Forensic Plagiarism Detection and Authorship Attribution: on the linguists' achievements and the challenges for computerized analysis

M. Teresa Turell
ForensicLab - IULA
Universitat Pompeu Fabra

Malcolm Coulthard
Centre for Forensic Linguistics
Aston University

Session development

- **Part I: The forensic linguist's achievements**
  - Conceptual and methodological context of real forensic cases
- **Part 2: Examples of Linguists' Achievements**
  - Plagiarism
  - Authorship
- **Part 3: Challenges for computerized analysis**
  - Possible collaboration of forensic and computational researchers

Part 1

The forensic linguist's achievements

The conceptual and methodological context of real forensic linguistic cases

Forensic Linguistics

- Language of the Law
- Language of the Court
- **Language as Evidence**

Language as Evidence
The expert witness in court

- **Tasks:**
  - What a text, either spoken or written, says.
  - Who is the author of that text (plagiarised or original).
  - What is the linguistic profile of a text.
- **Research domains**
  - Forensic voice comparison leading to reliable speaker identification (Forensic phonetics & acoustics).
  - ***Forensic written text comparison leading to***
    - ***reliable authorship attribution.***
    - ***reliable plagiarism detection.***
  - Trademark litigation.

Language as Evidence
Premises and assumptions

- Language provides oral and written information of several kinds.
- The linguistic production of individual speakers and writers can reveal an individual's socio-individual and socio-collective traits.
- Each individual has an idiosyncratic idiolectal style, which has to do with
  - a) how a **language** , shared by lots of people, is used in a distinctive way by a particular individual (Turell 2010).
  - b) the speaker/writer's production, which appears to be 'individual' and 'unique' (Coulthard 2004).
  - c) Halliday's (1989) proposal of 'options' and 'selections' from these options.

-

Language as Evidence
Object of Study

- Language as it occurs in real forensic contexts:
  - **Real FL case data**: legal investigative proceedings

- Language as it occurs in the real world:
  - **Real W text data**: linguistic research leading to controlled experiments, and thus to more validity and reliability in both plagiarism detection and authorship attribution.

Language as Evidence

The nature of the linguistic material involved

Types of text

plagiarism studies/ literary authorship

- Long
- Non-spontaneous
- Addressed to a big audience
- Planned

Types of text

criminal authorship

- Short
- Incidental and spontaneous
- Addressed to a limited audience
- Production limited by space and time
- Emotional

- **Context of production:**

Minimal proportion of an individual's style ☞ Inadequacy of linguistic fingerprint

Language as Evidence
Models/Hypotheses

- **Theory of Language Variation**

  a) **inter-writer** rather than intra-writer variation.

  b) Idiolectal style

  - **quite stable throughout time.**
  - **not so stable according to textual genre.**

Language as Evidence
Required methodology

- **Qualified opinions:**
  - based on scientific methodologies.
  - fundamented on both:
    - **Qualitative** methods (derived from the linguist's knowledge).
    - **Automatic/semi-automatic** and **Quantitative** methods (to introduce reliability and accountability).

Part 2

Examples of Linguists’ Achievements

Uniqueness of Encoding: Plagiarism

“Plagiarism is a form of cheating in which the student tries to pass off someone else's work as his or her own..... Typically, substantial passages are ‘lifted’ verbatim from a particular source without proper attribution having been made.”

Discuss the kind of policy a primary school should have towards bilingualism and multilingualism

It is essential for all teachers to understand the history of Britain as a multi-racial, multi-cultural nation. Teachers, like anyone else, can be influenced by age old myths and beliefs. However, it is only by having an understanding of the past that we can begin to comprehend the present.

