Journal of Law and Policy

Volume 21
Issue 2
SYMPOSIUM:
Authorship Attribution Workshop

Article 10

2013

Towards an Index of Idiolectal Similitude (Or Distance) In Forensic Authorship Analysis

M. Teresa Turell, Ph.D.

Nuria Gavalda

TOWARDS AN INDEX OF IDIOLECTAL SIMILITUDE (OR DISTANCE) IN FORENSIC AUTHORSHIP ANALYSIS

M. Teresa Turell and Núria Gavaldà**

I. INTRODUCTION

Forensic linguistics is a discipline concerned with the study of language in any judicial context. The framework for the present article is the area of forensic linguistics known as Language as Evidence, where a sample or several samples of oral or written linguistic productions of one or more individuals may constitute evidence in a judicial process. In these cases, linguists acting as expert witnesses in court must compare two (sets of) samples, i.e., the nondisputed sample, the authorship of which cannot be questioned, and the disputed sample, the authorship of which is questioned, to determine the linguistic differences and similarities that the samples show and to try to reach a conclusion regarding the possibility that they have been produced by the same individual.

Linguistic evidence is not like other kinds of evidence such as DNA or fingerprints, in the sense that language is intrinsically variable. Sociolinguists have shown for decades that languages are in a state of constant change and that any language is intrinsically variable in all its levels, even at the idiolectal level.[1] In other words, the linguistic production of a single

* ForensicLab, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (Barcelona, Spain).

[1] *See, e.g.*, WILLIAM LABOV, SOCIOLINGUISTIC PATTERNS 122, 127, 271–72, 319–25 (1972); *see also* J.K. CHAMBERS, SOCIOLINGUISTIC THEORY: LINGUISTIC VARIATION AND ITS SOCIAL SIGNIFICANCE 33–37 (2009); M. Teresa Turell Julià, *La base teòrica i metodològica de la variació lingüística*, *in* LA SOCIOLINGÜÍSTICA DE LA VARIACIÓ 17, 20–22 (M. Teresa Turell ed.,

speaker or writer will generally show some variation. Consequently, when comparing two samples, the expert witness must ponder whether the degree of variation present is likely to be due to interspeaker/writer differences or to intraspeaker/writer differences. To do this, the linguist must analyze as many linguistic parameters as possible in order to reliably reach such conclusions.

Research in the last forty years has successfully identified parameters that can contribute to this endeavor. In the field of forensic speech comparison, where oral samples (recordings) are analyzed, both acoustic and linguistic parameters are normally considered. On the one hand, phoneticians analyze the acoustic nature of individual sounds (vowels and consonants) together with parameters related to the fundamental frequency (related to the pitch of the voice), voice quality, and suprasegmental patterns such as intonation or linguistic rhythm.[2] On the other hand, phonological variables are related to individual choices that each individual makes depending on their place of origin and other social factors such as gender, education, and class.[3] Moreover, variables related to the particular syntactic, morphological, or lexical patterns that an individual shows can also shed light on the differences or similarities between oral samples. In the field of forensic text comparison, or authorship analysis, where written texts are analyzed, variables related to lexical density, lexical richness, and syntactic and morphological patterns have been proven to be reliable markers of authorship.[4]

1995).

[2] *See, e.g.*, Peter French, *An Overview of Forensic Phonetics with Particular Reference to Speaker Identification*, 1 FORENSIC LINGUISTICS 169, 174–76, 178 (1994); *see also* Erika Gold & Peter French, *International Practices in Forensic Speaker Comparison*, 18 INT'L J. SPEECH LANGUAGE & L. 293, 295–96 (2011).

[3] *See, e.g.*, Paul Foulkes & Peter French, *Forensic Phonetics and Sociolinguistics*, *in* CONCISE ENCYCLOPEDIA OF SOCIOLINGUISTICS 329, 330 (Rajend Mesthrie ed., 2001).

[4] *See, e.g.*, David Woolls & Malcolm Coulthard, *Tools for the Trade*, 5 INT'L J. SPEECH LANGUAGE & L. 33, 37 (1998); *see also* Harald Baayen et al., *Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution*, 11 LITERARY & LINGUISTIC COMPUTING 121, 128 (1996); M. Teresa Turell, *Textual Kidnapping Revisited: The Case of*

Also, other features related to the deep structure of language, such as the analysis of parts of speech via *n*-grams,[5] have also been shown to account for idiosyncratic characteristics.

This article proposes an Index of Idiolectal Similitude (or Distance) (hereinafter IIS) as a new tool to carry out forensic speech and text comparison.[6] Part II provides some of the premises and hypotheses underlying the study of forensic linguistics. Part III contains an overview of the study, including descriptions of its objectives, theoretical framework, hypotheses, and methodology. Finally, Part IV presents the result of the study and is followed by an assessment of the results and discussion on the future of the study.

II. PREMISES AND HYPOTHESES

The study of idiolectal similitude or distance is based on two fundamental premises: 1) language provides oral and written

---

Plagiarism in Literary Translation, 11 INT'L J. SPEECH LANGUAGE & L. 1, 19–20, 24 (2004).

