Identificación de estratos sociolingüísticos en el Corpus PRESEEA-Valencia mediante aprendizaje automático

Adrián Cabedo Nebot

Universitat de València

Créditos

Proyecto de investigación (PID2023-148371NB-C42). Estudio de los condicionantes sociales del español actual en el centro y norte de España: NUEVAS IDENTIDADES, NUEVOS RETOS, NUEVAS SOLUCIONES (ECOS-C/N)

https://ecoscn.org/

Universitat de València

Email: adrian.cabedo@uv.es

Introducción

- **Objetivo**: Examinar elementos de variación prosódica influyen en la caracerización del sexo y la edad en español hablado en Valencia (secundariamente, sentimientos y morfosintaxis).
- **Datos**: Corpus PRESEEA; 136 hablantes con educación superior, 42.001 grupos entonativos.
- **Variables clave**: Frecuencia fundamental (F0), duración, intensidad, velocidad de habla, rango tonal.

Objetivo

Este estudio combina aspectos acústicos y lingüísticos con metodologías de análisis computacional.

Quilis, Antonio. 1993. Tratado de Fonética y Fonología españolas. Gredos.

Garrido Almiñana, Juan María. 2018. «Using Large Corpora and Computational Tools to Describe Prosody: An Exciting Challenge for the Future with Some (Important) Pending Problems to Solve». Pp. 3-43 en Methods in prosody: A Romance language perspective. Language Science Press.

Hipótesis y marcos teóricos

- **Hipótesis**: Existen patrones entonativos (sentimientos, morfosintaxis) específicos asociados al sexo y la edad.
- **Marcos teóricos**: AMH (Análisis Melódico del Habla) y TOBI (Tonos e Índices de Ruptura).
- MAS analiza contornos completos; TOBI se centra en acentos tonales y tonos de frontera.

Referencias más importantes

Cantero Serena, Francisco José, y Dolors Font-Rotchés. 2009. «Melodic analysis of speech method (MAS) applied to Spanish and Catalan». Phonica 5:33-47. doi: 10.1344/phonica.2009.5.33-47.

Estebas, Eva, y Pilar Prieto. 2008. «La notación prosódica del español: una revisión del Sp- ToBI». Estudios de fonética experimental (17):263-83.

Relación con la sociolingüística

! Cita de Johnstone

But contexts are never equivalent, because no two speakers could possibly be linguistically identical. In the inevitable sense that no two people share exactly the same linguistic memories, no two people speak alike: every speaker is idiosyncratic. Class, sex, age, region, the nature of the linguistic task, and the makeup of the audience all have an important bearing on how people sound; but they do not determine how people sound. These social facts, along with other factors such as ethnicity, ideology, and identity, provide (or withhold) resources among which individuals choose as they decide how to be and talk. (Johnstone and Bean 1997, 236)

Johnstone, Barbara, and Judith Mattson Bean. "Self-Expression and Linguistic Variation." Language in Society, vol. 26, no. 2, 1997, pp. 221–46. JSTOR, http://www.jstor.org/stable/4168762. Accessed 31 Mar. 2025.

Datos y metodología

- Corpus: Entrevistas PRESEEA (Valencia). https://preseea.uah.es/
- Transcripción y alineación: Whisper y Montreal Forced Alignment.
- Extracción acústica: PRAAT, etiquetado prosódico con Oralstats.

Moreno Fernández, Francisco. 2021. Metodología del "Proyecto para el estudio sociolingüístico del español de España y de América" (PRESEEA). Alcalá de Henares: Universidad de Alcalá.

Cabedo Nebot, Adrián. 2022. Oralstats.

Variables

1. filename	2. id_ip	3. ip_spk	4. ip_tier
5. ip_text	6. ip_tmin	7. ip_tmax	8. ip_duration
9. ip_mean_pitch	10. alargamiento	11. multiple_tonics	12. ip_range_st
13. ip_inflexion	14. ip_body_value	15. ip_toneme_prev	16. ip_toneme_value
17. TOBI	18. ip_anacrusis_displacement	19. ip_anacrusis	20. ip_body
21. ip_body_displacement	22. ip_toneme	23. ip_toneme_displacement	24. pattern
25. edad	26. sexo	27. sex	28. age
29. category	30. qanger	31. qanticipation	32. qjoy
33. qsad	34. qsurprise	35. qtrust	36. qfear
37. qdisgust	38. qdet	39. qnoun	40. qadj
41. qadv	42. qaux	43. qccon	44. qpron
45. qverb	46. qadp	47. qscon	48. qintj
49. qnum	50. qpropn		

Orientación computacional del estudio

La integración computacional permite un análisis eficiente y escalable. (Anthony 2020)

Anthony, Laurence. 2020. «Programming for Corpus Linguistics». Pp. 181-207 en A Practical Handbook of Corpus Linguistics, editado por M. Paquot y S. Th. Gries. Cham: Springer International Publishing.