a. It is essential for all teachers to understand the history of Britain as a multi-racial, multi-cultural nation. Teachers*, like anyone else, can be* influenced *by* age old myths and beliefs However, it is only by having an understanding of the past that we can begin to comprehend the present

b. In order for teachers to competently acknowledge the ethnic minority, it is essential to understand the history of Britain as a multi-racial, multi- cultural nation. Teachers *are prone to believe* popular myths and beliefs; however, it is only by understanding and appreciating past theories that we can begin to anticipate the present

c. *It is very important for us as educators to realise* that Britain as a nation *has become* both multi-racial and multi-cultural. Clearly it is vital for teachers *and associate teachers* to ensure that popular myths and *stereotypes* held by the wider community do not influence *their teaching. By* examin*ing British history* this will assist our understanding and in that way be better equipped to deal with the present and the future

Plagiarism - UCAS Personal Statements

- *234 statements related a dramatic incident involving "burning a hole in pyjamas at age eight“.*
- *175 contained a statement which involved "an elderly or infirm grandfather".*
- *370 statements contained a sentence including "a fascination for how the human body works..."*

Example of a Personal Statement

Ever since I accidentally burnt holes in my pyjamas after experimenting with a chemistry set on my 8th Birthday, I have always had a passion for science. Following several hospital visits during my teenage years to explore my interest, the idea of a career that would exploit my humanity and problem-solving abilities always made medicine a natural choice.

Instanced Personal Statement

Ever since I burnt holes in my dress after experimenting with my brother's chemistry set when I was 10, I have always been passionate about the sciences. Following several visits to the local hospital during my teenage years as a result of minor accidents, the idea of a career that would help people always made physio-therapy a natural choice.

Uniqueness of linguistic encoding

Stat: I asked her if I could carry her bags

Int: I asked her if I could carry her bags

Stat: I picked something up like an ornament

Int: I picked something up like an ornament

(Appeal of Robert Brown)

Uniqueness of linguistic encoding

I asked	2,170,000
I asked her	284,000
I asked her if	86,000
I asked her if I	10,400
I asked her if I could	7,770
I asked her if I could carry	7
I asked her if I could carry her	4
I asked her if I could carry her bags	0

Uniqueness of linguistic encoding

I asked	75,000,000
I asked her	6,090,000
I asked her if	1,110,000
I asked her if I	110,400
I asked her if I could	78,700
I asked her if I could carry	15
I asked her if I could carry her	7
I asked her if I could carry her bags	5

Suspect Text Messages

Thought u wer grassing me up.mite b in trub wiv **me** dad told mum i was lving didnt giv a shit. been**2** kessick camping was great.ave**2** go **cya**

Hi jen tell jak **i am** ok now ever 1s gona b mad tell them **i am** sorry.living in scotland wiv my boyfriend.shitting **meself** dads gona kill me mum dont give a shite.hope nik didnt grass me up.keeping phone **of.**tell dad car jumps out of gear and stalls put it back in auction.tell him **i am** sorry

Jenny’s Text Choices Compared

I am	im	i am
I have	ive	ave
my	my	me
off	off	of
to	#2#	#2
see you	cu	cya

JENNY NICHOLL HISTORIC MESSAGES

Sum black+pink k swiss shoes and all the other shit like socks.We r goin2the indian.Only16quid.What u doin x

Yeah shud b gud.i just have2get my finga out and do anotha tape.wil do it on sun.will seems keen2x

Shit is it.fuck icant2day ive allready booked2go bowling.cant realy pull out.wil go2shop and get her sumet soon.thanx4tdlin me x

No reason just seing what ur up2.want2go shopping on fri and2will`s on sun if ur up2it

Sorry **im** not out2nite havnt seen u 4a while aswel.ru free2moro at all x

No **im** out wiv jak sorry it took me so long ive had fone off coz havnt got much battery

Only just turned my fone.havnt lied bout anything.no it doesnt look good but ur obviously jst as judgmental than the rest.cu wen i cu&i hope its not soon

I havnt lied2u.anyway **im** off back2sleep

I know i waved at her we wer suppose2go at4but was a buffet on later on so waited.anyway he had a threesome it was great cu around

Im tierd of defending myself theres no point.bye

Happy bday!wil b round wiv ur pressent2moz sorry i cant make it2day.cu2moz xxx

IM

COMPARED TO

I AM

SUSPECT TEXT MESSAGES

Thought u wer grassing me up.mite b in trub wiv me dad told mum i was lving didnt giv a shit.been2 kessick camping was great.ave2 go cya