[5] *N*-grams are sequences of grammatical categories. For example, "the man" is a bigram (sequence of two grammatical categories (article + noun)) and "the man is" is a trigram (sequence of three parts of speech (article + noun + verb)). *See, e.g.*, Maria S. Spassova & M. Teresa Turell, *The Use of Morpho-syntactically Annotated Tag Sequences as Forensic Markers of Authorship Attribution*, PROCEEDINGS OF THE SECOND EUROPEAN IAFL CONFERENCE ON FORENSIC LINGUISTICS / LANGUAGE AND THE LAW 229, 229–37 (2007); *see also* Maria Stefanova Spassova, El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español 59–63 (2009) (unpublished Ph.D. dissertation, Universitat Pompeu Fabra), *available at* http://tesisenred.net/bitstream/handle/10803/7512/tmss.pdf.pdf?sequence=1.

[6] The research presented in this article is based on the findings of two research projects, *Idiolectometría aplicada a la lingüística forense*, funded by the Spanish Ministry of Science and Education (EXPLORA-HUM2007-29140-E; PI: M. Teresa Turell, 2007–08), and the FFI project, *Idiolectometría forense e Índice de similitud idiolectal*, funded by the Spanish Ministry of Science and Innovation (FII2008-03583/FILO; PI: M. Teresa Turell, 2008–11). *See generally* FORENSICLAB—UNITAT DE VARIACIÓ LINGÜÍSTICA, FORENSIC IDIOLECTOMETRY AND INDEX OF IDIOLECTAL SIMILITUDE (2013), http://www.iula.upf.edu/rec/forensic_isi/docums/forensic_isi_en.pdf.

information of several kinds and can reveal an individual's socio-individual and socio-collective traits; and 2) each individual seems to have a unique idiosyncratic use of language that distinguishes him or her from the rest of language users in his or her community. This individual use of language has traditionally been referred to by forensic linguists as "idiolect."[7] This article follows the more recent concept of "idiolectal style" proposed by Turell, which is defined as follows:

> [The] concept "idiolectal style," following the use of the term "style" in pragmatics, is proposed as a notion which could be more relevant to forensic authorship contexts. "Idiolectal style" would have to do primarily, not with what system of language/dialect an individual has, but with a) how this system, shared by lots of people, is used in a distinctive way by a particular individual; b) the speaker/writer's production, which appears to be "individual" and "unique" (Coulthard 2004)[8] and also c) Halliday's (1989) proposal of "options" and "selections"[9] from these options.[10]

Regarding forensic authorship analysis, there have been some recent objections to current work, in particular with approaches involving qualitative analyses of the data. These objections deal with the fact that qualitative approaches may be considered nonscientific and subjective, that they are rarely testable, and that their rate of error has never been established.[11]

---

[7] *See* J.R. Baldwin, *Phonetics and Speaker Identification*, 19 MED. SCI. & L. 231, 231 (1979); *see also* GERALD R. MCMENAMIN, FORENSIC LINGUISTICS: ADVANCES IN FORENSIC STYLISTICS 53–54, 112 (2002); Malcom Coulthard, *Author Identification, Idiolect, and Linguistic Uniqueness*, 25 APPLIED LINGUISTICS 431, 431 (2004).

[8] Coulthard, *supra* note 7, at 445.

[9] M.A.K. HALLIDAY & RUQAIYA HASAN, LANGUAGE, CONTEXT AND TEXT: ASPECTS OF LANGUAGE IN A SOCIAL-SEMIOTIC PERSPECTIVE 55–56, 113–15 (1989).

[10] M. Teresa Turell, *The Use of Textual, Grammatical and Sociolinguistic Evidence in Forensic Text Comparison*, 17 INT'L J. SPEECH LANGUAGE & L. 211, 217 (2010).

[11] *See, e.g.*, Carole E. Chaski, *Empirical Evaluations of Language-Based Author Identification Techniques*, 8 FORENSIC LINGUISTICS 1, 2 (2001); *see*

In this sense, if we compare this area with other forensic linguistic sciences, such as forensic phonetics and acoustics, forensic authorship analysis does not count on a common framework regarding the definition of the nature, number, and size of the samples to be used before one can attribute authorship safely. Moreover, it is also necessary to agree on what comparison baseline is needed before one can achieve degrees of reliability. Thus, there is a general need in all languages, as well as in all operational areas of Language as Evidence, to be able to count on corpora consisting of all possible existing spoken or written idiolectal styles of each speaker or writer, even if this is a daunting, almost impossible, endeavor.

Meanwhile, forensic authorship analysis can benefit from a complementary combination of both qualitative and quantitative methods.[12] In other words, until the Likelihood Ratio framework[13] for written texts can be adopted in forensic authorship analysis, among other quantitative methods, different approaches that complement each other—i.e., cumulative evidence—will have to be used in the comparison of disputed and nondisputed texts. Studies have shown that there are several techniques that can be used in forensic authorship analysis,

---

also Tim Grant & Kevin Baker, *Identifying Reliable, Valid Markers of Authorship: A Response to Chaski*, 8 FORENSIC LINGUISTICS 66, 68–76 (2001).

[12] *See* Turell, *supra* note 10, at 218, 220.

