classified them
- Examined responses by Mr. C to the Agent
- Derived Spanish version of Miranda from interview and examined English translation for readability (total and each statement)
- Lined up Comprehension Checks and Assists with Miranda statements

More Detailed Descriptions of Conversation and Readability Analyses
Conversation Analysis tools

First, all communication breakdowns in the interview were marked as the initial locations for beginning the analyses. Then communicative functions (what people do with language) were examined to better understand what the Agent was doing with language in the interview. Comprehension Checks and Assists by the agent were selected as the

functions. They were then identified and counted. Mr. C's responses to these were also noted.

All instances of Comprehensive Checks and Assists were then classified according to form. Only the standard Comprehension Check function was used: 'Do you understand?' 'There were five types of Reading Assists by the Agent: (1) filling in a missing word, (2) starting to read the statement, (3) completing the statement, (4) responding to a point of confusion, and (5) full <u>re</u>-reading.

Readability formulas

Two readability formulas were applied to the overall Miranda text and to each of the Miranda statements. The Flesch Ease of Reading and Flesch-Kincaid Grade Level were the formulas used. (Microsoft Word tools were used to apply these formulas.) The English translation of the Miranda was used (see *Findings* for an explanation). In addition, the number of words, sentence lengths, and the number of passive sentences were noted for each statement. Passive constructions are considered cognitively more difficult to process than active constructions.

Matching assistance to Miranda

The Comprehension Checks and Assists were then matched to the Miranda statements where they had occurred. As the Miranda statements were already marked with the readability scores and the other data mentioned above, this matching allowed further observations about reading difficulty.

Using third-party evidence

Once all the data had been compiled, reference was then made to the third-party evidence: Mr. C's Spanish first language reading score comes from the bilingual neuropsychologist's report.

Findings

Assistance by the Agent

Findings from the analyses of the interactions between the Homeland Security agent were classified according to Comprehension Checks and Assists. The Homeland Security agent provided limited to minimal assistance to Mr. C. in terms of checking his comprehension and assisting him in his reading of the Miranda. There were 11 Confirmation Checks and 16 Assists. The types of assistance and totals are discussed below. (Also, see Appendices A and B for additional details.)

Not surprising, all 11 of the Comprehension Checks, were very close variations in form of 'Do you understand?' as this is the standard follow-up for each Miranda statement. There was no follow-up to verify comprehension beyond this even when there were indications early on that Mr. C was having problems. The pace of the interview also did not allow Mr. C the conversational space to easily ask for clarifications.

Of the 16 Reading Assists, the most frequent was 'filling in a missing word' (9/16). There were also two instances of each type of "partial reading" (starting or completing a statement). There were also two attempts to respond to a specific point of confusion.

Finally, there was one example of the Agent fully re-reading a statement. This appeared after the statement where the greatest assistance was needed: 3 Comprehension Checks and 5 Reading Assists. Perhaps the Agent felt that with all the confusion, he wanted to be sure a fully coherent statement was read aloud and recorded. This occurred with the Continuing Rights statement.

> If you decide to answer our questions now, you still retain the right To stop the interrogation at any time or stop the interrogation for the purpose of consulting with your attorney.

> *Si decide responder a nuestras preguntas ahora, usted retiene el derecho de detener el interrogatorio en cualquier momento o de detener el interrogatorio para el propósito de consultar con su abogado.*

This linguistic evidence raised a clear question of Mr. C's comprehension of this particular right.

Several of the 16 Assists appear to have caused confusion. One of these occurred at the beginning when Mr. C was informed that they were going to read his rights.

Mr. C: *What _are_ my rights?*

Agent: *We're going to read them!*

It appeared that the Agent assumed that Mr. C needed no background knowledge to understand the text of the Miranda statements. The Agent appeared to believe that if Mr. C would simply read a text aloud, or would listen to the Agent reading parts of it aloud, this would be evidence of comprehension on the part of Mr. C.

Notably, best practices were not applied. The Agent did not give Mr. C the chance to paraphrase each Miranda statement in his own words. This strategy has been recommended as best practice for testing for Miranda comprehension with juveniles and vulnerable adults (Grisso, 2003) and is a standard in some interview. This is also best practice in educational and training to asses comprehension of concepts.

Data from the readability formulas that are presented in the next section raised further questions about Mr. C's comprehension of his rights, about whether Mr. C 'knowingly' and 'intelligently' waived his rights.

Readability

Observations about the readability of the Miranda are organized around three topics: (1) readability data, (2) Spanish Miranda, and (3) additional readability issues.

Readability data

As noted earlier, there is no single, standard form of Miranda. Thus, the readability scores for the various parts of this Miranda should not automatically be assumed for all forms of Miranda across the country. The Miranda used in this interview requires at least a 10.5 grade reading level for the overall rights statement and as high as a 17.3 reading grade level for the Waiver. Mr. C had a 5th grade Spanish reading level. The contrast

between Mr. C's reading ability and the reading difficulty of the Miranda is illustrated in the grid below.

[scale=0.70]Img/Image5

In Table 1 below the readability statistics are given for the Miranda. As can be seen the easiest to read is the right to remain silent statement. However, Rogers and his colleagues (2007; 2011) note that even by US citizens, "silence" can be interpreted in various ways in the context of a law enforcement communication. The remaining Miranda statements are all beyond Mr. C's reading level. The Continuing Right statement (6) and the Waiver (7) are similar in terms of being the most difficult. They exhibit similar readability features and scores.