Hi jen tell jak **i am** ok know ever 1s gona b mad tell them **i am** sorry.living in scotland wiv my boyfriend.shitting meself dads gona kill me mum dont give a shite.hope nik didnt grass me up.keeping phone of.tell dad car jumps out of gear and stalls put it back in auction.tell him **i am** sorry

Y do u h8 me i know mum does.told her i was goin.i aint cumin back and the pigs wont find me.**i am** happy living up here.every1 h8s me in rich only m8 i got is jak.txt u couple wks tell pigs **i am** nearly 20 aint cumin back they can shite off

She got me in this shit its her fault not mine get blame 4evrything.**i am** sorry ok just had 2 lve shes a bitch no food in and always searching me room eating me sweets.ave2 go ok **i am** very sorry x

Plagiarism by Intralingual Translation

With all of these problems it was little short of a miracle that the "stichting" board was ready to lay the cornerstone for the building in the summer of 1907 at the opening of the Second Hague International Conference. It then took six more years before the Palace was completed during which time there continued to be squabbles over details, modifications of architectural plans and lengthy discussions about furnishings. For ten years the Temple of Peace was a storm of controversy, but at last, on 28 August 1913, the Grand Opening ceremonies were held. (J F Wall, *Andrew Carnegie)*

The foundation stone was not laid until the summer of 1907, in nice time for the opening of the Second Hague International Conference. Actual construction of the palace took a further six years, delayed and exacerbated by constant bickering over details, specifications and materials. For an entire decade the Peace Palace was bedevilled by controversy, but finally, on 28 August 1913, the opening ceremony was performed. (J Mackay, *A Life of Andrew Carnegie)*

Authorship Attribution Extortion Case Complementary Linguistic Evidence

- **Qualitative** textual analysis
- **Corpus linguistic analysis** of
  - grammatical evidence
  - sociolinguistic evidence
- **Statistical analysis** of sequences of linguistic categories.

Authorship Attribution Extortion Case

- Corpus

. Extortion found in one of a number of Spanish emails (DT@).

. Authorship denied later on.

. Supposedly sent to the company when also writing some faxes (NDTfax).

. Clear-cut authorship attribution context.

. Helping a Spanish civil court to decide whether the author of 4 NDTfax texts could also be the author of the DT@ texts, whose authorship this individual denied after he had been dismissed by his company for extorting them.

