Corpus Applications in Forensic Computational Linguistics

Ángela Almela, PhD

UNIVERSIDAD DE
MURCIA

Applied linguistics as a science

UNIVERSIDAD DE MURCIA

- Linguistics is a science
  - Linguistic methods are tested by experiments.
  - Some linguistic methods can be automated in software.
- Corpus linguistics involves several disciplines and can be applied to different linguistic aspects.
  - Leech (1992): *language in use*

LAEL 50 years

Forensic computational linguistics

UNIVERSIDAD DE MURCIA

- Forensic linguistics attempts to solve forensic issues by using tools from:
  - Linguistics
  - Corpus linguistics
  - Computational linguistics

- Forensic computational linguistics
  - Developed out of linguistic theory and computational linguistics
  - Using language as evidence

- It is crucial that the expert witness provides error rates starting with known data, i.e., ground truth data.

LAEL 50 years

Forensic computational linguistics

UNIVERSIDAD DE MURCIA

- Innovations in forensic science
  - Examples of standard procedures from non-forensic sciences:
    - Forensic Toxicology from Chemistry
    - Forensic DNA Identification from Paternity Testing
    - Forensic Linguistics from Computational Linguistics

LAEL 50 years

Forensic computational linguistics

UNIVERSIDAD DE MURCIA

- Standard linguistics procedures, e.g.,
  - Normalizing frequency instead of using raw frequency when comparing corpora of different sizes.
  - Corpora annotation
    - Automatic
    - Manual
    - Semi-automatic
      - Post-editing human correction is recommended as a standard industry practice (Cantos-Gómez 2013; McEnery & Hardie 2012)

LAEL 50 years

Corpus linguistics for forensic purposes

UNIVERSIDAD DE MURCIA

But WHY USING CORPORA IN FORENSIC COMPUTATIONAL LINGUISTICS?

- Empirical analysis grounded in linguistic theory that can be replicated:
  - Coulthard (1994) advocated for the use of corpus in forensic linguistics given the possibilities that the empirical exploration of corpora can provide in terms of evidence and investigation.
  - Chaski (1997) developed the first specific corpus for forensic authorship identification
    - Funded by USDOJ and available to researchers who meet research proposal standards.

LAEL 50 years

Chaski Writing Sample Database (Chaski 1997)

Task ID	Topic
1	Describe a traumatic or terrifying event in your life and how you handled it
2	Describe someone or some people who have influenced you
3	What are your career goals and why?
4	What makes you really angry?
5	A letter of apology to your best friend
6	A letter to your sweetheart expressing your feelings
7	A letter to your insurance company
8	A letter of complaint about a product or service
9	A threatening letter to someone you know who has hurt you
10	A threatening letter to a public official or celebrity whom you do not know

Corpus linguistics for forensic purposes

UNIVERSIDAD DE MURCIA

The four corners of forensic linguistics

Identification
(speaker / author / language)

Text-typing, e.g.,
Is this a real threat?
Is this truthful or false?

Inter-textuality:
Are these texts related to each other?
How similar are these two texts?

Linguistic profiling, e.g.,
Age, Dialect, Gender, Education, L1

LAEL 50 years

Corpus linguistics for forensic purposes

UNIVERSIDAD DE MURCIA

Today's webinar

Identification
(author)

Text-typing:
Is this truthful or false?

LAEL 50 years

Corpus linguistics for forensic purposes

UNIVERSIDAD DE MURCIA

Text-typing:
Is this truthful or false?

LAEL 50 years

Text-typing: Is this truthful or false?

UNIVERSIDAD DE MURCIA

- Different tools for automated deception detection
  - Deception data can be experiment in lab versus real-life experience:
    - Low-stakes
      - No harm can be done
    - High-stakes
      - Real-life damages are possible and likely

LAEL 50 years

Text-typing: Is this truthful or false?

UNIVERSIDAD DE MURCIA

- Deception detection
  - Ground truth data that are forensically feasible
    - "Ground truth" data means data that we know what the correct answers are.
      - For deception detection, we need data where we know documents are true or false.
    - When a method is tested on ground truth data, we can accurately report its error rate.

LAEL 50 years

Text-typing: Is this truthful or false?

