

### Towards a morphosyntactic base-rate knowledge: apparent and real time dimensions of idiolectometric analyses of Catalan

M. Teresa Turell  
teresa.turell@upf.edu  
Home page: <http://www.upf.edu/pdi/iula/teresa.turell/>  
HUM 2007-29140-EXPLORA  
FFI 2008-03583-FILO  
Spanish Ministry of Science and Technology  
[http://www.iula.upf.edu/rec/ideolec/index\\_eng.htm](http://www.iula.upf.edu/rec/ideolec/index_eng.htm)

8th IAFEL Conference - Amsterdam - July 7, 2009

### Modern Forensic Linguistics: some challenges

- Undertaking forensic voice and written text comparison analyses,
- by using data from real world texts, reference corpora texts and real case documents,
- which could help the FL to attain more reliable and robust reports in forensic speaker identification and authorship attribution contexts.

8th IAFEL Conference - Amsterdam - July 7, 2009

### Our approach

- Forensic voice and written text comparison methodology should be based upon both
  - qualitative approaches
  - automatic (or semi-automatic) and quantitative approaches
- to real-world and real-case texts.

8th IAFEL Conference - Amsterdam - July 7, 2009

### Modern Forensic Linguistics: specific challenges

- a) to establish the Base Rate Knowledge of population distribution
  - in as many languages as possible
  - for as many as voice parameters and variables
  - for as many written markers of authorship as possible;
- b) to establish the threshold level of idiolectal similitude or distance above or below which one could reliably consider that two or more particular voices or written samples could have been produced by the same individual;
- c) to extend the LR Bayesian framework to forensic written text comparison.

8th IAFEL Conference - Amsterdam - July 7, 2009

### Paper's content

- Experimental work in progress derived from two research projects on Forensic Idiolectometry.
- The overall aim is to propose a protocol for the establishment of an **Index of Idiolectal Similitude (IIS)** applied to
  - phonological/phonetic, morpho-syntactic and discourse-pragmatic modules
  - of Catalan, Spanish and English
- that could be used in forensic spoken or written text comparison.

8th IAFEL Conference - Amsterdam - July 7, 2009

### General background (1) Projects' general aim

- To study the speakers' **idiolect** in its application to forensic voice comparison and forensic written text comparison.

8th IAFEL Conference - Amsterdam - July 7, 2009

### General background (2) Idiolect

- “The set of options that speakers/writers take from the linguistic repertoire (phonological, morphosyntactic, pragmatic) available to them speakers/writers of a specific language” (Nolan 1994: 331).
- A speaker’s idiolect seems to be individual and unique (Coulthard 2004).

8th IAFEL Conference – Amsterdam – July 7, 2009

### General background (3) Idiolectometry

- Emerging discipline which studies the idiolect.
- Forensic and non-forensic idiolectometry has a long-standing development, particularly
  - in the area of phonetics and acoustics and in measurements of written idiolects (Foster 2001; Baayen & van Halteren et al. 1996, 2005; Feiguina & Hirst 2007; Spassova & Turell 2007)
  - in identification of idiolectal style (Chaski 2001; Grant & Baker 2001; McMenamin 2002).
  - in stylistics (McMenamin 2002).
  - vocabulary studies by core, hapax legomena and dislegomena, lexical density, lexical richness (Coulthard 2004; Turell 2008).

8th IAFEL Conference – Amsterdam – July 7, 2009

### General background (4) Focus

- Measurement of the linguistic differences existing between idiolects and each individual’s idiolectal distance, so that an **Index of Idiolectal Similitude (IIS)** can be obtained.
- Establishment of what kind of idiolectal similitude one needs to have before one can safely and reliably say that two linguistic samples (spoken or written) have been produced by the same person.

8th IAFEL Conference – Amsterdam – July 7, 2009

### General background (5) Methods 1 and 2

- **Method 1** consists in comparing the percentages of
  - a) marked variants
  - b) non-standard variants
 – in variable or categorical realizations, obviously with variables that occur in the speakers’ modality of discourse under comparison.
- **Method 2** involves the same procedure as for M1 but attributing a specific weight to groupings of variables (50% ~ 50%; 40% ~ 60%; 30% ~ 70%; 20% ~ 80).

8th IAFEL Conference – Amsterdam – July 7, 2009

### General background (6) Methods 3 and 4

- **Method 3** measures the Euclidian distance between 2 speakers, for bi-dimensional and n-dimensional contexts, normalises the results in order to obtain a range between 0 and 1 and calculates the IIS.
- **Method 4** consists in running cross-tabulation in SSPS and looking for the Adjusted Residual Value (ARV) for each variable.

8th IAFEL Conference – Amsterdam – July 7, 2009

### The experimental study (1) Description

- Experimental idiolectometric morpho-syntactic analysis.
- Methods M1 and M2.
- Set of recordings in AT and RT.
- Speakers from a dialectal area of Central Catalan:
  - **Tarragoní**, around the area of the Roman city of Tarragona.

8th IAFEL Conference – Amsterdam – July 7, 2009

## The experimental study (2) Objectives and hypotheses

- **Twofold objective:**
- To provide the preliminary results obtained in the calculation of the IIS for the morpho-syntactic module of Catalan between speakers of the same dialect in two different times of measurement (AT: at the end of the eighties; RT: 20 years later (2007-2008), by applying calculation methods M1 and M2.
- To test two working hypotheses used in forensic studies of the idiolect, namely
  - a) that there seems to be **more inter-speaker than intra-speaker variation**, although inter-speaker variation will be lower for speakers from the same dialect.
  - b) that it **a speaker's idiolect doesn't vary substantially throughout time**, in terms of its morphosyntactic structure.