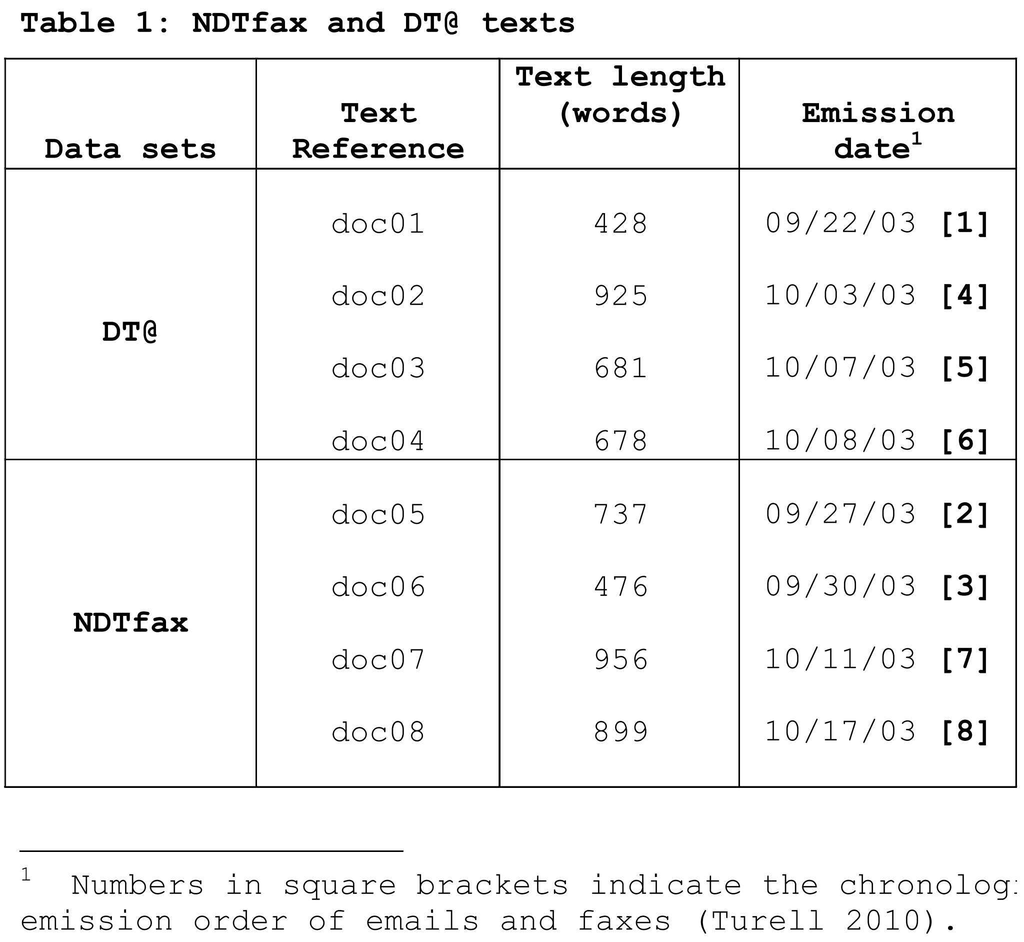


Table 1: NDTfax and DT@ texts

Data sets	**Text Reference**	**Text length (words)**	**Emission date**[1]
DT@	doc01	428	09/22/03 **[1]**
	doc02	925	10/03/03 **[4]**
	doc03	681	10/07/03 **[5]**
	doc04	678	10/08/03 **[6]**
NDTfax	doc05	737	09/27/03 **[2]**
	doc06	476	09/30/03 **[3]**
	doc07	956	10/11/03 **[7]**
	doc08	899	10/17/03 **[8]**

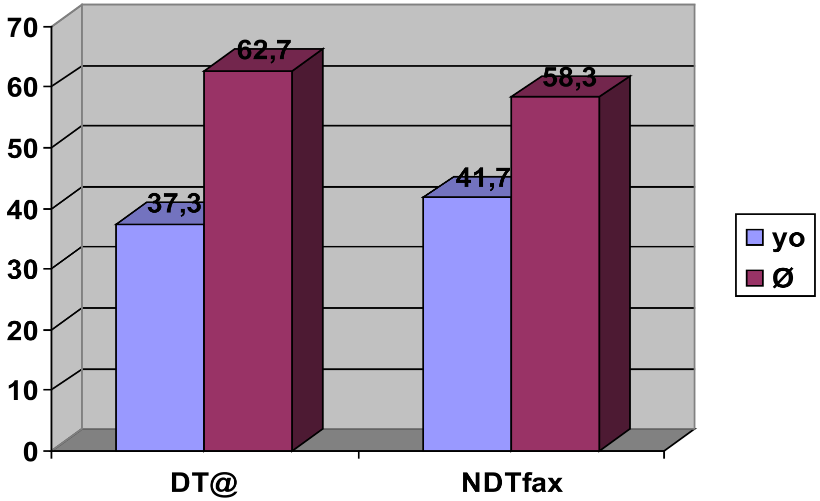
[1] Numbers in square brackets indicate the chronolog: emission order of emails and faxes (Turell 2010).