UNIVERSIDAD DE MURCIA

Computational classification of written statements as true or false

Automatic extraction of lexical features for different purposes, e.g.:

- General Inquirer (Stone et al. 1962)
- LIWC: Linguistic Inquiry and Word Count (Pennebaker et al. 2001)
- UMTextStats (García-Díaz et al. 2020)

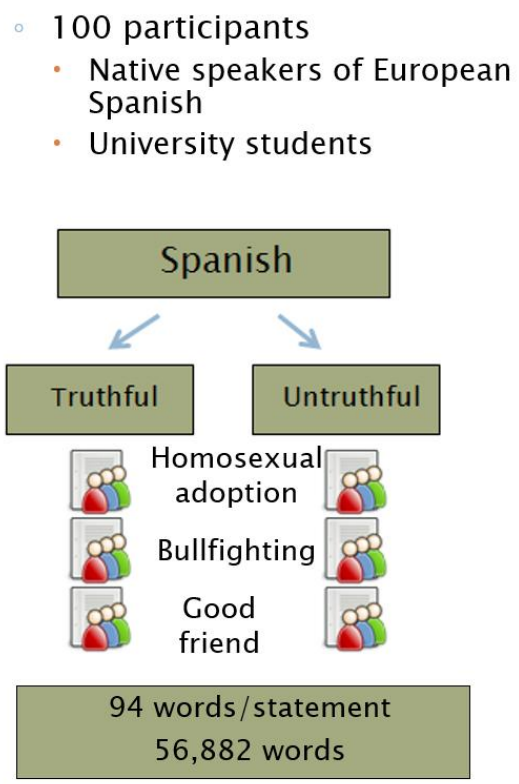
Software specifically developed for linguistic deception detection, e.g.:

- VERIPOL (Quijano et al., 2018)
- WISER (Chaski et al., 2015)

Text-typing: Is this truthful or false?

UNIVERSIDAD DE MURCIA

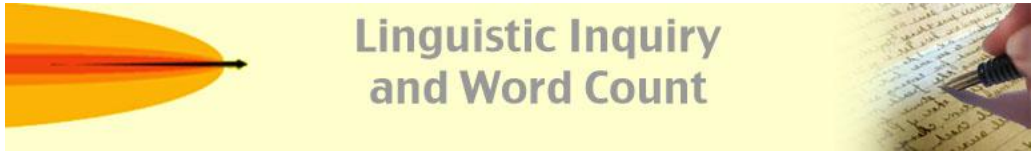
- Corpus collection for deception detection (Almela, Valencia-García, & Cantos 2013)



LAEL 50 years

Text-typing: Is this truthful or false?

UNIVERSIDAD DE MURCIA



- Automatic extraction of lexical features: LIWC categories
  - 4 broad dimensions
    - Linguistic processes
    - Psychological processes
    - Relativity
    - Personal concerns

➢ Relationship between language and…
  - ➢ State of mind
  - ➢ Mental health
  - ➢ Truth value

LAEL 50 years

Text-typing: Is this truthful or false?