Table 1. Readability Statistics for Miranda

English Miranda	Number of Sentences	Number of Words	Number of Passive sentences (relatively more complex than Active sentences)	Flesch Ease of Reading (the higher, the easier, with 100 being easiest)	Flesch-Kincaid Grade Level
Total Miranda Rights/ Waiver	7	140	28%	54.8	10.5
1.Silence	1	7	0	90.9	2.3
2.Used Against	1	16	100%	68.9	7.6
3.Consult an attorney	1	18	0	56.9	9.7
4.During Interrogation	1	14	0	59.6	8.4
5.Can't Pay for an attorney	1	21	100%	64.6	9.4
6.Continuing Right: Stop to Consult attorney	1	31	0	38.9	15.5*
7.Waiver	1	33	0	29.7	17.3*

Spanish Miranda

The Miranda rights were given in Spanish and translated for the bilingual transcript. The readability statistics given in Table 1 and discussed here are for the English version. Comparable readability statistics (following the Flesch Ease of Reading and Flesch-Kincaid Grade Level tools) were not available for Spanish. A different tool was considered for the Spanish one but was not used in this analysis. The assumptions about that tool differed from those of the Flesh and-Kincaid formulas, thus, it was felt they could not be appropriately compared.

Also, the Border Patrol agent who first Mirandized Mr. C did so in Spanish. Later in a motion hearing in court he translated the Miranda, on-the-spot, into English. This was acceptable to the court.

Furthermore, when comparing the text data available for both versions in the bilingual transcript, similarities could be seen for number of words, sentences, and Passive Sentences. Both also use a number of more formal, less common vocabulary items. For this case there was not enough time to do a detailed comparison of the Spanish and English form of Miranda beyond that shown in Appendix C.

Nevertheless, some informal observations were made. Mr. C had difficulty simply decoding some of the formal, less common words in Spanish. Thus, it is unlikely then that he clearly understood the actual meaning. Additionally, two Spanish verbs, *deber* (modal) and *conocer* caused confusion. Further discussion of these problems are beyond the scope of this paper.

Additional readability issues

The neuro-psychologist's report provided additional information on Mr. C's reading ability, beyond his Spanish reading scores. He has difficulty comprehending and understanding information because of his low intellectual skills. This could affect his reading comprehension at a meaningful level. Also, apparently Mr. C had little familiarity with the US legal system. This meant that he probably did not have the necessary background knowledge/schema for understanding Miranda at a meaningful level.

Closing Observations

Conclusions

I wish to conclude with my official opinion in the formal expert report on this case:

> 'In my professional opinion and to a reasonable degree of certainty, it is highly unlikely that Mr. C clearly understood his Miranda Rights and that he could have knowingly and intelligently waived them.'

This opinion was based on conclusions resulting from linguistically-based analyses of the language evidence combined with readability assessments and other relevant theory and practice from the field of reading. There was also additional confirming information from a forensic psychologist's evaluation.

Readability research has been the main tool for examining comprehension of Miranda, but researchers in this area (including Rogers *et al.*, 2007, 2011) recognize the limitations of readability formulas for this purpose. However, in this case adding analyses of the functions of the agent's delivery of Miranda has been shown to be a very productive strategy for examining the criteria of 'knowingly' and 'intelligently.' It would be valuable to test it further in other cases where video recordings are available.

If adequate language evidence is available this strategy might also be relevant in cases where English is the medium of communication, involving English first language speaking suspects as well as non-native speakers. It could, of course, also apply in other language combinations.

Post script: What happened to Mr. C.?

As Mr. C could not afford an attorney, his case had been assigned by the court to an attorney. Minimal expenses were allowed for the attorney's work and the hiring of two experts.

Once the judge read the expert reports, he apparently found the reports were strong. As the prosecution could not find a counter-expert for the linguistics report, the judge then suggested both sides settle out of court and not waste more time and money.

Plea bargaining took place. Mr C was charged with a lesser crime, that of lying to a federal agent, apparently a lie of omission. Mr. C had not told the agent the other reason he wanted to go to California besides for work. He wanted to see a certain young lady who had moved there with her family. Love blinds judgment! Mr. C was sentenced to he time he had already served in jail and was on a bus that afternoon, back to Mexico in time for the Christmas holidays.

In closing, I wish to recall the legend of the Cock of Barcelos (Galo de Barcelos) in recognition of the location of the 2012 IAFL Conference in Portugal. *Galo de Barcelos* is a beloved national symbol. The legend involves a miraculous intervention by an already-cooked rooster to save the life of a man falsely accused and sentenced to be hanged. The accused proclaimed his innocence, declaring that the rooster on the judge's dinner table would crow. On hearing the rooster crow, the judge set the man free.

While linguists cannot persuade judges with miracles, as in the legend of the *Galo de Barcelos*, sometimes we can persuade them with linguistic analyses.

References

Ainsworth, J. (1993). In a Different Register: The Pragmatics of Powerlessness in Police Interrogation. *Yale Law Review*, 103(2), 259–266.