[13] The Bayesian likelihood ratio represents the framework within which other forensic sciences such as analysis of DNA are being developed. This statistical method calculates the probability of the evidence considering the hypotheses given by both the defense and the prosecution. However, one of the most important limitations by which this method cannot be used in present-day authorship analysis is that it needs a Base Rate Knowledge of population distribution in order to make decisions regarding how significant certain differences and similarities between linguistic samples are, which is only available for very limited linguistic features. This Base Rate Knowledge implies the collection of data regarding the general usage of the linguistic parameters being considered by a relevant population, or group of language users from the same linguistic community, with which the specific behavior of the speakers or writers under comparison can be compared.

including textual qualitative analytical procedures,[14] the analysis of lexical density and lexical richness,[15] and the use of reference corpora to account for the rarity of linguistic variables.[16] Furthermore, the use of semiautomatic analyses of "deep-structure" linguistic variables (such as Discriminant Function Analysis of sequences of annotated linguistic categories) has also proved to be a reliable technique.[17] Finally, the measurements of idiolectal similitude/distance such as those involved in the use of the IIS proposed here may also be a good approach to carry out forensic authorship analysis.

---

[14] *See, e.g.*, Ol'ga Feiguina & Graeme Hirst, *Authorship Attribution for Small Texts: Literary and Forensic Experiments*, PROC. SIGIR'07 INT'L WORKSHOP ON PLAGIARISM ANALYSIS, AUTHORSHIP IDENTIFICATION, & NEAR-DUPLICATE DETECTION, 2007, at 236, 236–39; David I. Holmes, *Authorship Attribution*, 28 COMPUTERS & HUMAN. 87, 87–106 (1994); Spassova & Turell, *supra* note 5, at 229–37; Hans van Halteren et al., *Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution*, 11 LITERARY & LINGUISTIC COMPUTING 18, 18–24 (1996).

[15] *See, e.g.*, Woolls & Coulthard, *supra* note 4, at 37–38 (describing a method of authorship identification that focuses on lexical richness, average sentence length, and grammar); *see also* Coulthard, *supra* note 7, at 435 (discussing the value of measuring the percentage of lexical types in detecting plagiarism); Turell, *supra* note 4, at 24 (summarizing findings measuring uniqueness of used terms and phrases by measuring density); M. Teresa Turell, The Disputed Authorship of Electronic Mail: Linguistic, Stylistic and Pragmatic Markers in Short Texts (2004) (unpublished conference paper).

[16] *See, e.g.*, Malcom Coulthard, *On the Use of Corpora in the Analysis of Forensic Texts*, 1 FORENSIC LINGUISTICS 25, 28–29 (1994) (explaining how corpora may be used to, for example, determine how likely it is for a word to occur, both individually and with other words); *see also* Turell, *supra* note 10, at 216, 218 (describing linguistic variables and their influence on forensic text comparison).

[17] *See, e.g.*, Spassova, *supra* note 5; *see also* Núria Bel et al., *The Use of Sequences of Linguistic Categories in Forensic Written Text Comparison Revisited*, PROC. INT'L ASS'N FORENSIC LINGUISTS' TENTH BIENNAL CONF., 2012, at 192, 192–93, 197–98, 200, *available at* http://www.forensic linguistics.net/iafl-10-proceedings.pdf (reporting positive findings through the use of qualitative and semi-automatic and quantitative approaches, based on various analyses, including Discriminant Function Analysis); Feiguina & Hirst, *supra* note 14.

III. THE STUDY

A. Main Objectives

This article presents a study that explores and develops the possibility of measuring the linguistic differences existing between idiolectal styles and each individual's idiolectal similitude or distance, with the aim of establishing an IIS which will compare several linguistic samples and calculate the linguistic distance between them. The main objective of the establishment of the IIS is to a) create a technique that allows researchers to compare several linguistic samples in terms of the variables that the protocol contemplates, b) calculate the linguistic similitude or distance between them, and c) determine what kind of idiolectal similitude is needed in order to say as definitively as possible that two linguistic samples have been

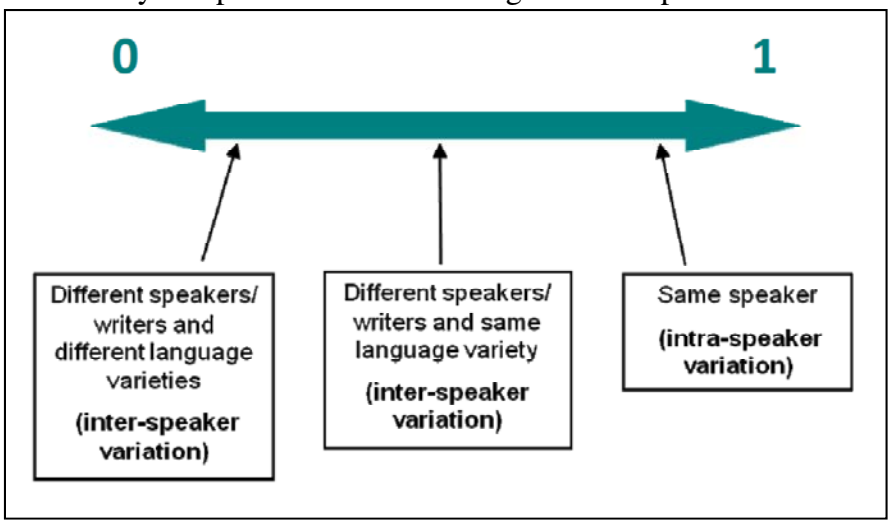

Figure 1: Representation of the IIS as a continuum

produced, or not, by the same individual. The final aim of this project is to be able to apply the IIS methodology to real forensic cases, where instead of comparing two samples from real world data, i.e., where we know who the authors or speakers are, one would compare one disputed and one nondisputed sample or several disputed and nondisputed sample sets.