Example: LIWC Dimension I – Standard linguistic categories

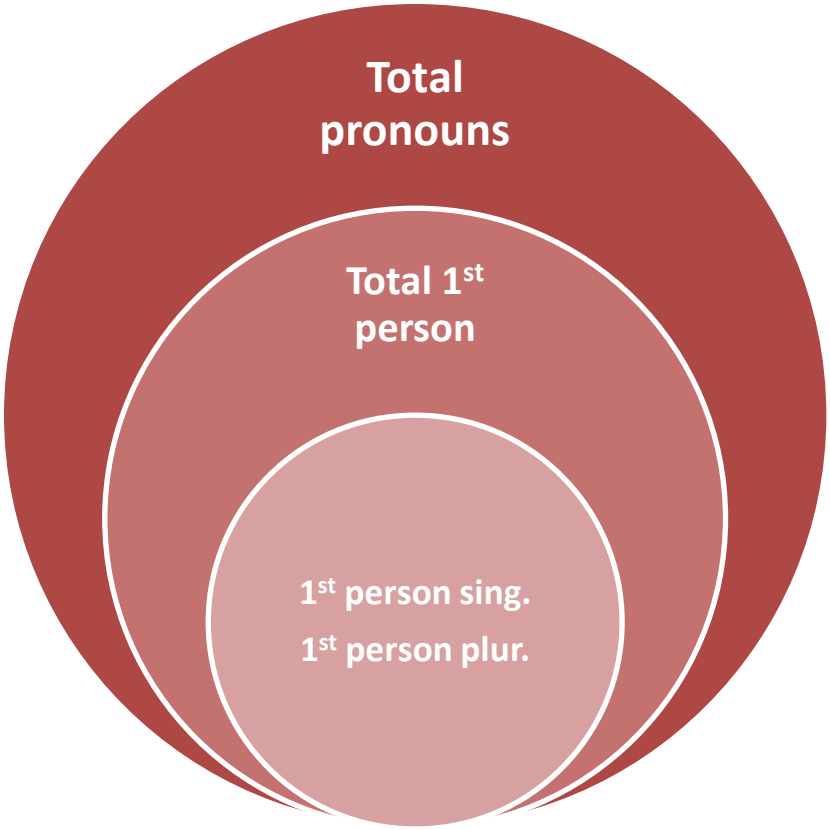
UNIVERSIDAD DE MURCIA

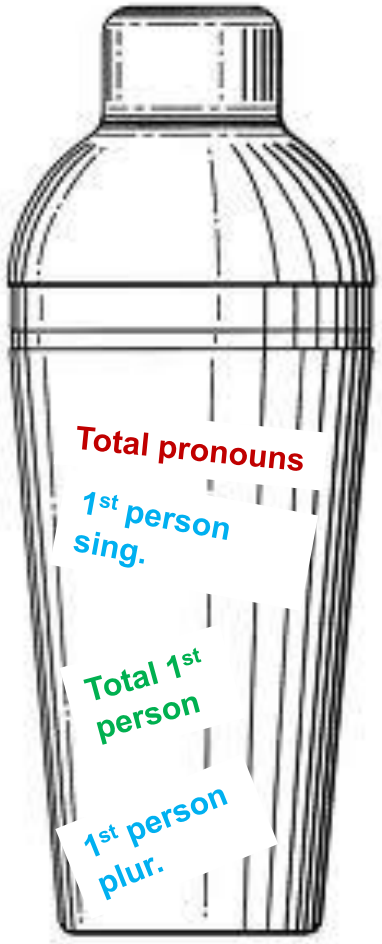
Category	Abbrev.	Examples
Word count	WC	
Words per sentence	WPS	
Sentences ending with ?	Qmarks	
% words longer than 6 letters	Sixltr	
Total pronouns	Pronoun	*I, our, they, you're*
Total first person	Self	*I, we, me*
1st person singular	I	*I, my, me*
1st person plural	We	*we, our, us*
Total second person	You	*you, you'll*
Total third person	Other	*she, their, them*
Negations	Negate	*no, never, not*
Assents	Assent	*yes, OK, mmhmm*
Articles	Article	*a, an, the*
Prepositions	Preps	*on, to, from*
Numbers	Number	*one, thirty, million*

Text-typing: Is this truthful or false?

Example: LIWC Dimension I – Standard linguistic categories

UNIVERSIDAD DE MURCIA





LAEL 50 years

Not to mix categories with redundant information in automatic classification experiments

Text-typing: Is this truthful or false?
Example: Stylometric dimension

UNIVERSIDAD DE MURCIA

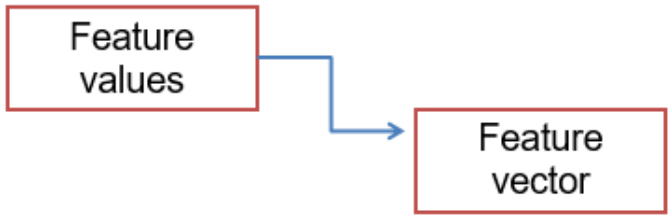
Standardized type/token ratio
Mean word length
Sentences/WC
1-letter words/WC
2-letter words/WC
3-letter words/WC
4-letter words/WC
5-letter words/WC
6-letter words/WC
7-letter words/WC
Complex words/WC

LAEL 50 years

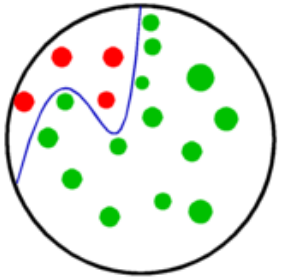
Text-typing: Is this truthful or false?
Machine learning techniques

UNIVERSIDAD DE MURCIA

- Goal of automatic classification (Almela et al. 2013)
  - To use an object's characteristics to identify which class it belongs to
  - Classification decision based on the value of a linear combination of the characteristics



