Propuesta computacional de análisis sociolingüístico

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

Copyright y derechos:

Presentación al Congreso XVI CILC2025 por Adrián Cabedo Nebot is licensed under CC BY 4.0 

Introducción

- **Objetivo**: Desarrollar un entorno de análisis de transformación automática (Oralstats) que combine información prosódica, morfológica, léxica y pragmática.
- **Datos**: Corpus PRESEEA-Valencia; 136 hablantes con educación superior, 122744 grupos entonativos (luego realizamos una selección). 24 se recogieron a finales de 1990; el resto del año 2021.
- **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.

Hipótesis y marcos teóricos

Marcos teóricos

- **Perspectiva de análisis variacionista** (Moreno Fernández, Cestero Mancera, López Morales...; PRESEEA)
- **Marcos de análisis prosódico**: AMH (Análisis Melódico del Habla) y TOBI (Tonos e Índices de Ruptura).

Marcos de análisis prosódico

- **Hipótesis**: Existen patrones entonativos (sentimientos, morfosintaxis) específicos asociados al sexo y la edad.

Datos y metodología

- Corpus: Entrevistas PRESEEA (Valencia), todavía no disponibles online. Transcritas y alineadas al audio. El resto del corpus PRESEEA de otras ciudades puede encontrarse, junto con una muestra de Valencia, en https://preseea.uah.es/
- PRESEEA Valencia (esta presentación). 24 hablantes: 8 hombres y 8 mujeres de nivel alto y con tres franjas de edad: 4 x 3 edades (jóvenes, mediana edad, edad más avanzada).

Referencias

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.

Orientación computacional

Orientación computacional del estudio

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

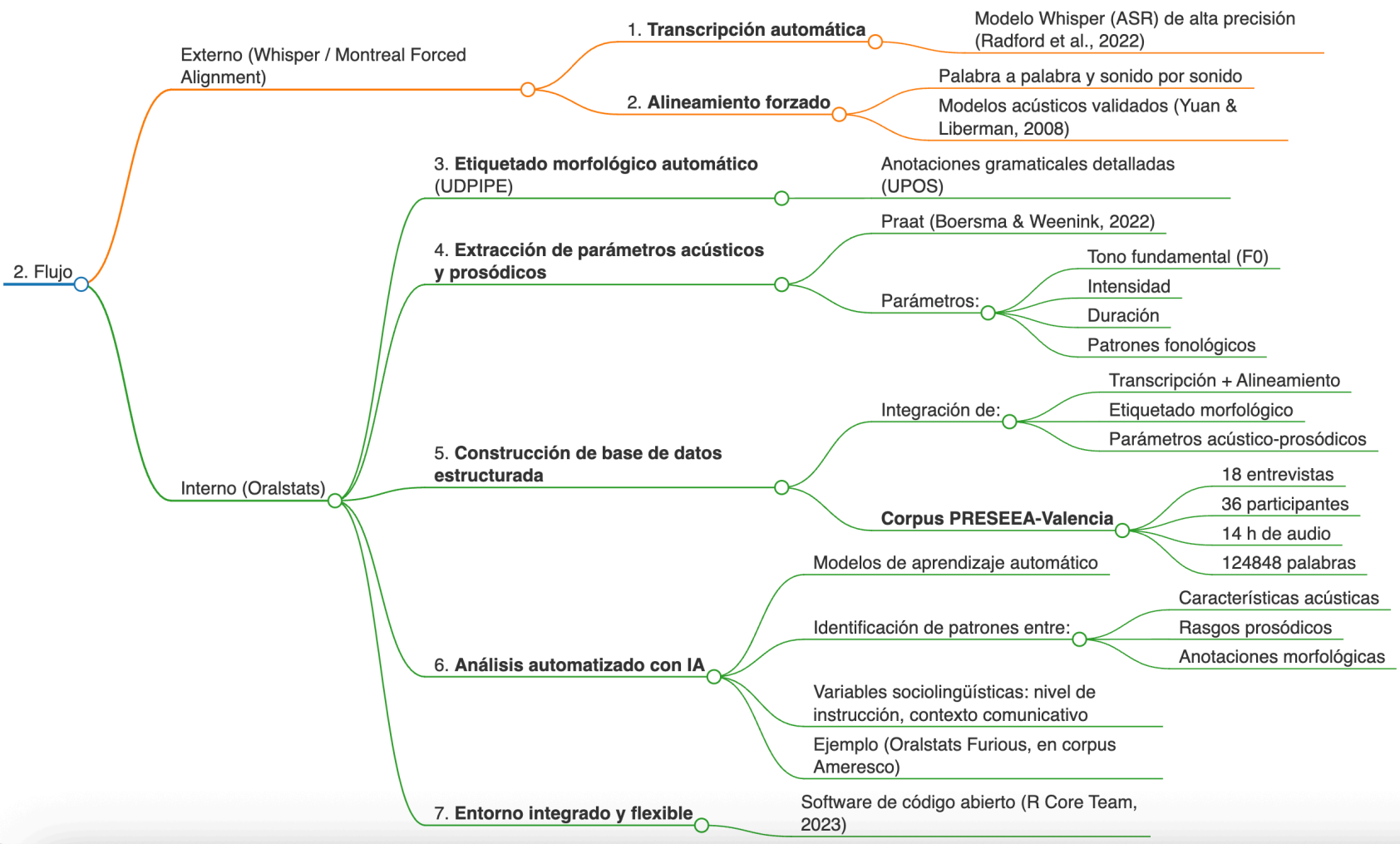
Anthony, Laurence. 2020. «Programming for Corpus Linguistics». Pp. 181-207 en A Practical Handbook of Corpus Linguistics, editado por M. Paquot y S. Th. Gries. Cham: Springer International Publishing.