The IIS is conceived as a continuum (see Figure 1) between 0 and 1, where 0 indicates maximum difference and 1 indicates minimum difference. According to this concept, when two (sets of) linguistic samples, either oral or written, are compared, and the IIS is applied, a result closer to 0 indicates that the two samples under comparison were produced by different individuals and that these samples exhibit interspeaker/writer variation. A value at an intermediate position along the continuum indicates that there is also interspeaker/writer variation, but the slight increase in similarity may indicate that the two individuals share the same linguistic variety. Finally, a value close to 1 would mean that there exists an expected intraspeaker/writer variation but would lead the expert to conclude that the two samples are so similar that they could have been produced by the same individual.

B. Theoretical Framework

The theoretical framework behind the IIS proposal draws from the Theory of Language Variation and Change ("TLVC") developed by William Labov during the 1960s. The TLVC maintains that language is in a state of constant change and that changes in language can be perceived synchronically by means of variation present at all levels of language. In this sense, linguistic variation was demonstrated not to be random, as previous theories of language had maintained, but proved to be systematic and patterned. This correlates to internal linguistic characteristics such as the particular phonetic context in which a specific sound appears and also external social factors such as gender, age, social class, and level of income.[18] Labov's theory

[18] *See, e.g.*, WILLIAM LABOV, SOCIOLINGUISTIC PATTERNS 111, 120–21,

is, according to Turell, "theory building" in terms of three main dimensions: first, in terms of the basic aim stated, which is to describe linguistic variation and change; second, regarding the data it analyzes, which is an individual's most spontaneous variety, (that is, his or her vernacular); and third, as regards the methods it applies in order to measure this variation, namely observation, description, and explanation.[19] The TLVC studies both individual and group (speech community) variation.[20] This individual–speech community binomial has proved to be very useful, not only in studies of linguistic variation but also in other areas of applied linguistics such as the linguistic profiling aspects of forensic linguist expert witness work. For the purposes of further applications of the IIS to real forensic data, one relevant issue drawn from this theory is the exploration of single dimensions of variation through the binary division of linguistic internal factors, and when relevant, of social factors as well.[21] Also of relevance are the use of multivariate analyses to show the simultaneous effect of all relevant independent variables and the use of cross-tabulation to give a more refined view of the distribution of the data and the degree of independence of intersecting variables.[22]

---

161 (1972) (providing an overview of factors impacting linguistic variation); *see also* 1 WILLIAM LABOV, PRINCIPLES OF LINGUISTIC CHANGE: INTERNAL FACTORS 5 (1994) ("To explain a finding about linguistic change will mean to find its causes in a domain outside of linguistics . . . ."); 2 WILLIAM LABOV, PRINCIPLES OF LINGUISTIC CHANGE: SOCIAL FACTORS 74–75 (2001) (distinguishing between former and current approaches to assessing variation).

[19] *See* M. Teresa Turell, William Labov Laudatio, Universitat Pompeu Fabra (June 15, 2012), *available at* http://www.upf.edu/enoticies/1112/_pdf/laudation_turell_angles_.pdf.

[20] *See, e.g.*, WILLIAM LABOV ET AL., ATLAS OF NORTH AMERICAN ENGLISH 69, 157, 285, 303 (2006).

[21] *See, e.g.*, LABOV, *supra* note 1, at 110–121, 160–182 (examining the relationship of sociology and linguistic variations). *See generally* 1 LABOV, *supra* note 18 (discussing the internal factors affecting linguistic variation); 2 LABOV, *supra* note 18 (noting the role of socioeconomics on changes in linguistics).

[22] *See, e.g.*, LABOV, *supra* note 1, at 7–8, 11, 41, 72, 108, 226 n.30 (presenting studies of linguistic variables and the sociolinguistic characteristics these variables reveal); *see also* WILLIAM LABOV, WHAT IS A

In addition to this, drawing from what is now known as forensic sociolinguistics, it can be stated that the linguistic production of an individual can provide clues regarding social factors such as their age, gender, occupation, education, religion, political background, their geographical origin, their ethnicity or race,[23] their nonnativeness when using a second or foreign language, and a variety of language reflecting markers of language contact.[24]

C. Hypotheses

The working hypotheses to be tested through the analysis of the observed linguistic parameters and variables are the following:

1. Interspeaker/writer variation will be higher than intraspeaker/writer variation. In this sense, IIS results obtained when comparing samples from the same speaker or writer should be closer to 1 than those obtained when comparing samples from different individuals.

2. Despite the existing intraspeaker/writer variation, an individual's idiolectal style will be quite stable throughout time. Consequently, IIS results should be close to 1 when comparing two samples from the same individual from different measurement times.

3. An individual's idiolectal style will also remain relatively stable despite the use of different genres or textual registers but possibly not as stable as it might be throughout time. Therefore, when comparing samples from the same individual involving

---

LINGUISTIC FACT? 12 (1975) (noting the need for improvement in linguistic data methodology as well as the scope of linguistic variation).

[23] *See* Sharon S. Smith & Roger W. Shuy, *Forensic Psycholinguistics: Using Language Analysis for Identifying and Assessing Offenders*, FBI L. ENFORCEMENT BULL., Apr. 2002, at 16–21, *available at* http://diogenesllc.com/statementlinguistics.pdf (noting the ability of language to reveal characteristics of the speaker).