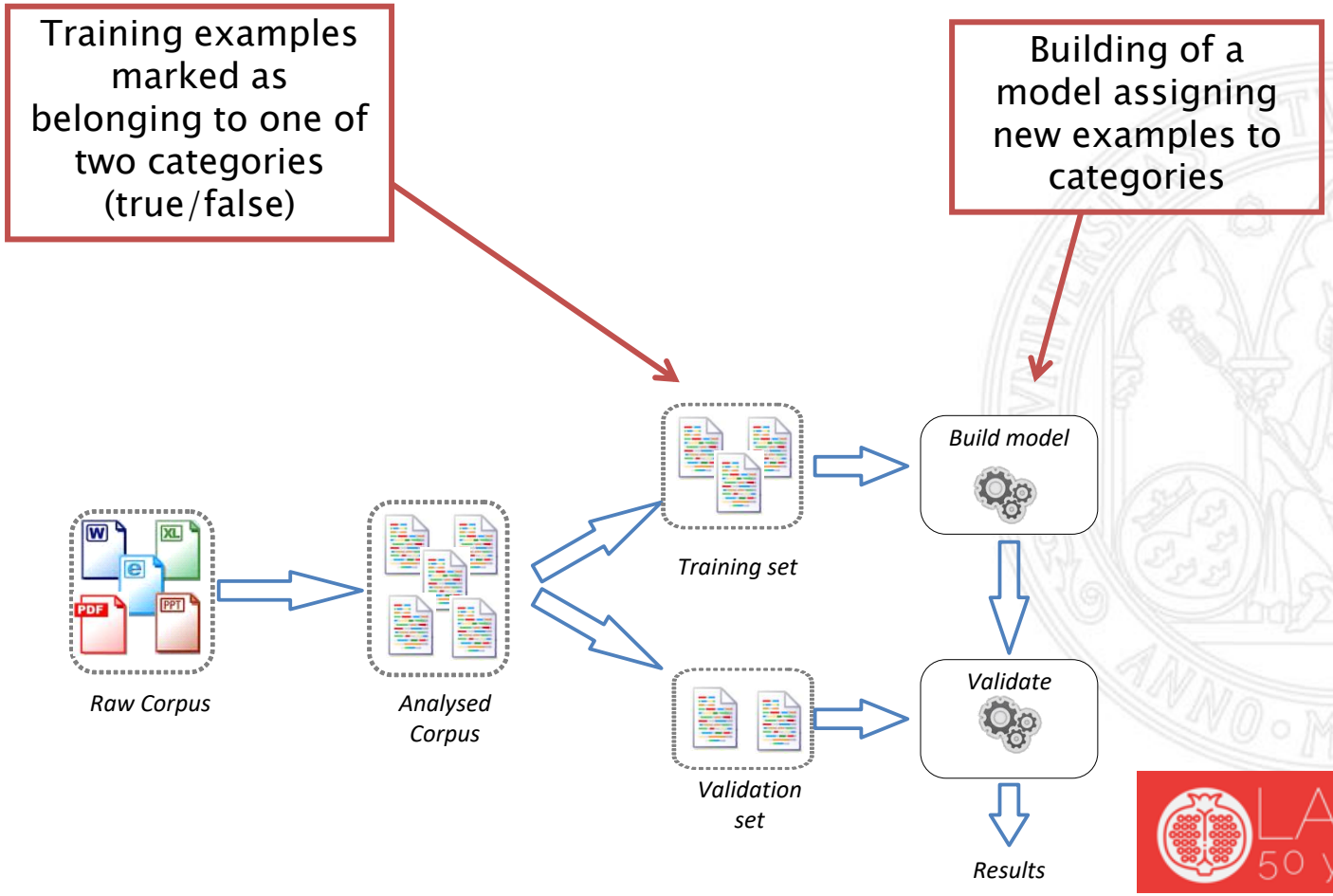
  - Support Vector Machine (SVM)
    - Representation of examples as points in space



Text-typing: Is this truthful or false?
Machine learning techniques

UNIVERSIDAD DE MURCIA

- 10-fold cross-validation



LAEL 50 years

Text-typing: Is this truthful or false?
Machine learning techniques

UNIVERSIDAD DE MURCIA

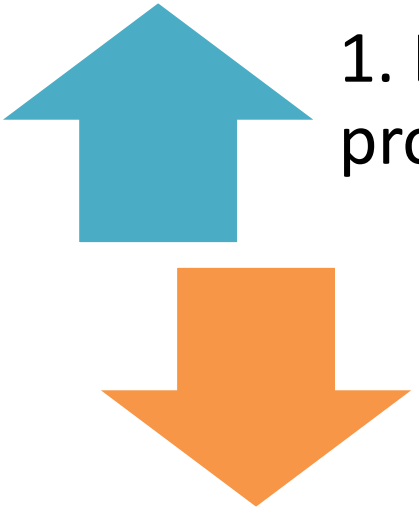
- Feature vectors = 31 classifiers
  - LIWC dimensions including terminal categories
  - Stylometric dimension

1	1+ styl.
2	2+ styl.
3	3+ styl.
4	4+ styl.
1_2	1_2+ styl.
1_3	1_3+ styl.
1_4	1_4+ styl.
2_3	2_3+ styl.
2_4	2_4+ styl.
3_4	3_4+ styl.
1_2_3	1_2_3+ styl.
1_2_4	1_2_4+ styl.
1_3_4	1_3_4+ styl.
2_3_4	2_3_4+ styl.
1_2_3_4	1_2_3_4+ styl.
Styl.	