Briere, E. (1978). Limited English speakers and the Miranda rights. *TESOL Quarterly*, 12(3), 235–245.

Carrell, P. and Eisterhold, J. (1983). Schema Theory and ESL Reading Pedagogy. *TESOL Quarterly*, 17(4), 553–573.

Eades, D. (2010). *Sociolinguistics and the Legal Process*. Toronto: Multilingual Matters.

Einesman, F. (2010). Cultural issues in motions to suppress. In L. Ramirez, Ed., *Cultural Issues in Criminal Defense*. Huntington, NY: Juris Publishing, 3 edition.

Grisso, T. (2003). *Evaluating competencies: Forensic assessments and instruments*. New York: Kluwer Academic/ Plenum Publishers.

Rogers, R., Harrison, K., Shuman, D., Sewell, K. and Hazelwood, L. (2007). An Analysis of Miranda Warnings and Waivers: Comprehension and Coverage. *Law and Human Behavior*, 31(2), 177–192.

Rogers, R., Rogstad, J., Steadham, J., Drogin, E. and Eric, Y. (2011). In plain English: Avoiding recognized problems with Miranda miscomprehension. *Psychology, Public Policy, and Law*, 17(2), 264–285.

Shuy, R. (1997). Ten Unanswered Language Questions about Miranda. *Forensic Linguistics*, 4(2), 175–196.

Solan, L. and Tiersma, P. (2005). *Speaking of Crime: The Language of Criminal Justice*. Chicago: University of Chicago Press.

Appendix A: Agent's Confirmation Checks During Reading of Miranda Rights/ Waiver

Text references and coding are based on the expert report. The Spanish from the transcript/video are used. Any standardization or 'corrections' of the Spanish might affect the word counts; therefore, 'errors' remain in the tables.

Wording of Miranda Rights/Waiver	Number of words	Number of Confirmation Checks (CC) & Type	Confirmation Checks
Opening: 0. **Antes de que hagamos cualquier pregunta usted debe conocer sus derechos**.	11 words	1 CC (#2)	(p.18:21) Do you understand that? *(?Si entiende eso?)*
1. **Usted tiene derecho de permanecer callado**.	6 words	1 CC (#2)	(p.19:1) Do you understand that? *(?Si entiende eso?")*
2. **Cualquier cosa que, que usted diga su contra por el tribunal o en cualquier otro procedimiento**.	16 words	1 CC (#2)	(p.19:10) Do you understand that? *(?Si entiende eso?")*
3. **Usted tiene derecho a consultar algún abogado antes de que haga cualquier declaración o conteste cualquier pregunta**.	17 words	2 CCs (#2) (#2)	(p.19:18) Do you understand that? *(?Si entiende eso?)* (p.20:3) Do you understand that? *(?Entiende eso?")*
4. **Usted tiene derecho a tener un abogado presente con usted durante el interrogatorio**.	13 words	1 CC (#2)	(p.20:18) Do you understand that? *(?Si entiende eso?")*
5. **Si no puede pagar un abogado se le proporcionara uno antes de que le hagamos cualquier pregunta se usted lo desea**.	21 words	1 CC (#2)	(p.20:29) Do you understand that? *(?Si entiende eso?)*

6. Si decide responder a nuestras preguntas ahora, usted retiene el derecho de detener el interrogatorio en cualquier momento o de detener el interrogatorio para el propósito de consultar con su abogado.	31 words	3 CCs (#2) (#2) (#2)	(p.21:17) Do you understand that? *(?Si entiende eso?)* (p.21:33) Do you understand? *(?Entiende?)* (p.22:7) But do you understand that right? *(Pero, ?si entiende ese derecho?)*
Waiver: 7. **Me han leído y explicado esta declaración de mis derechos y entendido completamente estos derechos. Renuncio a ellos libre y voluntariamente sin ser amenazado a mi intimidado.** (15+12= 27 words)	27 words	1 CC (#2)	(p.23:1) Do you understand that *(?Si entiende eso?)*
Closing Affirming signature page Began reading closing form wording + info on time and date **Fue detenido a las once, trein..hor, fecha….**	**Total words bolded:** **142**	**Total Confirmation Checks** **11**	

Appendix B: Agent's Assists During Reading Aloud of Miranda Rights/Waiver

Text references and coding are based on the expert report. The Spanish from the transcript/video are used. Any standardization or 'corrections' of the Spanish might affect the word counts; therefore, 'errors' remain in the tables.

Wording of Miranda Rights/Waiver	Number of words	Number of reading "assists" and types	Specific Assists & Types **#1 By filling in a word** **#2 By starting to read the statement** **#3 By completing the statement** **#4 By clarifying a misunder-standing, responding to confusion expressed** **#5 By re-reading aloud (completely)**
Opening: 0. **Antes de que hagamos cualquier pregunta usted debe conocer sus derechos.**	11 words	**2 assist** (#1) (#4)	 p 18:16) "conocer" (p.18:29) Twice Mr. C interrupts Agent's efforts to check his comprehension and to initial the form. 1) Mr. C ,'Yes, but they are my rights.' Agent, having gotten the "yes," tries to get Mr. C initial, 'ponga.' 2) Mr. C still appears confused, 'What are my rights?' Agent assists by saying "We're going to read them now." (18:29)
1. **Usted tiene derecho de permanecer callado**.	6 words	0 assists	
2. **Cualquier cosa que, que usted diga puede ser usada en su contra por el tribunal o en cualquier otro procedimiento**.	20 words	1 assist (#3)	(p19:8/9) Overlap of final word "procedimiento"
3. **Usted tiene derecho a consultar algún abogado antes de que haga cualquier declaración o conteste cualquier pregunta.**	17 words	3 assists (#1) (#4) (#5)	(p19:12) "Usted tiene derecho.." (p.19:31) when Mr. C said he hadn't consulted any attorneys, the Agent clarified that Mr. C had that right. (p. 19:35-20:2) Agent re-readings right but with one error/corrects it