[24] Turell, *supra* note 10, at 220–25 (noting the ability to use linguistic production to identify users from different geographical regions and users whose first language is not Spanish).

different genres, IIS should also be close to 1 (but perhaps not as close as results in hypothesis 2).

D. Methodology

The analysis of idiolectal distance that is presented here is based on research carried out in two projects.[25] Each project involved several stages where different numbers of subjects and methods were analyzed. This article presents final results obtained in the last stage, in which six individuals were studied per each module, and a final list of variables, ranging between 10 and 18 depending on the module, were selected after some preliminary studies where some other variables were discarded. Moreover, a total of four different methods were explored, but only three were involved in the final stage. The remaining method, which was based on the Euclidean distance, was finally discarded, and it is not included in this account.

1. Linguistic Modules and Variables

The protocol devised to calculate the IIS has explored, so far, three different linguistic levels, or modules: the phonological module, the morphosyntactic module, and the discourse-pragmatic module. The phonological module involves the analysis of phonological processes related to insertion, elision, or change of sounds, such as yod-coalescence in English (a process by which a word like *duke* can be pronounced [dju:k] or [dʒu:k]). The morphosyntactic module considers variables related to morphological and syntactic patterns, such as the presence or absence of the conjunction *that* in a sentence like *I thought (that) it was nice*. Finally, the discourse-pragmatic module considers discursive and pragmatic phenomena, such as the choice of the intensifier *really* in contrast with other intensifiers such as *absolutely* or *completely*, as in *I was really/absolutely/completely terrified*.

---

[25] *See supra* note 6.

Regarding the variables, the IIS is concerned with discrete variables[26] that occur in the idiolectal style of the two speakers or writers under analysis, and they all show variation, which is structured in two main variants, either variant A or B, or the presence or lack of the process, following the most standard formulations of linguistic variation analysis.[27] For example, the variable that deals with the process of yod-coalescence explained above contemplates two variants: 1) the presence of the process, by which all instances where yod-coalescence occurs are calculated and 2) the lack of process, by which all the instances where yod-coalescence could occur but does not, are calculated.

2. Corpora

Different corpora have been used to test the formulated hypotheses, and all, in one way or another, have involved the elicitation of semispontaneous speech,[28] except for the morphosyntactic module of Spanish, which was analyzed by using a written corpus. Moreover, all the corpora (except that of the discourse-pragmatic module of Spanish) contain data from the same adult men and women collected in two measurement times ("MT1" and "MT2," respectively) with a lapse of ten to twenty years depending on the module, in order to investigate the subjects' idiolectal style throughout time.

The corpus of study for the Catalan modules contains data on Eastern Catalan and consists of sociolinguistic interviews recorded in La Canonja, a Catalan speech community in the

---

[26] In statistics, variables may be a) discrete, meaning that they take a limited number of values, such as gender (either male or female) or social class; and b) continuous, which implies any value within a range of values on a scale, such as age, for example.

[27] *See, e.g.*, LABOV, *supra* note 1, at 192–93; WILLIAM LABOV, THE SOCIAL STRATIFICATION OF ENGLISH IN NEW YORK CITY 31 (2d ed. 2006).

[28] Semispontaneous speech implies the speech resulting from an interview, where the electronic equipment such as microphones or cameras may make the speaker aware of the situation and inhibit them from using completely spontaneous speech, or their *vernacular*, as it is referred to in sociolinguistics.

Tarragona area in a real-time, Labovian study.[29] In the Spanish modules, several corpora were used: the Mexican Spanish HETA corpus[30] was used to analyze the phonological module, the written Peninsular Spanish corpus[31] was used to analyze the morphosyntactic module and, finally, the Peninsular Spanish corpus,[32] only available for MT1, was used to analyze the discourse-pragmatic module. Regarding the English modules, a corpus containing data on Southern British English in MT1 and MT2 was compiled by means of radio and TV interviews, and the subjects are world-known artists, whose recordings are available online.

3. Methods

The three phonological modules in Catalan, Spanish, and English were analyzed following the auditory-acoustic approach,[33] and the three morphosyntactic and discourse-

---

[29] Oral corpus of La Canonja (1987–92), compiled by Juan José Pujadas, Mercè Pujol, and M. Teresa Turell, through 2 CICYT research projects (PBS90-0580 and SEC93-0725).

[30] Fernanda López, El análisis de las características dinámicas de la señal de habla como posible marca para la comparación e identificación forense de voz: Un estudio para el español de México (2010) (unpublished Ph.D. dissertation, Universitat Pompeu Fabra), *available at* http://www.tdx.cat/bitstream/handle/10803/42940/tfle.pdf;jsessionid=AA3F9AC40961A1652DA3E5E543E32BD9.tdx2?sequence=1.

[31] Maria S. Spassova, Las marcas sintácticas de atribución forense de autoría de textos escritos en español (May 2006) (unpublished PhD dissertation, Universitat Pompeu Fabra).

[32] PRESEEA, http://preseea.linguas.net/ (last visited Feb. 6, 2013).