LAEL 50 years

Text-typing: Is this truthful or false?
Machine learning techniques

UNIVERSIDAD DE MURCIA



1. Linguistic processes

4. Personal concerns

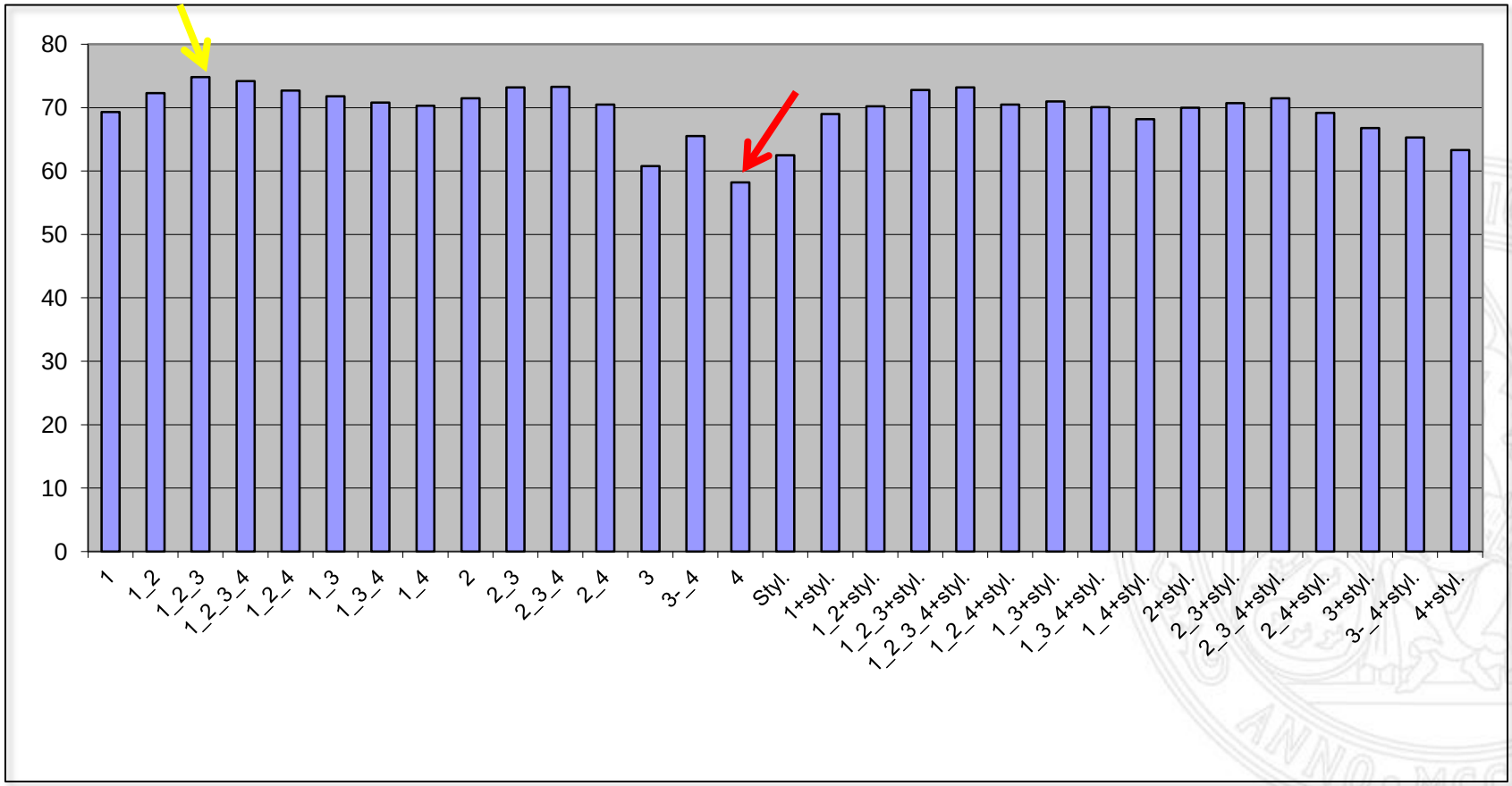
"Function words can provide powerful insight into the human psyche."

Chung & Pennebaker (2007: 344)

LAEL 50 years

Text-typing: Is this truthful or false?
Machine learning techniques

UNIVERSIDAD DE MURCIA



LAEL 50 years

Text-typing: Is this truthful or false?
Real life applications: VeriPol

UNIVERSIDAD DE MURCIA

- VeriPol (Quijano-Sánchez, Liberatore, Camacho-Collados, & Camacho-Collados 2018)
  - Assessment of false violent robbery cases for the Spanish National Police (CNP)
  - NLP and ML methods in a decision support system providing police officers with the probability that a given report is false
  - Ground truth data: Corpus of 588 false robbery reports and 534 truthful robbery reports

LAEL 50 years

Text-typing: Is this truthful or false?
Real life applications: WISER (WItness Statement Evaluation Rank)