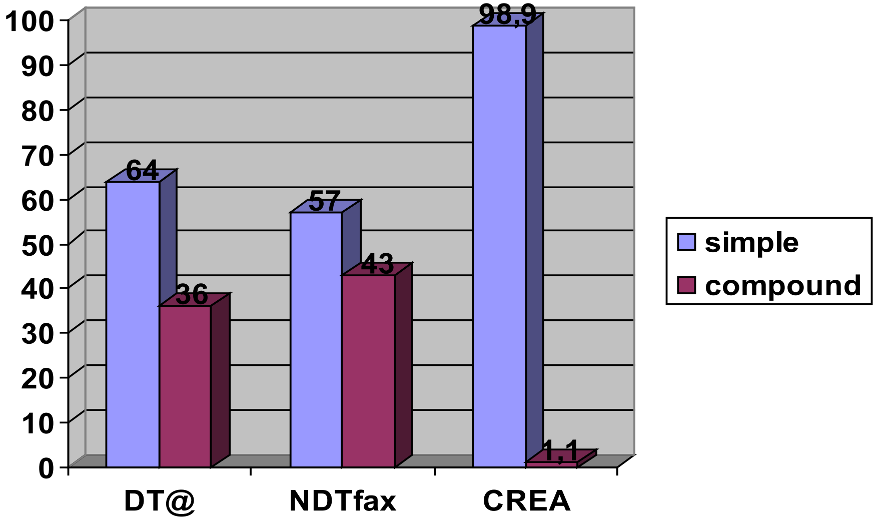
Corpus Linguistics

The use of corpora to analyse grammatical and sociolinguistic evidence (Turell 2010)

The Spanish first person singular pronoun (1PSP)

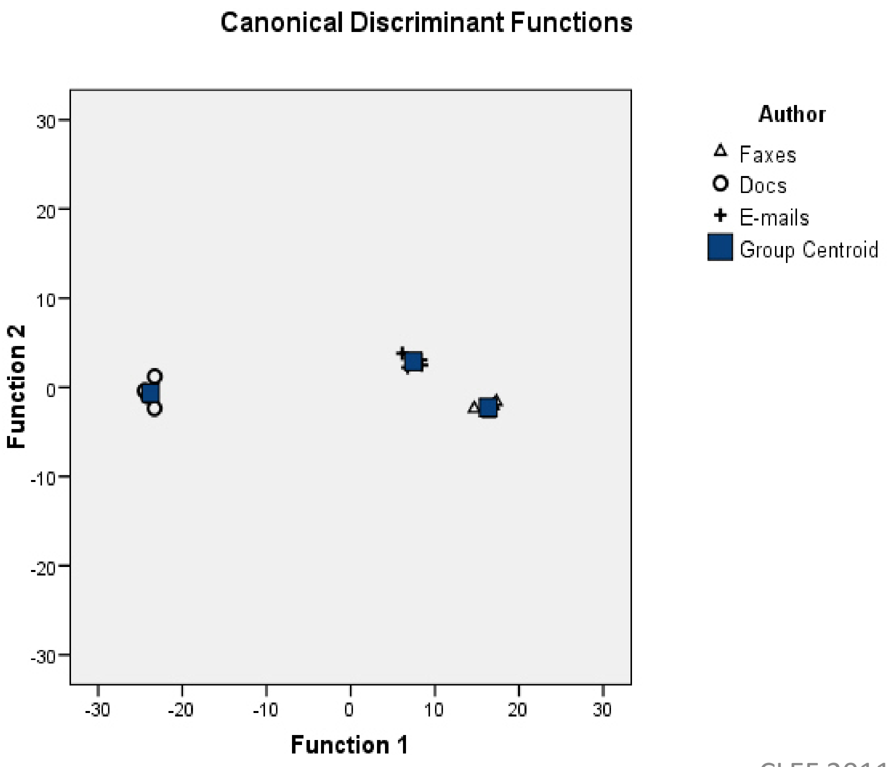
The Spanish relative pronoun (single** que **/ compund** el cual**)