[33] The auditory-acoustic approach to forensic phonetics is the combination of two main approaches. On the one hand, in order to carry out an auditory analysis, phoneticians make use of their knowledge about general phonetics and phonology and the phonetics and phonology of the linguistic system at hand for the interpretation of the acoustic samples being analyzed. On the other hand, an acoustic analysis involves the use of specially developed techniques—normally involving specialized computer software aimed at the acoustic analysis of speech—together with the phonetician's knowledge of physics and the acoustic properties of the speech signal, especially those characteristics most relevant to the language under analysis. For further information, see Francis Nolan, *Speaker Recognition and*

pragmatic modules in these same languages were coded for the different linguistic variables that had been located by their discreteness. Method 1 involves the calculation of an average of the difference in the percentage of occurrences of each variant. On the other hand, method 2 is based on the Adjusted Residual Value ("ARV") obtained after running cross-tabulations, which is a number indicating the difference in the distribution of the variables in the samples compared. Finally, method 3 is based on the Phi Coefficient, which is a coefficient that ranges from 0 to 1 and provides an indication of the strength of the relationship between the variables considered.

IV. RESULTS

The results obtained by using the three methods were very similar. However, method 3, which is based on the Phi Coefficient, proved better at accounting for intra- and interspeaker/writer results.

Regarding the phonological modules, hypothesis 1, which stated that intraspeaker results would be higher in the IIS continuum than interspeaker results, is confirmed by all three methods in all three languages. In this article, only results from method 3 will be shown and discussed for all the modules. Figure 2 shows interspeaker IIS results with method 3, where each point in the graph corresponds to an IIS value after comparing samples from two different speakers. Results show that all interspeaker IIS values are relatively low in general (between 0.2 and 0.8), which is an expected result considering that, except for the Catalan corpus, all speakers belong to the same dialectal area. Method 3 has proved useful in the case of the phonological module of Catalan in order to observe that when the IIS is calculated between speakers of different varieties, the interspeaker IIS values are lower than when the speakers compared belong to the same dialectal area, a result

---

Forensic Phonetics, *in* THE HANDBOOK OF PHONETIC SCIENCES 744, 744–67 (William Hardcastle & John Laver eds., 1994); French, *supra* note 2, at 295–96.

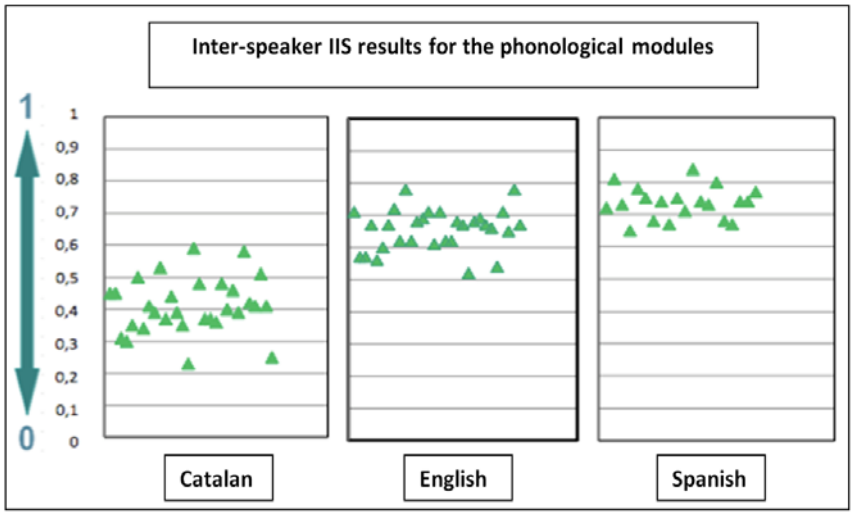

Figure 2: Interspeaker IIS results for the phonological modules

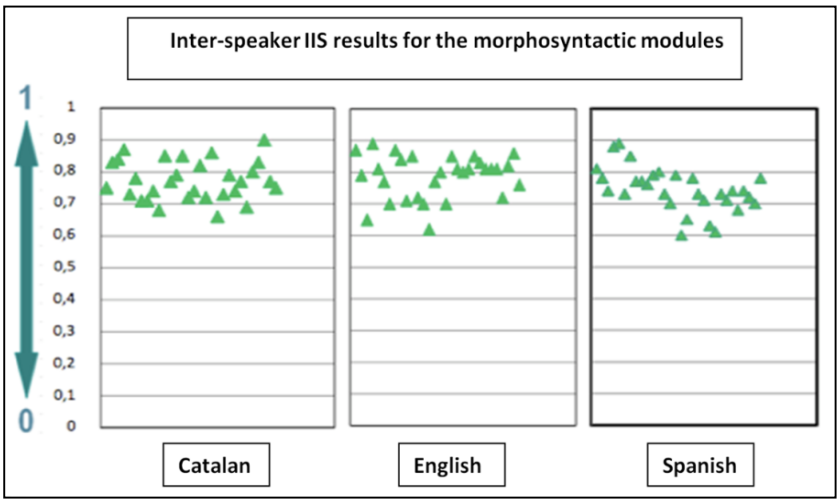

Figure 3: Interspeaker IIS results for the morphosyntactic modules

which is very relevant in real forensic cases concerned with linguistic profiling.

Hypothesis 1 is also confirmed by all 3 methods for the morphosyntactic modules (Figure 3) and the discourse-pragmatic modules (Figure 4). In both modules in the three languages, all interspeaker/writer IIS values are relatively low in general (they range between 0.6 and 0.8), which is an expected result considering that all the subjects belong to the same dialectal area.