UNIVERSIDAD DE MURCIA

- **ALIAS Technology, LLC (CEO: Carole E. Chaski, PhD)**
  - **Task:** Does this text contain deceptive language?
  - **Uses:** It can help investigators prepare for interrogations by analyzing witness statements after the interview but before the interrogation.
  - **Speed:** WISER1 runs very quickly, in minutes.
  - **Notes:** Law enforcement agencies who enter into a research relationship with the Institute for Linguistic Evidence (ALIAS Technology's sister for R&D) can obtain access to WISER1 without cost for a negotiated period of time.
  - **Accuracy:** It currently attains over 90% accuracy distinguishing truthful from false witness statements from actual criminal investigations. However, the Institute for Linguistic Evidence is conducting ongoing research on new text collections to determine under what conditions WISER can continue this high level of accuracy.
  - **Current languages:** English
  - **Research-in-progress languages:** Spanish

LAEL 50 years

Text-typing: Is this truthful or false?
Real life applications: WISER (WItness Statement Evaluation Rank)

UNIVERSIDAD DE MURCIA

ALIAS Technology

- Chaski, Almela, Holness, & Barksdale (2015):

1. Experimental data: students writing two narratives of a traumatic experience, one truthful and the other false → 71% accuracy

2. High-stakes data, actual statements from real criminal investigations with non-linguistic evidence of their veracity or falsehood → 93% accuracy

LAEL 50 years

In a nutshell…

UNIVERSIDAD DE MURCIA

- Importance of contextualized study of deception
- Ground truth data (collaboration with law enforcement)
- Existing tools = NOT INFALLIBLE
- As linguists, we should keep on testing what is used in real life and trying to improve it with our linguistic knowledge.

LAEL 50 years

ALIAS: A system for linguistic evidence

UNIVERSIDAD DE MURCIA

Linguistic Evidence Data

- System maintained on secured, web-accessible server.
- Subscribers access system to input/store linguistic data.
- Text, voice and image data can be stored.

Forensic Text Analysis

- ALIAS modules perform high-level linguistic analysis.
- Each module is tuned for its specific task.
- Modules can be added rapidly when R&D warrants.