4. **Usted tiene derecho a tener un abogado presente con usted durante el interrogatorio.**	13 words	2 assists (#1) (#3)	(p.20:9) after Seguimos. Otra vez. (beginning another right "Usted tiene.." (p 20:13-14) A tener un abogado presente con usted durante [el interrogatorio [OL w/ C]
5. **Si no puede pagar un abogado se le proporcionara uno antes de que le hagamos cualquier pregunta se usted lo desea.**	21 words	0 assists	
6. **Si decide responder a nuestras preguntas ahora, usted retiene el derecho de detener el interrogatorio en cualquier momento o de detener el interrogatorio para el propósito de consultar con su abogado.**	31 words	5 assists (#1) (#1) (#2) (#1) (#1)	(p20:34) "decide" (p21:2) "responder" (continuing same statement as "decide") (p21:5) "Usted…" [OL with "usted tiene el derech…" (p21:10) "aun…" [okay.] (p21:25) decide responder… (repeating C's "decide")
Waiver: 7. **Me han leído y explicado esta declaración de mis derechos y entendido completamente estos derechos. Renuncio a ellos libre y voluntariamente sin ser amenazado a mi intimidado.**	27 words	3 assists (#2) (#1) (#1)	(p22:12) [Okay. Lea ese parrafo:] "Me han…" (p22:21): "intimidado" (p22:39) "immunidad"
Closing Affirming signature page Began reading closing form wording + info on time and date **Fue detenido a las once, trein..hor, fecha**….	**Total words bolded: 142**	**Total Assists: 16**	

Appendix C: Spanish Readability for Miranda: Text Data Using Microsoft Tools

Spanish Miranda	Number of Sentences	Number of Words	Number of Passive sentences (relatively more complex than Active sentences)	Flesch Ease of Reading (Not available for Spanish)	Flesch-Kincaid Grade Level (Not available for Spanish)
Total Miranda Rights/ Waiver (opening not included here)	7	131	37.5%		
1.Silence	1	6	0		
2.Used against You	1	16	100%		
3.Consult an attorney	1	17	0		
4.During interrogation	1	13	0		
5.Can't pay for an attorney	1	21	100%		
6.Continuing right: Stop to Consult an attorney	1	31	0		
7.Waiver	2	27(15+12)	50%		

Language and Law: ways to bridge the gap(s)

Virgínia Colares *

Abstract. *This essay explores the gap(s) between language and law regarding the concepts of language and how they affect legal practice. It seeks to identify and bridge the gap(s) between the two fields of knowledge, articulating the elements that permeate across, over and through the two disciplines in an attempt to understand such complexity. The reflection focuses on how the following three topics are viewed both from the perspective of the study of language and from the perspective of law: (1) the meanings of "pragmatic" and "pragmatism" with historically diverse backgrounds and deep epistemological differences in terms of scientific practices, (2) as a consequence, the different conceptions of language , language, text, speech etc., and (3) the lack of discussion on the research methodology, interpretation and production of meaning in the two sciences. The conclusion is that the bridge between language and law has not yet been built because the many attempts that have been made towards that direction focused on abstract models of ideal languages from the perspective of linguistics; and Law has built its point of view by developing a language concept based on the common sense of language teachers. All of this being governed by the paradigm of correction or based on classical logic and rhetoric.*

*Keywords: **Philosophy of Language, Linguistics Pragmatics, Legal Pragmatics.***

Introduction

What are the gaps between Language and Law? Identifying these gaps is the first step. I have observed that since the 1980s, in Brazil, the relationship between Language & Law or Law & Language, has not been consolidated as a research topic. We have just started an exchange between the two areas. The main obstacle is the lack of familiarization of lawyers with studies of linguists and vice versa. At the conference "Language & Law: the multiple turns and new research agendas in Law", carried out at the Law Department of the Catholic University of Pernambuco, in Recife, from the 3rd to the 6th of September, 2012, we were able to observe some of the difficulties in understanding how each field

*Universidade Católica de Pernambuco – UNICAP

builds its knowledge. We meet again here, on the other side of the ocean, at another law school, trying to bridge the gap(s) between Language and Law. I noticed at that past event, as well as in the works that I reviewed for this 3rd European Conference, that the main gap(s) between our different fields of knowledge are:

1. The knowledge we build from the various meanings of "pragmatics" and "pragmatism", with historically different origins and profound epistemological differences in our scientific work;
2. And consequently, different conceptions of language, words, text, discourse, interpretation etc.;
3. The lack of debate on the methodology of investigation of the two sciences, since its formulation in ancient Greek thought until today.