## The experimental study (3) General and specific IIS hypotheses

- a) that minimum deviance of idiolects will occur in **intra-speaker** comparisons of samples, in both apparent and real time, and so the expected IIS's will range between 0.70 and 1;
- b) small deviance will show in comparisons **between speakers from the same dialect** (inter-speaker), and so the IIS's will also be somewhat high, ranging between 0.40 and 0.70;
- c) the highest deviance will be observed in **inter-speaker** comparisons of speakers **from different dialects**, with IIS's between 0.10 and 0.40.

## The experimental study (4) Corpus Table 2

Speakers	Apparent Time (No. words)	Real Time (No. words)
VC	9,000 (originally 16,000) (recorded in 1988)	5,500 (recorded in 2007)
MG	8,000 (originally 11,000) (recorded in 1988)	6,900 (recorded in 2008)

## The experimental study (5) Instrument of data collection

- A revised version of the instruments used in variation sociolinguistics, that is, the **sociolinguistic interview or life story**.
- The AT and RT interviews differed in length, so the latter were shortened to **control for unevenness of data**.
- **Panel study**, within the Labovian methodological framework (Labov 2001; Turell 2003).

## The experimental study (6) Nature, typology and distribution of the variables Table 5

- Nature of variables described in tables 3 and 4

	Macro	Micro	Total
Generalising rules	4	1	5
Idiolectal features	2	6	8
Total	6	7	13

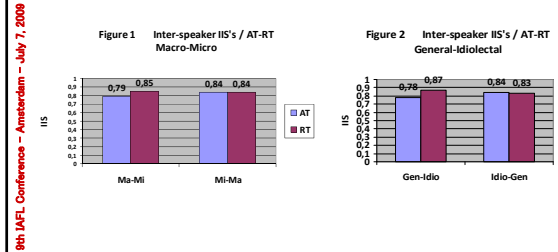
## The experimental study (7) Experiment 1

- **Experiment 1**
- Variables grouped by their MACRO or MICRO nature.
- IIS's for inter and intra-speaker variation in AT and RT.
  - To test **inter-speaker** variation:
    - a. Between Speaker VC and speaker MG in AT.
    - b. Between Speaker VC and Speaker MG in RT.
  - To test **intra-speaker** variation:
    - a. Between Speaker VC in AT and Speaker VC in RT.
    - b. Between Speaker MG in AT and Speaker MG in RT.

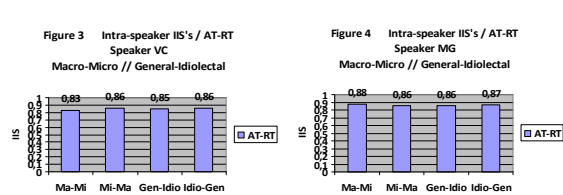
## The experimental study (8) Experiment 2

- Experiment 2
- Variables grouped by GENERALISING/DIALECTAL rules and IDIOLECTAL features.
- IIS's for inter and intra-speaker variation in AT and RT.
  - To test **inter-speaker** variation:
    - a. Between Speaker VC and speaker MG in AT.
    - b. Between Speaker VC and Speaker MG in RT.
  - To test **intra-speaker** variation:
    - a. Between Speaker VC in AT and Speaker VC in RT.
    - b. Between Speaker MG in AT and Speaker MG in RT.

## Results (1) (Weighing 80% ~20%)

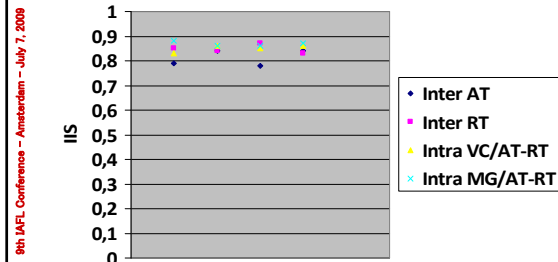


## Results (2) (Weighing 80% ~20%)



## Results (3)

Figure 5 All IIS calculations



## Conclusions

- Very preliminary experimental results.
- Conclusions and the paper itself would have benefited from richer and more conclusive results,
  - had we been able to make
    - many more comparisons with more speakers.
    - more comparisons from speakers from different dialects in real world recordings.
  - had we been able to start applying the technique to real forensic case recordings.

## Interest

- Idiolectometric measurements of spoken variables beyond phonetics and acoustics, in this case morphosyntactic variables found in spoken texts
  - seem to be justified.
  - can contribute to making forensic spoken text comparison more reliable and rigorous.
- This kind of research can make a methodological contribution to the **base rate knowledge of population distribution** which is needed for linguistic variables other than the phonological and phonetic ones.

### Future perspectives

- To set up a data base by saving
  - all realizations of all the variables studied and to be studied in the future.
  - all the calculated IIS's, of say 100-200 Catalan speakers, or more, from different dialectal areas of Catalan.
- so that when we get a disputed recording, or several disputed recordings, in a real forensic case, we are able to compare them
  - not only to non-disputed recordings,
  - but also to this base rate knowledge of population distribution.

**Thank you**