Distribución de hablantes

	Edad	Count (%)
Mujer	middleage	12 (8.8%)
Mujer	older	15 (11.0%)
Mujer	young	45 (33.1%)
Mujer	**Total**	**72 (52.9%)**
Hombre	middleage	16 (11.8%)
Hombre	older	8 (5.9%)
Hombre	young	40 (29.4%)
Hombre	**Total**	**64 (47.1%)**
Total		**136 (100.0%)**

- Subconjunto equilibrado (8 hablantes por grupo; total 48).

Datos

Unidades en general

42.001 grupos entonativos 206.284 palabras 384.940 vocales

Unidades por hablante

Media de 875 grupos entonativos por hablante (DE = 311), 4298 palabras (DE = 1.676) y 8.020 vocales (DE = 3.067).

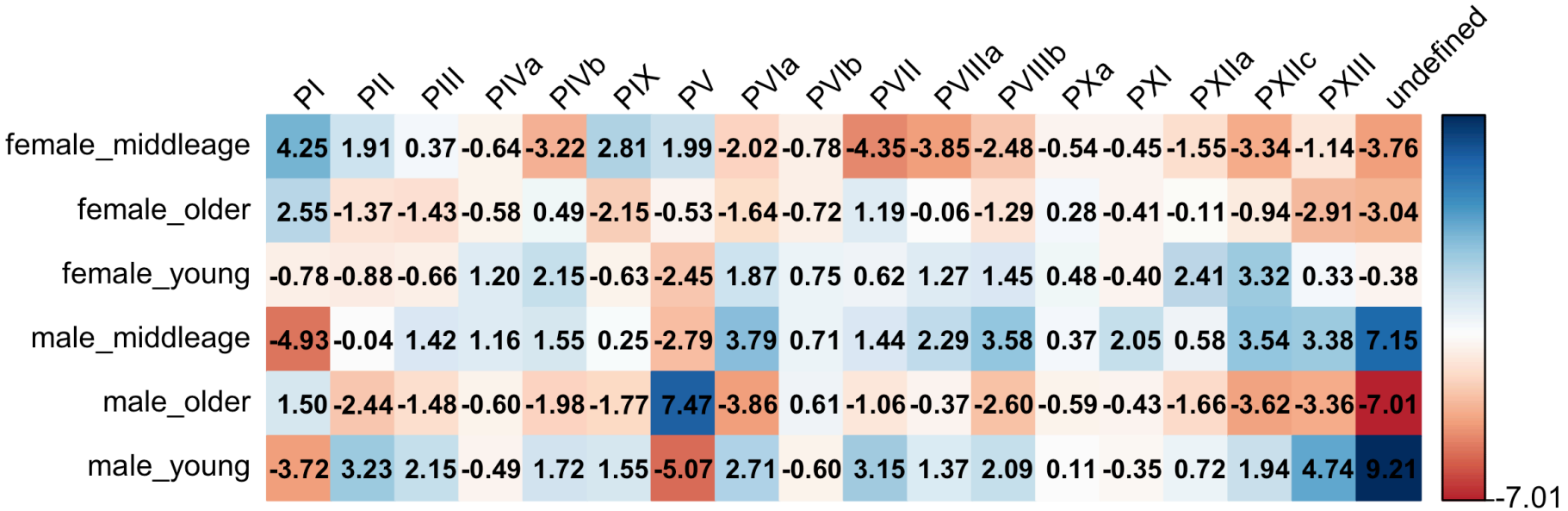
Duración

Duración total: 31 horas y 18 minutos, con una duración media de 39 minutos por hablante (DE = 8.5).