Hypothesis 2 stated that an individual's idiolectal style would stay relatively stable despite the course of time. In order to

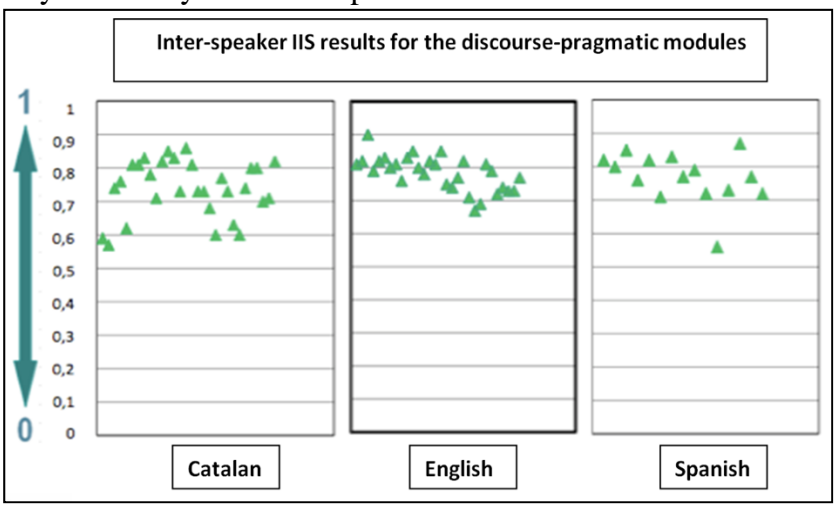

Figure 4: Interspeaker IIS results for the discourse-pragmatic modules.

confirm this hypothesis, samples from the same individual in MT1 and MT2 were compared with each other. In Figures 5–7, points in the graph indicate an intraspeaker/writer IIS result, i.e., an IIS value after comparing samples from the same subjects in two separate points in their lives.

Results show that this second hypothesis is confirmed for both the phonological and the morphosyntactic modules. Figures 5 and 6 illustrate results in these two modules for the three languages. As can be seen, IIS results for all the modules range between 0.8 and 0.9, which is high, as expected, since 1 on the IIS continuum means maximum similarity.

With regard to the discourse-pragmatic modules, hypothesis 2 could only be tested for the Catalan and English modules, since the Spanish corpus for this module did not contain data in two measurement times. Hypothesis 2 is also confirmed with all three methods of Catalan and English. With method 3 (Figure 7), all IIS values are quite high, as expected, with the majority ranging between 0.9 and 0.7.

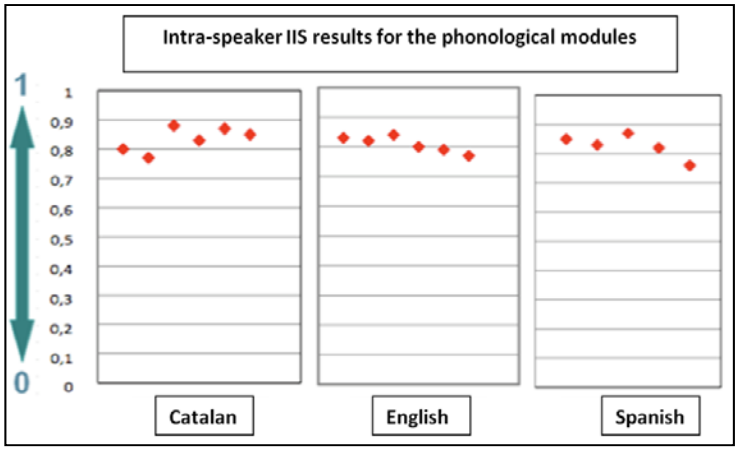

Figure 5: Intraspeaker IIS results for the phonological modules.

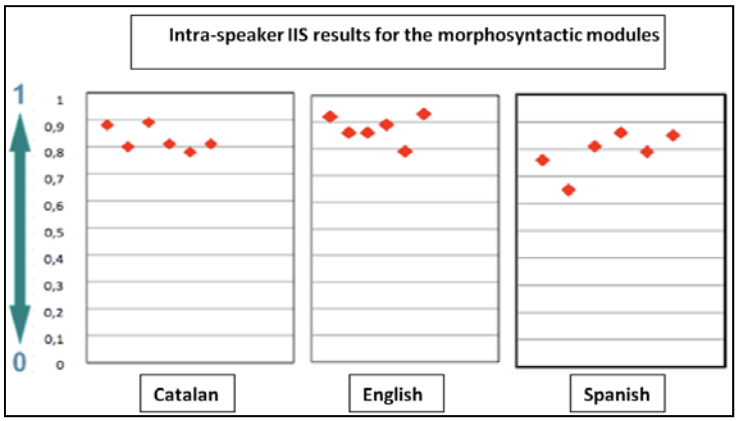

Figure 6: Intraspeaker IIS results for the morphosyntactic modules.