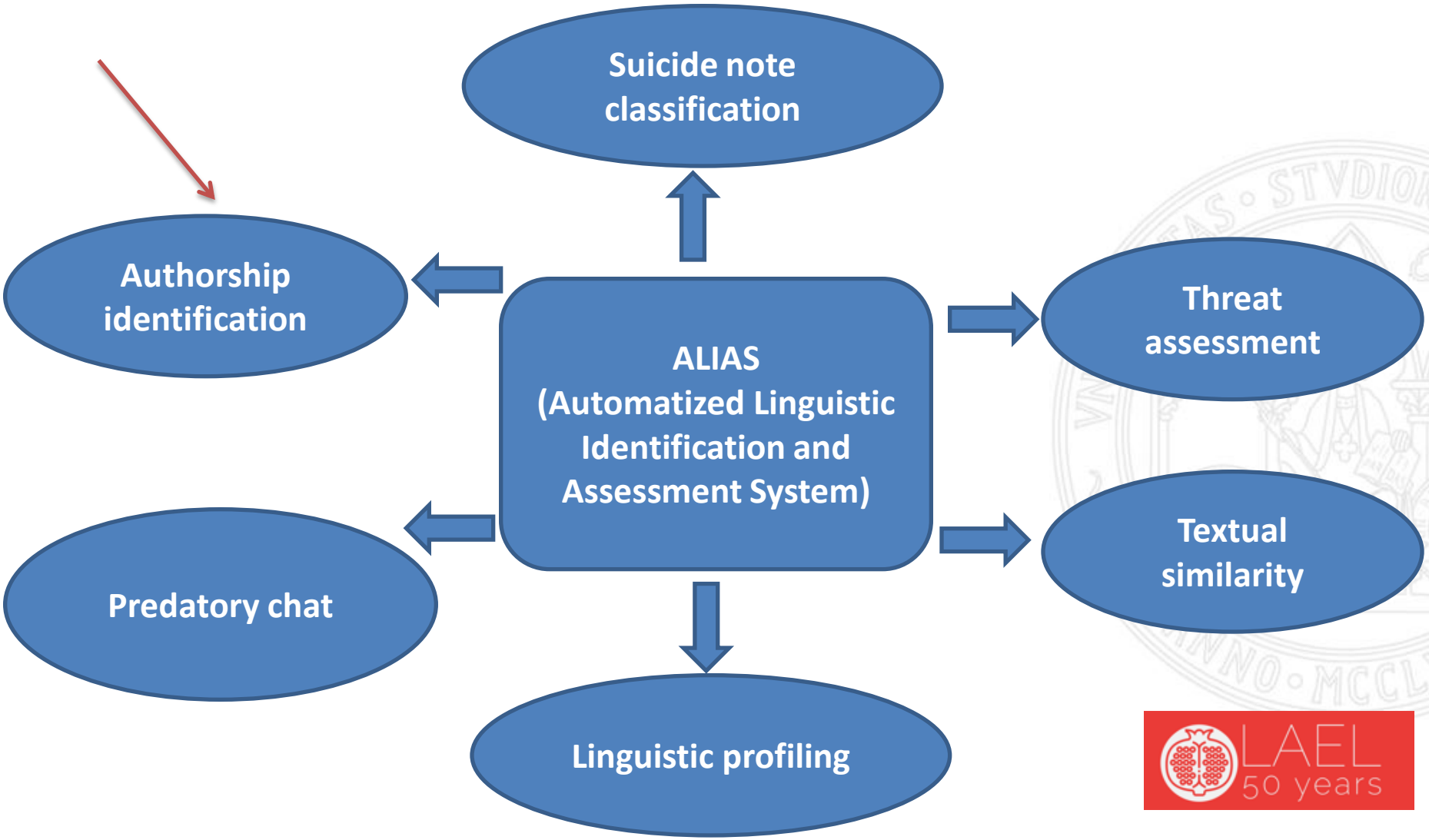
Results for Investigation & Testimony

- R&D includes testing on ground-truth data.
- R&D produces quantity requirements and error rates.
- ALIAS modules meet US legal evidence standards.

LAEL 50 years

ALIAS: A system for linguistic evidence

UNIVERSIDAD DE MURCIA



LAEL 50 years

Authorship identification

UNIVERSIDAD DE MURCIA

LAEL 50 years

ALIAS SynAID: Syntax-Based Author ID

UNIVERSIDAD DE MURCIA

- Syntactic method
  - Syntactic analysis
    - Identification of the roles that words and punctuation play in a given phrase structure or set of phrase structures;
    - Identification of the relationships between them.
  - Syntactic analysis also takes into account the complexity of phrase structures: **markedness** (Battistella 1990)
    - Unmarked phrase → "the professor"
    - Marked phrase → "the Swedish professor who you met at the conference"

LAEL 50 years

ALIAS SynAID: Syntax-Based Author ID

UNIVERSIDAD DE MURCIA

- The reality of phrase structure in our cognition
  - Common experience of being able to finish another person's sentence
  - Reality of phrase structure in our cognition
    - Demonstrated neurologically and accepted by linguists of all schools
  - Another common experience: not being able to repeat verbatim what has been said
    - Phrase structure degrades in memory within milli-seconds: we remember the meaning instead of the form, since the purpose of language is **communication of meaning**
    - Syntactic structures are not easy to imitate, because they are not easy to remember
      - They can be measured in all authors, since they must be used in producing language

LAEL 50 years

ALIAS SynAID: Syntax-Based Author ID

UNIVERSIDAD DE MURCIA

- Data analysis procedure: Main steps (Chaski 2005)
  - KD are analyzed linguistically using the SynAID method:
    - Each word is tagged for its syntactic category (noun, verb, adjective, etc.)
    - The combination of tags is categorized as “marked” or “unmarked”
    - Each document is measured for 26 syntactic features
    - The 26 SynAID linguistic categories are quantified so that each bundle and each individual text has a numerical profile
    - The numerical profiles of the known authors are tested statistically using Linear Discriminant Function Analysis to determine if SynAID can differentiate the different authors (cross-validated accuracy)
    - Using that statistical model, it is applied to classify the QD, reporting error rate

LAEL 50 years

ALIAS SynAID: Syntax-Based Author ID

UNIVERSIDAD DE MURCIA

- Syntactic method (Chaski 2005)
  - SynAID has been used in ~50 cases, with average accuracies >95%
  - Admitted as scientific evidence in Federal, State and Military Courts after evidence hearings
  - Used in Canada, Australia and Europe

LAEL 50 years

Institute for Linguistic Evidence (ILE) paradigm

UNIVERSIDAD DE MURCIA

1. Empirical testing of method independent of litigation.
2. Method is grounded in linguistics.
3. Method is tested on ground truth data that are forensically feasible.
4. All known and all questioned texts are analyzed the same way, by computer software whenever possible, or protocol.
5. Data are not contaminated.
6. Statistical procedures are in method, and follow standard principles of statistics including cross-validation.
7. A conclusion/prediction from testing the forensic data is stated in the report and in testimony.