Scripts de análisis

Oralstats

Versión 1.6 (próximamente en Github). Pueden consultarse versiones anteriores.
https://github.com/acabedo/oralstats

Esquema proceso computacional



Input de entrada

- Archivos json o srt o txt (whisper o whisperX)
- Archivos de pitchtier (PRAAT)
- Archivos de intensitytier (PRAAT)
- Opcional (metadatos para asignar sexo o edad de los hablantes)

Ejemplos input (json de WhisperX)


```
{
  "segments":[
    {
      "start":1.929,
      "end":2.989,
      "text":" Hola, ¿cómo estás?",
      "words":[
        {
          "word":"Hola,",
          "start":1.929,
          "end":2.109,
          "score":0.62,
          "speaker":"SPEAKER_01"
        },
        {
          "word":"¿cómo",
          "start":2.129,
          "end":2.309,
          "score":0.621,
          "speaker":"SPEAKER_01"
        },
        {
          "word":"estás?",
          "start":2.349,
          "end":2.989,
          "score":0.786,
          "speaker":"SPEAKER_01"
        }
      ],
      "speaker":"SPEAKER_01"
    },
    {
      "start":3.049,
      "end":4.11,
      "text":"Muy bien, ¿y tú?",
      "words":[
        {
          "word":"Muy",
```

Ejemplos input (headersheet csv de pitchtier)

1990_alto_01.PitchTier

1	4.0103061224489194	382.95152637871269
2	4.0203061224489192	382.2892371023338
3	4.0303061224489189	373.03970722971582
4	4.0803061224489197	425.89824602791703
5	4.0903061224489194	424.71296628695836
6	4.1003061224489192	460.34337890641427
7	4.110306122448919	446.89945023983915
8	4.6203061224489197	75.397689583991394
9	4.7603061224489194	84.367722687127355
10	4.7703061224489192	89.790285962742004
11	4.7803061224489189	90.416200085602711
12	4.8103061224489192	109.86289726414644
13	4.820306122448919	110.60541839715962
14	4.8303061224489197	110.91784461132241
15	4.8403061224489194	110.327555984082
16	4.8503061224489192	109.56604464164731
17	4.860306122448919	109.05746340825615
18	4.8703061224489197	109.45915350274986

Ejemplos input (headersheet csv de intensitytier)

1990_alto_01.IntensityTier

rowLabel	Time (s)	Intensity (dB)
?	0.0333061224904296	41.125240612226285
?	0.0413061224904296	37.7316091150997 86
?	0.0493061224904296	30.621426741414076
?	0.05730612244904296 4	24.96390867329518
?	0.0653061224904296	23.918283555892405
?	0.0733061224904297	26.578216737230335
?	0.0813061224904296	24.32044782353678
?	0.0893061224904296	24.2943608818128
?	0.0973061224904296	25.886879732537473
?	0.1053061224904297	27.481531018739517
?	0.1133061224904296	29.44742043765871
?	0.1213061224904296	28.879850745593806
?	0.1293061224904297	27.261618323096094
?	0.1373061224904297	25.967276253636648
?	0.1453061224904295	24.049138539877 54
?	0.1533061224904296	23.418532772795075
?	0.1613061224904297	22.48451825309992
?	0.1693061224904297	20.29932935940542

Bases de datos creadas

Perspectiva colapsable, las unidades mayores se generan a partir de las inferiores.

- Ips
- Words
- Phones

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		

Entorno Shiny

https://adrin-cabedo.shinyapps.io/preseeasample2025/