I will reflect upon these three topics, trying to identify gaps and connect the two knowledge domains, articulating elements that pass between, beyond and through the two disciplines, in an attempt to understand the complexity.

Linguistics Applied to Law

The application of linguistics in situations of judgment reflects a discourse trying to modernize justice. The studies of language in the legal area each represent a possible application of the assumptions and postulates of linguistics to the language 'of' courts and to the language 'in' courts in the various contexts of use, oral or written, in legal practice. Such studies do not yet characterize a specific and systematic exchange as in psycholinguistics, sociolinguistics, legal sociology and many other areas in which the theoretical and methodological apparatus are brought into cooperation for the interdisciplinary construction of another object of study.

Apparently, this exchange, so important for building bridges, will not take place so easily. In his article, Hutton (1996: 205–214), in a vehement tone, questions and makes objections on the relevance of the linguistic concepts for interpretive disciplines such as literary criticism, and especially Law. Regarding the contributions of forensic linguists, Hutton lists three types of obstacles:

(1) Linguists' theoretical concepts, and their central postulates as a discipline, building specific metalanguages:

(a) providing evidence to the court,

(b) producing and observing transcriptions and (c) identifying individual voices, constitute a specialized discourse on the idealizations of linguists and on the abstractions that linguists have been making outside the current "language behavior."

Hutton states that linguists propose that transcripts have the same importance of oral materials produced in court, under questionable assumptions that the courts operate interpreting fragments of speech or reading transcriptions. The principal objection for him is that the linguistic analysis does not account for the meaning of a specific statement on a specific occasion. "Linguistics has, in effect, developed a third realm of nature for itself, treating language as sui generis and developing for it a corresponding unique methodology and terminology, shared neither by the social sciences nor by the natural sciences" (Hutton, 1996: 209).

(2) Linguistics is not a science in the same proportion as Chemistry is. "Linguistic intuitions were a means into the system, not a means of analyzing the system. In this sense

linguistic analyses are subject neither to empirical confirmation or disconfirmation, nor are they accountable to social judgments about their correctness, relevance, plausibility or importance" (Hutton, 1996: 209). The author introduces his argument taking as a starting point the fact that linguistics is not an experimental science and does not work with invariants. Its method does not have the property of repeatability, nor is finalistic (practical purpose) or timeless, etc. Hutton accuses Linguistics of making use of *post factum* procedures, intuitions, simplification, using internal theories from heuristic categories and procedures:

By searching for fundamentals in the philosophy of science to develop his critique, the author has omitted at least thirty years in the history of schools of thought outlined by epistemology. The crisis of the humanities, the sciences that have man as an object of study, to which the author refers, went through successive historic moments: Humanism, Positivism (Comte), Historicism (Dilthey), Relativism (Max Weber). The specificity of the objects of study of the humanities seeks its foundations in the notion of phenomenon, in a refusal to treat societies as stages of culture and civilization in a universal historical process as the natural sciences do based on Darwin's notion of evolution. Structuralism shows that the exchange and circulation of specific objects is a way to build society as a whole, so the exchange and circulation of the word, of the linguistic systems, organizes and relates to other symbolic systems and defines the general and specific structure of a society, organizing social relations. The postulates of Marxism allow us to understand the human elements (historical and social institutions) with their links on historicity and materiality of human existence, allowing the rational interpretation of the various plans that overlap.

The anachronism of the ideal of science conceived by Hutton (1996) in the arguments developed in order to disallow linguistics as a competent domain of knowledge to give support to Law, however, gives rise to the possibility of working in an interdisciplinary way to point out the third obstacle:

(3) Linguists work with the idealization of linguistic behavior in the pursuit of certain invariant categories among all the categories, in the world's 'disorder', without depleting the complexity of the phenomena under study, based on remote conceptions of language. In this respect – idealization – the author establishes equivalence between the attitudes of lawyers and linguists, when faced with the categorization process (imputed to linguists) and when faced with the act of classification inherent to judicial rulings (imputed to law professionals). Because "Neither lawyers nor linguists have a monopoly of the truth, and both could learn from each other, and benefit from the chance to examine each other's presupposition about language". (Hutton, 1996: 209)

"Pragmatics" and "Pragmatism"

The first topic that I would like to consider is the various meanings of "pragmatics" and "pragmatism" that are adopted by both linguists and lawyers. The terms "pragmatics" and "pragmatism" have been used interchangeably or equivalently, and they do not mean the same thing. *Pragma*, a word of Greek origin, means thing or object, especially in the sense of something made or produced, and the verb *pracein* means to act or to do. The Romans translated *pragma* into the Latin word *res*, which is the generic term for thing, losing, in the translation, the "act or do", a meaning that was present in the original

Greek term. That explains the existence of similar names for practices/interpretations deeply different from each other.