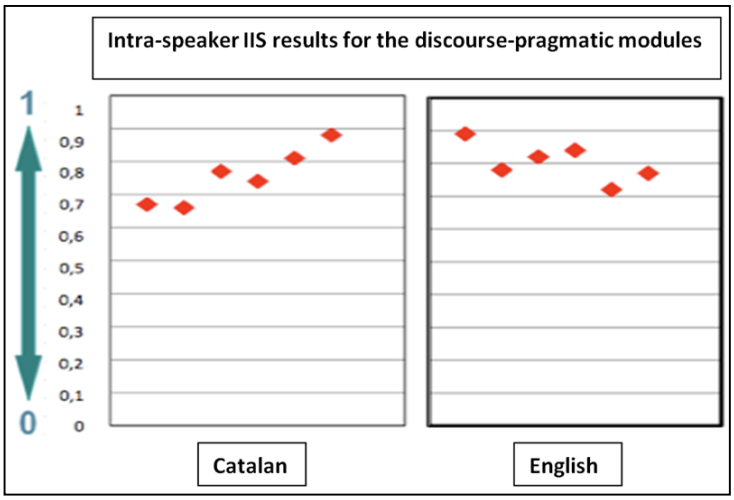

Figure 7: Intraspeaker IIS results for the discourse-pragmatic modules.

CONCLUSIONS

The first conclusion that we can draw from our results, which has already been noted above,[34] is that method 3, based on the Phi Coefficient, turned out to be the most reliable method, in the sense that it triggered the most robust results for both intra- and interspeaker/writer variation, and in particular in the phonological modules, although with some exceptions. Moreover, hypothesis 1 is confirmed for all modules and languages in that there seems to be more variation, and thus more idiolectal distance, between different individuals than between two samples of the same individual. Also, hypothesis 2 is also confirmed in that samples from the same individual at two measurement times seem to show pretty stable patterns, which would seem to confirm that an individual's idiolectal style (spoken or written) does not appear to vary much throughout time.

If we look more closely into interspeaker/writer IIS results, some IIS values seem to be too high, or at least higher than expected, especially for the morphosyntactic and the discourse-pragmatic modules. In this sense, it should be borne in mind that, except for the phonological module of Catalan, all the subjects considered belong to the same language variety; therefore, high results placed at a middle point along the IIS continuum were expected. However, it is true that in some cases, the IIS methodology does show unexpected results in that some of these interspeaker/writer values are certainly as high as intraspeaker/writer results. We believe that these unexpected results have to do with certain methodological difficulties that we encountered in the process of our research. First, the sample stratification regarding genre, time, language variety, and gender might have had some influence. Not all corpora were stratified for different genres (and at the same time, for different measurement times), and so, for the time being, it has not been possible to test hypothesis 3, which stated that an individual's idiolectal style should be quite stable in spite of the use of different genres. This hypothesis will be explored in the future.

---

[34] *See supra* Part IV.

Regarding time, the phonological module of Spanish only had five speakers in MT2, whereas the discourse-pragmatic module of Spanish contained data in MT1 for all the speakers. As for language variety (or dialect), even if it was not formulated as a hypothesis, the analysis of the phonological module of Catalan, stratified with speakers from two dialects, has proven very robust in its ability to account for interspeaker variation, so it would be desirable to be able to count on all the other modules stratified by language variety. Finally, as regards gender, for the IIS itself and also in order to contribute to the Base Rate Knowledge of population distribution, it would be interesting to test whether there is more interspeaker/writer variation when all speakers are considered together or when a distinction is made in the comparison between female and male speakers or writers.

Another difficulty for comparative purposes—naturally not exclusively related to the IIS measure but which could affect the internal validity of results—has to do with the nature of the variables, namely the different nature that morphosyntactic and discourse-pragmatic variables have in comparison with phonological variables. On the one hand, morphosyntactic and discourse-pragmatic variables have a lower frequency than phonological variables, which could affect final results. On the other hand, the discreteness of morphosyntactic and discourse-pragmatic variables (i.e., their capacity for being formulated as discrete variables with two variants) is much more difficult to establish than that of phonological variables.

Furthermore, it is also possible that the nonparallel nature of the corpora under analysis may have had an effect on the final results. Only in the case of the English (internet TV/radio samples) and the Catalan (La Canonja) IIS calculation, the same corpus was used to analyze the three modules under investigation, while the three linguistic modules of Spanish each contemplated different corpora.

Robust results seem to be associated with the choice of the variables, the establishment of their discreteness, and the number of variables. The more variables, the better IIS results seem to be. The robustness of the IIS will be better grasped when other relevant results are tabulated (for example, when pattern similarity in all modules for each pair of speakers or

writers compared is applied). In other words, for two samples to be attributed to the same speaker or writer, the IIS values must all be near 1 in all modules; for two samples to be attributed to different speakers or writers, IIS values must all be between 0.7 and 0.5 (same speech variety) or between 0.5 and 0.3 (different speech varieties).

The disparity of results obtained in some of the IIS values has had a direct effect on the design of further experiments and on future data collection. Future research will focus on increasing the number of languages as objects of analysis (e.g., Arabic), the sample size (i.e., more subjects for each language), and also on the stratification of the corpora by genre in order to confirm hypothesis 3. Additionally, other indicators such as gender, age, or educational level will be examined to contribute to the Base Rate Knowledge of population distribution.

In conclusion, the IIS measure can provide reliability to the concept of idiolectal similitude or distance, and once the protocol for its calculation is consolidated, the IIS measure may be successfully complemented with other approaches to forensic speech and text comparison to be used in real forensic cases. In addition to this, research towards the establishment of the IIS measure can also provide forensic linguistics with a Base Rate Knowledge of population distribution as regards several linguistic variables for the three modules and the three languages under study, which is a fundamental issue in current forensic linguistic work.