D23 Formalizing Spanish Markedness: Working Toward a Spanish Version of the Automated Linguistic Identification & Assessment System (ALIAS) Syntax-Based Authorship Identification (SynAID)

Ángela Almela, PhD, Universidad de Murcia, Murcia 30071, SPAIN; Pascual Cantos, PhD, Universidad de Murcia, Murcia 30071, SPAIN; Moisés Almela, PhD, Universidad de Murcia, Murcia 30071, SPAIN*

Learning Overview: After attending this presentation, attendees will better understand the motivation, challenges, and feasibility of developing a syntax-based method for Spanish authorship identification.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by enabling multilingual English-Spanish authorship identification for criminal investigations.

The Institute for Linguistic Evidence (ILE) has embraced the forensic computational linguistic paradigm from its inception in 1998, confronting the tradition of academicians functioning as expert witnesses in forensic linguistics without sustaining a proper research agenda in the field or without testing their methods for error rate. An ILE team performs scientific analysis firmly grounded on linguistic theory, in line with Labovian linguistics that "validity and utility are not distinct, […] theory only has value to the extent that it is relevant to real-life problems."[1] Thus, ILE researchers develop and/or adopt scientific methods that have been devised and tested for accuracy outside of any litigation, since the ILE paradigm attempts to promote a real forensic science that is useful to the judicial system, involving reliable, replicable, and respectable research.[2]

Within this framework, research conducted at ILE is grounded on research questions arising from the observation of practical problems in the forensic setting.[3] ILE researchers make use of computational linguistic analysis software, specifically ALIAS, as well as statistical software, such as SPSS, DTREG, and statistical analysis algorithms in Python.[3-6]

The main current language in ALIAS is English, and Spanish is among the research-in-progress languages. Accordingly, the present proposal puts forward the ongoing project aiming to develop the Spanish version of this computational tool. In particular, research and development has focused primarily on developing a Spanish version of SynAID.

Specifically, SynAID applies standard techniques of syntactic analysis from the core field of linguistics known as syntax.[7-9] Syntax is the mental process of combining words into phrases and phrases into sentences, so that meaning can be conveyed through language. Syntactic analysis is the process for identifying, within a sentence or phrase, each word's syntactic relationship to the other words with which it is being combined. In the sentence, "The dog bit the mailman," the words *the* and *dog* combine to make the phrase *the dog*, a noun phrase; the words *bit*, *the*, and *mailman* combine to make the phrase *the mailman*, which itself combines with *bit* to make the verb phrase *bit the mailman*. and *bit the mailman* combines with *the dog* to make the sentence *the dog bit the mailman.* Although syntax is unconscious and automatic, it has been proven to be psychologically and neurologically real.[10,11] Further, the memory of syntax degrades in milliseconds, as we focus on the meaning rather than the form.[12] Syntax is a core field within linguistics and is accepted within all paradigms of linguistics as central to human language.[7-13]

Further, the SynAID method for authorship identification is based on the advanced analytical method known as syntactic markedness, that is to say, the complexity of phrase structures, materialized in linguistic phenomena, such as extrapositions, dislocations, inversions, and preposings and internal structures of phrases.[7,8,14]

For the syntactic design of structures in Spanish, every syntactic head was listed with all the possible patterns for that head; then, these syntactic patterns were subcategorized into marked and unmarked patterns by considering levels of complexity and other factors in markedness theory.[14] A considerable number of similarities were found across languages (English-Spanish); nonetheless, certain linguistic phenomena unique to Spanish deserved special attention, such as subjunctive mood and clitics.

As for the data collection process, data was elicited using an instrument designed for forensic linguistic evidence.[15] The questionnaire consisted of ten questions, in response to which participants authored a different type of text in Spanish. An Institutional Review Board approval was processed in order to protect the welfare, rights, and privacy of human subjects. Once collected, the data were not altered in any way, which led to protocols for handling severely degraded data. Furthermore, researchers complied with all the ethical requirements for forensic research, namely anonymity, confidentiality, preservation of the data, and restricted access to the data, as presented elsewhere.[16]

Reference(s):

1. Labov, W. (1988). The Judicial Testing of Linguistic Theory, in *Linguistics in Context: Connecting Observation and Understanding*. Norwood: Ablex Publishing Corporation.
2. Chaski, C.E. (2013). Best Practices and Admissibility in Forensic Author Identification. *Journal of Law and Policy, 21*(2), 332-376.
3. Chaski, C.E. (2012). Author Identification in the Forensic Setting. In L. Solan and P. Tiersma (Eds.), *The Oxford Handbook of Forensic Linguistics* (pp. 333-372). Oxford: Oxford University Press.
4. Norusis, Marija J. (1994) *SPSS Professional Statistics 6.1.* Chicago: SPSS Inc.
5. Sherrod, Philip H. (2003). *DTREG: Predictive Modeling Software.* Nashville: DTREG.
6. McKinney, W. (2012). *Python for Data Analysis.* Sebastopol, CA: O'Reilly Media, Inc.
7. Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.
8. Pollard, C. and Sag, I. (1994). *Head-driven Phrase Structure Grammar.* Chicago, IL: University of Chicago Press.

*Presenting Author

9. Fabb, N. (2005). *Sentence Structure*. Londres: Routledge.
10. Fodor, Jerry A. and Bever, Thomas G. (1965). The Psychological Reality of Linguistic Segments, 4 *Journal of Verbal Learning & Verbal Behavior* 414, 414–20 (1965).
11. Pallier, Christophe, Devauchelle, Anne-Dominique and Dehaene, Stanislas. (2011). Cortical Representation of the Constituent Structure of Sentences, *PNAS* February 8, 2011. 108 (6) 2522-2527; https://doi.org/10.1073/pnas.1018711108.
12. Anderson. J.R. (1974). Verbatim and Propositional Representation of Sentences in Immediate and Long-term Memory. 13 *Journal of Verbal Learning & Verbal Behavior*, 149, 149-162 (1974).
13. Carvell, H.T. and Svartvik, J. (1969). *Computational Experiments in Grammatical Classification.* Janua Linguarum, Series Minor 63.The Hague: Mouton.
14. Battistella, E.L. (1990). *Markedness: The Evaluative Superstructure of Language*. Albany: State University of New York Press.
15. Chaski, C.E. (2001). Empirical Evaluations of Language-based Author Identification Techniques. *International Journal of Speech, Language and Law. 8*(1): 1-66.
16. Almela, A., Alcaraz-Mármol, G., García, A., and Pallejá-López, C. (in press). Developing and Analyzing a Spanish Corpus for Forensic Purposes. *LESLI: Linguistic Evidence in Security, Law and Intelligence,* 3.

Forensic Computational Linguistics, Syntax, Linguistic Markedness