"Pragmatics" and "pragmatism" are notions used concurrently in Philosophy, Language studies and Law. In a broad sense, "pragmatism" or "pragmatic philosophy" refers to concepts of philosophy advocating not only a distinction between theory and practice, but above all, the primacy of practical reason over theoretical reason. Its fundamentals are based in Kant, whose last work was called *Anthropology from a Pragmatic Point of View* [1798], as well as in some of the contemporary philosophy schools. In the United States, the philosopher Charles Sanders Peirce (1977)applies the notion of pragmatism not only to the Sign Theory that he developed within his research into semiotics, but to the conception of truth that he advocates in his definition of science. According to Peirce, scientific theories are sets of hypothesis whose validity can only be determined by taking into account their effectiveness and success, i.e., their results, effects and consequences, thus the practice of science itself. The psychologist William James and Oliver Wendell Holmes (1995), an American judge of the Supreme Court of the United States embraced his ideas. William James (1997), while also adopting the criterion of truth as success and effectiveness, gave them a more psychological and moral nature. He developed a "pragmatic philosophy" or "utilitarian philosophy" which led Peirce into adopting the term "pragmaticism" in order to dissociate himself from James, and characterize his own conception of semiotics.

Later, John Dewey (1971) followed James' ideas, developing a philosophy focusing on the practice, in an ethical and practical sense, analyzing society and culture, and proposing a philosophical system that combined the scientific study of psychology with the German Idealism. He exercised great influence in the American philosophy in the 1930s and the 1940s. Richard Rorty (1982) stands out in advocating what has been characterized as neo-pragmatism. Harvard's philosopher Stanley Cavell (1995) can also be included among those who represent contemporary pragmatism.

Law's reality exists only through its verbal expression. That is the reason for the legal philosophers' concern with the linguistic turn, to which they repeatedly refer. However, the linguistic turn operated last century is based on the assumption that is common to many areas of human and social knowledge: the fact that language (in its syntactical, formal, logical, structural, semantic, discursive aspects) allows operations such as thinking, knowing, deducting; i.e., operations supposedly "mental" or "cognitive" derived from the split between thought and language are nothing more than a dichotomizing construct of positivism. In this sense, why call it "*neo-pragmatism*" if the approach is still positivist?

The so-called pragmatic turn is led, in German philosophy, by Jürgen Habermas (1987; 1990; 1994) and his friend and collaborator Karl-Otto Apel (1996) inspired by Peirce's pragmatism and pragmatic philosophy of language. They developed the pragmatics concept by focusing on the analysis of the conditions of possibility of communication, their assumptions and their implications, in the fields of ethics and politics, and considering the discourse theory law.

The Habermasian theory of communicative competence arises in the field of law as a new way to articulate and justify a broader conception of rationality in order to rethink the normative foundations of the legal theory in the society. Habermas, however, pre-

pares his theory of communicative competence, starting by defining universal pragmatics, based on Chomsky's transformational generative grammar (Chomsky, 1975). The author evokes the notions of linguistic competence and linguistic performance with a view to the linguistic universals theory (formal and substantive). The task of the theory of communicative competence is to explain the operations performed by speaker and listener, with the aid of pragmatic universals, when they use sentences (or extra-verbal expressions) in verbalizations that Habermans called "utterances", in order to establish an understanding of the state of things. The contexts of certain speech situations also contain various extralinguistic elements (e.g., the speaker's psychic constitution, their factual knowledge, skills, etc.), which constitute the object of empirical pragmatics. Regarding this, Habermas also refers to the speech acts theory developed by Austin (1977) and Searle (1969).

The use of pragmatics, as a background, also applies to the so-called hermeneutic turn promoted by Heidegger (2003) and Gadamer (2002) in philosophical hermeneutics, a contemporary theory that emerged in the middle of the twentieth century and was characterized, in broad terms, by the idea that truth is the consequence of an interpretation. Pragmatics, in this case, leads to relativism and the criticism that is made is that relativism is equivalent to "anything goes", both from an ethical standpoint, as for knowledge production. In the legal field, relativism is criticized for generating social insecurity, where "everything" becomes contingent. We will not discuss here the question of the limits of hermeneutics and the idea of pre-understanding that the methodological principles of interpretation and explanation assume.

But the linguistic pragmatics presupposes a conception of language according to which the meaning is relative to certain contexts and should be considered based on linguistic terms and expressions used in these contexts. This is not equivalent to the "anything goes" of the relativism because the meaning is not seen as arbitrary, but as context dependent. Usage involves the determination of rules and conditions that characterize the specific contexts in which meaning is constituted. Claiming that meaning is "relative to the context" is not the same as confirming semantic, cognitive, or ethical "relativism", if "relativism" means that all points of view are equivalent and are equally valid. On the contrary, the consideration of rules, conventions and conditions excludes arbitrariness, explaining the process of constitution and alteration of the meaning of a word or linguistic expression as dependent, even on who produces it.

As can be seen, linguists and legal professionals build their objects of study under various theoretical perspectives and assumptions. In the field of linguistic pragmatics, Charles William Morris (1979) was the first to use this term, contemporaneously. The author distinguishes three levels of linguistic analysis: syntax (interrelation of the signs), semantics (relation between the signs and the world), and pragmatics to designate the study of the relation between signs and interpreters. Rudolf Carnap (2002), the logician and philosopher of science of German origin, with whom Morris has worked in Chicago, defined pragmatics as the study of language in relation to its speakers or users. More recently, the term "pragmatics" has come to encompass all language studies related to its use in communication.

The philosophy of pragmatic language values common language and the concrete use of language as its main research focus, considering semantics and syntax only as constructs of abstract theoretical. The ordinary language philosophy of Gilbert Ryle

(2002), the speech act theory of Austin, the conception of language games of Wittgenstein, among others, may be included in this perspective.