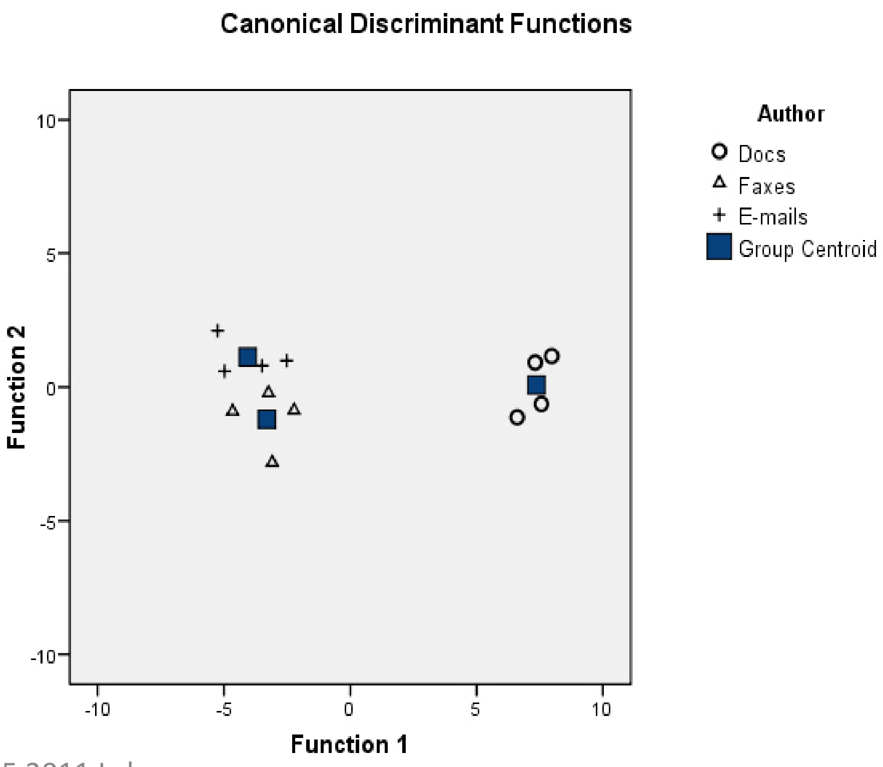


Statistical Analysis of Sequences of Linguistic Categories (Turell 2010)