ILE RESEARCH

https://linguisticevidence.org/

LAEL 50 years

Ongoing research

UNIVERSIDAD DE MURCIA

- Our research agenda for the WISER project:
  (1) Getting Spanish translations of ALIAS categories
  (2) Getting enough witness statements

- Our research agenda for the Spanish version of ALIAS SynAID (Almela, Cantos, & Almela 2020):
  (1) Contrastive analysis of English and Spanish
  (2) Measuring markedness categories in Spanish
  (3) Empirical testing of categories: Corpus test

LAEL 50 years

Final remarks

UNIVERSIDAD DE MURCIA

- Methods must be adjusted for realistic forensic use
- Methods must provide an error rate on forensic data
- Even within forensic science, methods should continue to be driven by research within the forensic setting

LAEL 50 years

References and recommended reading (I)

- Almela, A., Alcaraz-Mármol, G., & Cantos, P. (2015). Analysing deception in a psychopath's speech: A quantitative approach. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, *31*(2), 559-572.
- Almela, A., Cantos, P., & Almela, M. (2020, February). Formalizing Spanish Markedness: Working Toward a Spanish Version of the Automated Linguistic Identification & Assessment System (ALIAS) Syntax-Based Authorship Identification (SynAID). Paper presented at the *2020 Annual Scientific Meeting*, Anaheim, CA.
- Almela, A., Valencia-García, R., & Cantos, P. (2013). Seeing through Deception: A Computational Approach to Deceit Detection in Spanish Written Communication. *Linguistic Evidence in Security, Law and Intelligence*, *1*(1), 3-12.
- Battistella, E.L. (1990). *Markedness: The Evaluative Superstructure of Language*. Albany: State University of New York Press.
- Berber-Sardinha & Veirano, M. (Eds.) (2019). *Multidimensional analysis.* London: Bloomsbury Publishing: 97-124.
- Cantos, P. (2013). *Statistical methods in language and linguistic research*. London: Equinox.
- Carvell, H.T. & Svartvik, J. (1969). Computational Experiments in Grammatical Classification. *Janua Linguarum*, Series Minor 63. The Hague: Mouton.
- Chaski, C.E. (1997). Who wrote it? Steps toward a science of authorship identification. *National Institute of Justice Journal, 233*, 15-22.
- Chaski, C.E. (2001). Empirical Evaluations of Language-based Author Identification Techniques. *International Journal of Speech, Language and Law* (previously *Forensic Linguistics), 8*(1): 1-66.
- Chaski, C.E. (2005). Who's at the keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence, 4,* 1-13.
- Chaski, C.E. (2013). Best Practices and Admissibility of Forensic Author Identification. *Journal of Law and Policy*, *21*(2). Brooklyn Law School.
- Chaski, C.E., Barksdale, L., & Reddington, M. (2014). Collecting Forensic Linguistic Data: Police and Investigative Sources of Data for Deception Detection Research. Paper presented at the *Linguistic Society of America Annual Meeting*. Minneapolis, MN.
- Chaski, C.E., Almela, A., Holness, G., & Barksdale, L. (2015). WISER: Automatically Classifying Written Statements as True or False. Paper presented at the *American Academy of Forensic Sciences 67th Annual Scientific Meeting*. Orlando, FL.

References and recommended reading (II)

- Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.
- Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *International Journal of Speech, Language and Law*, *1*(1), 27-43.
- Fabb, N. (2005). *Sentence Structure*. London: Routledge.
- Labov, W. (1988). *The Judicial Testing of Linguistic Theory, in Linguistics in Context: Connecting Observation and Understanding*. Norwood: Ablex Publishing Corporation.
- Leech, G. (1992). Corpora and theories of linguistic performance. In Jan Svartvik (Ed.), *Directions in corpus linguistics.* Berlin: Mouton De Gruyter (pp. 105-122).
- McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice. Cambridge Textbooks in Linguistics.* Cambridge: Cambridge University Press.
- Olsson, J. (2008). *Forensic Linguistics (2nd edition).* London: Continuum International Publishing Group.
- Pennebaker, J.W., Francis, M.E., & Booth, R.J. (2001). *Linguistic Inquiry and Word Count.* Mahwah (NJ): Erlbaum Publishers.
- Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, J., & Camacho-Collados, M. (2018). Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police. *Knowledge-Based Systems*, *149*, 155-168.
- Ramírez-Esparza, N., Pennebaker, J.W., García, F.A., & Suriá, R. (2007). La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista Mexicana de Psicología 24*(1), 85-99.
- Stone, P.J., Bales, R.F., Namenwirth, J.Z., & Ogilvie, D.M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science,* 7, 484-494.
- Stone, P.J., Dunphy, D., Smith, M.S., & Ogilvie, D.M. (1966). *The General Inquirer: A computer approach to content analysis.* Cambridge, MA: MIT Press.
- van Halteren, H. (2007). Author verification by linguistic profiling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing, 4* (1), 1-17.

UNIVERSIDAD DE MURCIA

Thanks for your attention!

angelalm@um.es

aalmela@linguisticevidence.org