In summary, it is basically a philosophical view according to which the study of language should be conducted in a pragmatic perspective, i.e. as a concrete social practice, therefore examining the constitution of linguistic meaning in the interaction between speaker and listener, the context of use, the socio-cultural elements implied by use, and goals, and the effects and consequences of these uses. Pragmatics would not only be a segment of language studies, but its main area of research.

In my dissertation (1992) I mention the Solomon's judgment case based on the discussion between Weissbourd and Mertz (1985). The authors, to address the issue of linguistic pragmatics and relevance of context, refer to the Old Testament. They tell that two women, who lived together, each had a baby. On the first night, after childbirth, one of the babies died because the mother had lain down upon him. This mother then swapped her dead child for the living one without letting the other mother notice and there was no witness to this fact. The birth mother of the child who lived woke at dawn to breastfeed her child, and then was surprised. The two mothers claimed the child before King Solomon, who then ordered the guard to cut the child in two with his sword. Then one of the women begged the king to keep the boy alive, as she let the other woman keep him. The other woman said: "Neither mine nor yours, let it be cut in two." The king ruled: "Give the living child to the first woman, do not kill it, because she is his real mother."

From that situation, the researchers understand that Solomon dismissed the conflict of the two women who requested custody of the child based on inferences and world knowledge, and not based on the semantic content of the speech of the first woman. The decision to grant custody of the child to the woman who begged the king to keep the boy alive was due to the observation of contextual data: the altruistic behavior of the real mother.

One of the conclusions of Weissbourd and Mertz (1985), while doing research in African cultures, is that our Western cultures built their justice systems as if we had created another world based on institutionalized facts, a world consisting of individuals, corporations and semiotic properties. The authors regret that the effective use of speech that emerges from situations is not a reality for us.

The meaning of language for lawyers and for linguists

The second topic is perhaps one of the biggest gaps between language and the law: the different conceptions of language, speech, text, among others, that the two domains of knowledge have. According to a study conducted in 2002, from a sample that consisted of ten legal hermeneutics treaties, the conception of language that permeates law manuals is that of the representation of the world, of an instrument, in which the words have a literal meaning (Colares, 2002: 207-249).

The notion that permeates Streck *et al.* (2009) takes as a reference Homer's literary text "The Odyssey" as a metaphor or analogy between the interpretation of the Constitution and the reaction of Ulysses to the song of the sirens. In the article entitled "Ulysses and the sirens: About activisms court and the dangers of Creating a" new constitutional convention "by the judicial power" the authors state that:

> Since Ulysses knew of the effect of the sirens' enchanting song, he tells his subordinates to chain him to the mast of the ship, and, under no circumstances, obey any order he may later give to unchain him. In other words, Ulysses knew he would not resist, and therefore created a self-restraint not to succumb later. (Streck *et al.*, 2009: 76)

In the 31 other occurrences of the word "text", the term is used to refer to the Constitution, as a finished product prepared by Brazilian constituents of 1988. In none of the uses does it refer to a Text Linguistics (knowledge domain with legitimacy to define what a text is) concept of text. Thus, to the authors "constitutions serve as the chains of Ulysses, and through them the politicians set some restrictions not to succumb to the despotism of the future majorities (parliamentary or monocratic)."

The Constitution is considered as an aid that may imprison the significance attributed by the "politicians", the legislators, in a timeless manner, "Always remaining identical to itself" regardless of who reads it, as if the significance was glued to the constitutional text as a label affixed by the legislature, and as if every magistrate had only to "decode" it. In the same article there are 20 occurrences of the word "interpretation". The authors admit the difficulties of finding a theory of legal interpretation in the context of this movement caused by the appearance of new constitutions that established the Democratic State. According to Barretto (1999):

> this moment enabled the creation and redefinition of a series of legal institutions, such as the so-called "general clauses", the "vague legal concepts", "abstract rules" and, of course, the so-called "constitutional principles".

There is a disproportionate preference for measurement and verification issues in the relation between words and ontological units (objects, states of affairs, events), the main focus is logical truth, rationality, in spite of the non-classical logics (logic of knowledge or epistemic logics, doxastic logics; tense logics, modal logics (concepts of necessity and possibility) which are attempts of the logicians to apply the inference systems to natural languages.

In law, the *eternal*, or the alleged neutrality of legal discourse in relation to political and social issues of the country is based on the idea that judges are mere *law-abiding individuals, trapped* – as Ulysses was chained – by the dictates of the rules elaborated by the legislative, being the legal discourse responsible for implementing the principle of justice, under the authority of the state. The conception of language as an instrument, with a literal meaning stiffened by legal dogmatics, permeates the illusion of neutrality. And, seeing as neutrality is just a myth, "the speech intended to be 'neutral', naive, also contains an ideology – that of their own objectivity" (Koch, 2004: 17).

Some Conclusions

Methodologically, the reactions to the turns: linguistic, pragmatic and hermeneutic reflect the fear of facing the temporariness of the notion of truth. The truth would not be, as in logics, the result of the correspondence between the theoretical propositions and the nature of the reality they describe, as in positivism, i.e., known truths regardless of context, in a social vacuum. We cannot establish a set of propositions, i.e., a reality in itself, in a conclusive way, by comparison with a reality that does not depend on these propositions. Propositions are formulated in ordinary language, in certain languages historically located, and consist of the results and consequences of what they state about reality.