Discriminant Function Analysis
(NDTfax and DT@)
Bigrams



Discriminant Function Analysis
(NDTfax and DT@)
Trigrams

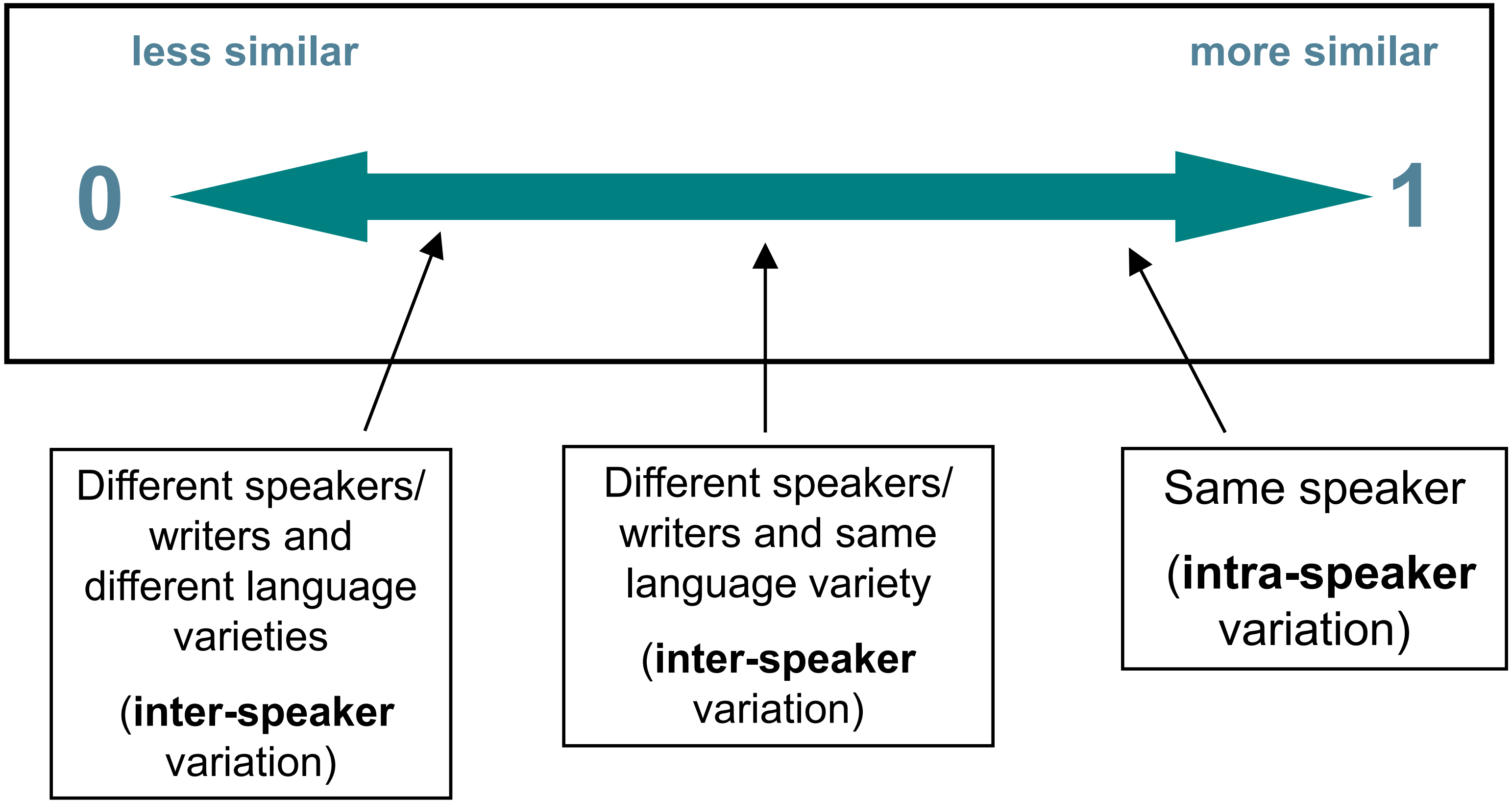


Index of Idiolectal Similitude (or Distance)

Research projects sponsored by the Spanish Ministry of Science and Technology

(EXPLORA -HUM2007-29140-E and **FFI2008-03583)**

- The IIS as a continuum



less similar

more similar

0

1

Different speakers/ writers and different language varieties

(**inter-speaker** variation)

Different speakers/ writers and same language variety

(**inter-speaker** variation)

Same speaker

(**intra-speaker** variation)

Part 3

Challenges for computerized analysis

Possible Collaboration of forensic and computational researchers

Plagiarism

- Plagiarism directionality between contemporary texts.
- Detecting plagiarism of meaning: pragmatic resources/figures of speech.
- Automatic detection of paraphrasing.
- Translingual plagiarism.

Authorship

- Base Rate population statistics.
- Bayesian LR for written texts.
- Identifying first language of non-native writers.
- Linguistics of impersonation: chatting like a 14-year old.
- Automatic analysis of SMS.
- Accounting for empty contexts (“**don't occur” variants**).

References

- Coulthard, R. M. (2004) Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25(4): 431-447.
- Coulthard, R. M. and Johnson, A. (2007) *An Introduction to Forensic Linguistics: Language in Evidence* London: Routledge.
- Grant, T. Txt 4n6: Idiolect free authorship analysis? *Routledge Handbook of Forensic Linguistics. Coulthard M and Johnson A,* (Eds) *London: Routledge, 508-22*
- Halliday, M.A.K. (1989) *Language, context and text. Aspects of language in a social-semiotic perspective.* Oxford: University Press, Oxford.
- Johnson, A. (1997) Textual kidnapping – a case of plagiarism among three student texts, *Forensic Linguistics: The International Journal of Speech, Language and Law* 4 (2) 210-25.
- 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. Forensic Linguistics* 17(2): 211-250.