Aplicación Shiny

https://adrin-cabedo.shinyapps.io/AESLA_sample/

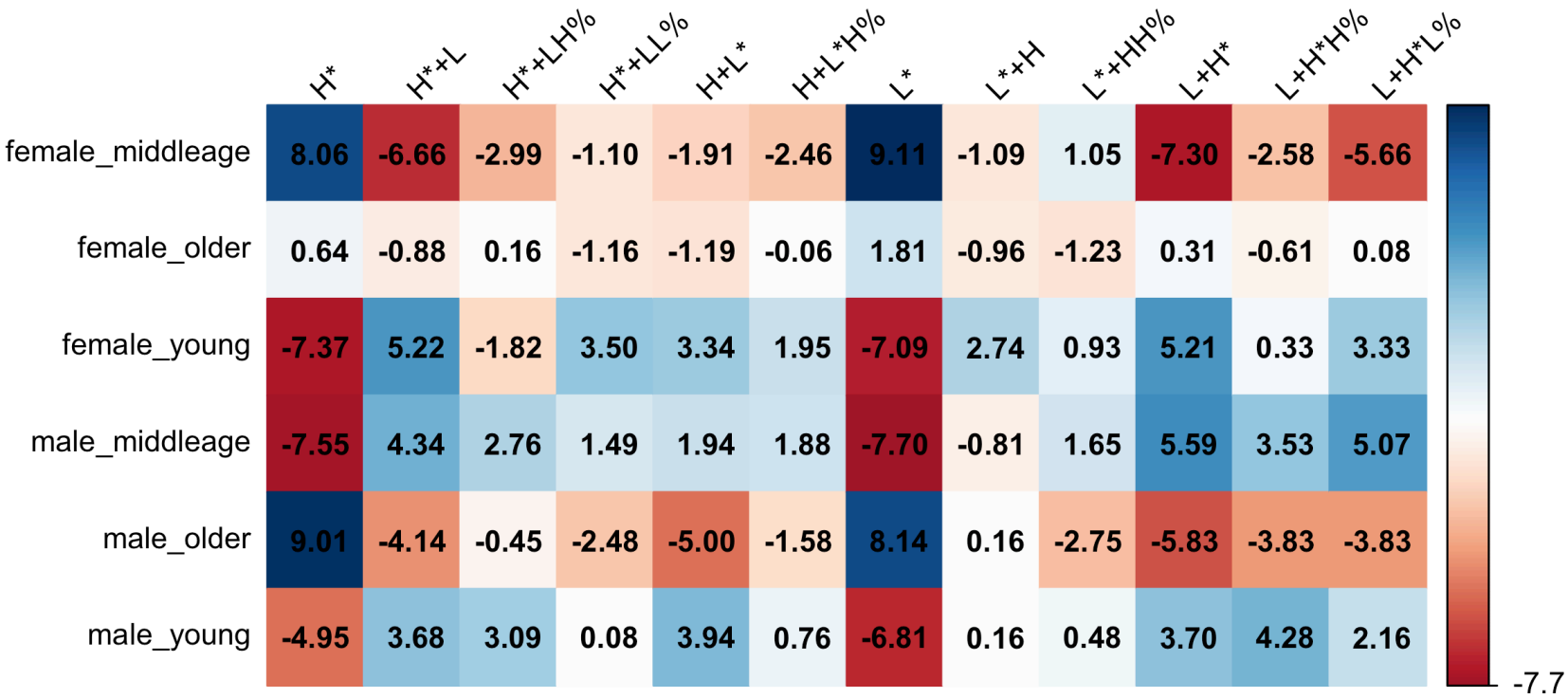
Patrones prosódicos (AMH) I



Patrones prosódicos (AMH) II

- Patrones dominantes: PI (54,4%), PV (20,1%), PVIIIa (7,2%).
- Mujeres jóvenes: contornos enfáticos y expresivos.
- Hombres mayores: contornos neutros o suspendidos.

Patrones prosódicos (TOBI) I



Patrones prosódicos (TOBI) II

- Patrones frecuentes: H*, L (22% cada uno), L+H* (18,6%).
- Mujeres jóvenes prefieren patrones ascendentes y enfáticos.
- Hombres mayores prefieren patrones descendentes básicos.

Modelo de clasificación Random Forest

- Clasificación combinada género-edad limitada (error OOB: 77,56%).
- Clasificación por género mejor (error OOB: 45,92%); edad intermedia (61,13%).
- Variables más importantes: rango tonal, patrones AMH, TOBI, duración.

Clasificación de estratos medio-baja

Los resultados evidencian una ligera relación entre prosodia y variables sociolingüísticas.

Comparación AMH vs TOBI

- AMH supera a TOBI en la caracterización de estratos sociales y de individuos.
- Contornos completos (AMH) diferencian mejor las variables sociales.
- TOBI, centrado en tonos de frontera, menos informativo por sí solo.

Discusión

- Las mujeres jóvenes muestran mayor expresividad prosódica, adoptando patrones entonativos más amplios y enfáticos (Crystal, 2013; Philips, 1980).
- Los hombres mayores tienden a patrones prosódicos más simples y graves, lo que podría relacionarse con posiciones de autoridad tradicionalmente atribuidas (Pillon et al., 1992).
- Estas diferencias apoyan parcialmente las teorías clásicas sobre diferencias lingüísticas por género, como las de Lakoff (1975) y Bucholtz (2002), que sugieren marcas lingüísticas distintivas según género.
- Sin embargo, nuestros resultados también revelan una complejidad que desafía enfoques exclusivamente biológicos (Plug et al., 2021).

Referencias bibliográficas

Bucholtz, Mary. 2002. «From "sex differences" to gender variation in sociolinguistics». *University of Pennsylvania Working Papers in Linguistics* 8(3).

Crystal, David. 2013. «Prosodic and paralinguistic correlates of social categories». Pp. 185-206 en *Social anthropology and language*. Routledge.

Lakoff, Robin. 1975. *Language and Woman's Place*. New York: Harper & Row.

Philips, Susan U. 1980. «Sex differences and language». *Annual review of anthropology* 523-44.

Pillon, Agnesa, Catherine Degauquier, y François Duquesne. 1992. «Males' and females' conversational behavior in cross-sex dyads: From gender differences to gender similarities». *Journal of Psycholinguistic Research* 21:147-72.

Plug, Ilona, Wyke Stommel, Peter LBJ Lucassen, Sandra van Dulmen, Enny Das, y others. 2021. «Do women and men use language differently in spoken face-to-face interaction? A scoping review». *Review of Communication Research* 9:43-79.

Conclusiones y perspectivas futuras

! Ideas para indagar

- AMH es un entorno completo para análisis prosódicos detallados.
- Los datos prosódicos solos no bastan para una categorización social precisa.
- Investigaciones futuras: incluir factores pragmáticos, emocionales y contextuales más amplios.