Well, the emergent or urgent approximation between language and the law consists of reviewing the methodologies of the two sciences. The assertion that the linguistic pragmatics would make science unviable refers to the traditional realist conception of science as a conclusive and definitive knowledge of a reality considered in it.

Science and scientific theories are, unlike the realist conception, to be considered as more of a "language game" according to Wittgenstein, with their own rules, conventions and objectives. A theory is an explanatory model of reality, a set of hypothetical propositions that aim to explain a particular domain of reality. The pragmatic notions of effects and consequences are critical to the evaluation of scientific results and experiments, and for falsifying and validating scientific hypotheses.

As demonstrated, language is the primary medium through which Legal actions are performed. Reiterating what I have already written at other times: the conditions of the use of language include multiple aspects, simultaneous and successive, in the institutional context of justice, creating a "new object", and should extrapolate the mere linguistic analysis to construct an object of study of an interdisciplinary nature: language uses governed by legal principles. The task of this "new" field of knowledge Language and the Law will be to describe the nature of this semiosis, considering the need to interpret authentic legal texts in social contexts. The conclusion is that the bridge between language and law has not yet been built because the many attempts that have been made towards that direction focused on *abstract models of ideal languages* from the perspective of linguistics; and Law has built its point of view by developing a language concept based on the common sense of language teachers. All of this being governed by the paradigm of correction or based on classical logic and rhetoric.

References

Apel, K.-O. (1996). *Ethics and the Theory of Rationality.* Atlantic Highlands, N. J.: Humanities Press.

Austin, J. L. (1977). *Philosophical Papers.* Oxford: Clarendon Press, 3 edition.

Barretto, V. d. P. (1999). Da interpretação à hermenêutica constitucional. In M. LACOMBE, Ed., *1988-1998: uma década de Constituição*, 369–394. Rio de Janeiro: Renovar.

Carnap, R. (2002). *The Logical Syntax of Language.* London: Open Court.

Cavell, S. (1995). *Philosophical Passages: Wittgenstein, Emerson, Austin, Derrida.* Cambridge: Basol Blackwell.

Chomsky, N. (1975). *Aspectos da teoria da sintaxe.* Coimbra: Arménio Amado.

Colares, V. (1992). A decisão interpretativa da fala em depoimento judiciais. Dissertação (mestrado em lingüística), Programa de Pós-Graduação em Letras e Lingüística da UFPE, Recife.

Colares, V. (2002). Direito, produção de sentido e o "regime de liberdade condicional". *Revista da Pós-graduação em Direito da UNICAP*, 1, 207–249.

Dewey, J. (1971). *Experiência e Educação.* São Paulo: Companhia Editora Nacional.

Gadamer, H.-G. (2002). *Verdade e Método.* Petrópolis, RJ: Vozes.

Habermas, J. (1987). *Teoría de la acción comunicativa. Racionalidad de la acción y racionalización social*, volume Tomo I. Madrid: Taurus.

Habermas, J. (1990). *Pensamento pós-metafísico.* Rio de Janeiro: Tempo Brasileiro.

Habermas, J. (1994). ¿Qué significa pragmática universal? In *Teoría de la acción comunicativa: complementos y estúdios prévios.* Madrid: Cátedra.

Heidegger, M. (2003). *Ser y tiempo.* Madrid: Trotta.

Holmes, O. W. (1995). *The Collected Works of Justice Holmes.* Chicago: University of Chicago Press.

Hutton, C. (1996). Law lessons for linguits? Accountability and acts of professional classification. *Journal of Language & Communication: an interdisciplinary journal,* 16(3), 205–227.

James, W. (1997). *Pragmatism: A New Name for Some Old Ways of Thinking / Four Related Essays Selected from "The Meaning of Truth".* New York/London/Toronto: Longmans/ Gren.

Koch, I. V. G. V. (2004). *A inter-ação pela linguagem.* São Paulo: Cortez, 9 edition.

Morris, W. (1979). *Logical Positivism, Pragmatism and Scientific Empiricism.* Paris, New York: Hermann, AMS Press.

Peirce, C. S. (1977). O que é pragmatismo. In *Semiótica.* São Paulo: Perspectiva.

Rorty, R. (1982). *Consequences of Pragmatism.* Minneapolis: University of Minnesota Press.

Ryle, G. (2002). *The Concept of Mind.* Chicago: University of Chicago Press.

Searle, J. (1969). *Speech Acts.* Cambridge: Cambridge University Press.

Streck, L. L., Barretto, V. d. P. and Oliveira, R. T. d. (2009). Ulisses e o canto das sereias: sobre ativismos judiciais e os perigos da instauração de um "terceiro turno da consti-tuinte". *Revista de Estudos Constitucionais, Hermenêutica e Teoria do Direito (RECHTD),* 1(2), 75–83.

Weissbourd, B. and Mertz, E. (1985). Rule-centrisme versus legal creativity: the skewing of legal ideology through language. *Law & Society Review,* 19(4), 613–759.

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford, New York: Basil Blacwell, Macmillan.