# Biblioteca de Babel

Revista de Filología Hispánica

Volumen extraordinario 2

2024

Los corpus orales como fuente de investigación en el habla coloquial del español

Mar Capilla Martín (coord.)

Editado por Biblioteca de Babel, Madrid, 2024.
Con el apoyo del Departamento de Filología Española de la UAM.

**ISSNe:** 2695-6349
**DOI:** https://doi.org/10.15366/bibliotecababel2024.extra2.

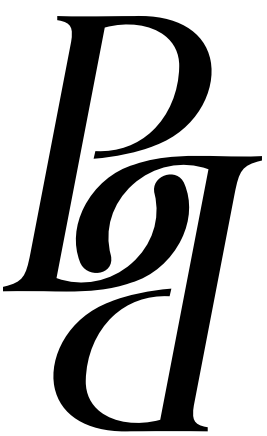
C/ Francisco Tomás y Valiente, 1
Módulo IV, 3.14
Facultad de Filosofía y Letras
Universidad Autónoma de Madrid
Ciudad Universitaria de Cantoblanco, 28049, Madrid.

https://revistas.uam.es/bibliotecababel
revista.biblioteca.babel@uam.es

Diseño original: Juan Cerezo Soler
Monograma: Ernesto Valerio
Maquetación: Aldar Ferrera-Lagoa

# Biblioteca de Babel

## Revista de Filología Hispánica



Volumen extraordinario 2

&

2024

Los corpus orales como fuente de investigación
en el habla coloquial del español

Mar Capilla Martín (coord.)

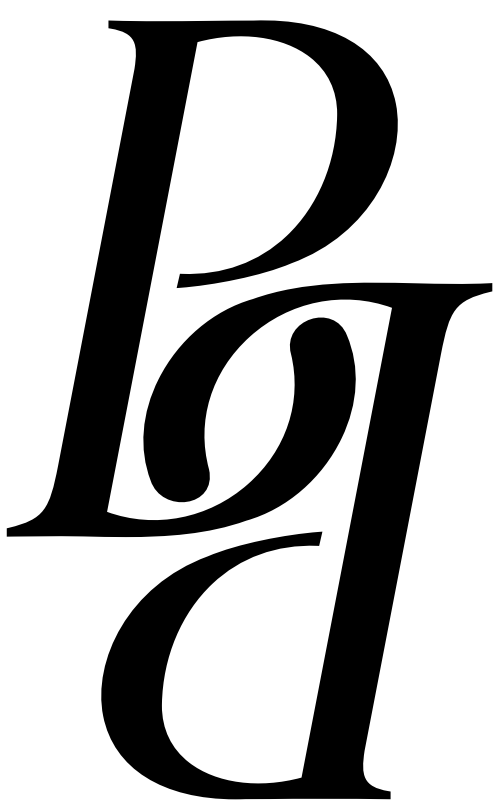
# Índice

## Introducción



## Artículos

# Introducción



# Introducción

Mar Capilla Martín
*Centro de Estudios de la Real Academia Española*
mcapilla@rae.es

Es para mí un gran honor presentar este monográfico dedicado a los corpus orales, un proyecto que me ha hecho especial ilusión coordinar debido a mi vínculo directo con el tema que aborda. Mi experiencia en el Corpus del Español del Siglo XXI (CORPES XXI), centrada en la selección y codificación de textos escritos y orales, me permite trabajar de cerca con los corpus lingüísticos. Por ello, dedicar este número de *Biblioteca de Babel*, revista del departamento de Filología Española de la Universidad Autónoma de Madrid, a los corpus, y especialmente a los corpus orales, supone un auténtico privilegio.

En el ámbito del español, los corpus lingüísticos constituyen una herramienta fundamental para el análisis de la oralidad, ya que facilitan un estudio sistemático y empírico del uso de la lengua en contextos reales de comunicación. A través de los datos recopilados en los corpus, es posible identificar y examinar fenómenos lingüísticos característicos del discurso oral que, mediante métodos tradicionales, resultarían difíciles de observar y analizar. Considerando la importancia de los corpus en la investigación lingüística y su notable desarrollo en las últimas décadas, resulta razonable que el tema central del monográfico haya sido precisamente el empleo de corpus orales para la investigación del habla coloquial en español. De esta manera, este monográfico reúne un conjunto de investigaciones que abordan, desde perspectivas teóricas, metodológicas y aplicadas, hasta retos y avances en el estudio de corpus orales. Estas contribuciones subrayan la riqueza de los datos empíricos para la investigación lingüística, así como el impacto de las nuevas tecnologías y metodologías en el análisis del discurso oral.

A lo largo de este monográfico, especialistas de diferentes áreas de la lingüística se encargarán de resaltar la relevancia de los corpus orales en las investigaciones lingüísticas centradas en el habla coloquial. Tenemos el honor de incluir una variedad de artículos que abordan enfoques diversos. El volumen comienza con un marco conceptual general y una exposición de las bases metodológicas necesarias para trabajar con corpus. A continuación, se presentan estudios centrados

en el análisis pragmático-discursivo, que reúnen investigaciones sobre el uso de marcadores pragmáticos desde perspectivas variadas, conectando estudios específicos con cuestiones más amplias. Finalmente, el monográfico concluye con una serie de artículos dedicados a corpus orales específicos, que ofrecen una perspectiva transversal e interesante desde distintos enfoques. La sección de artículos de este volumen se inaugura con el trabajo de Renata Enghels, profesora en la Universidad de Gante y codirectora del grupo de investigación CROS (Crossing the Border between English and Spanish), Marlies Jansegers, profesora en la Universidad de Gante y Nele Van Den Driessche, estudiante de doctorado en la misma universidad. Su contribución, titulada «Reflexiones metodológicas y teóricas sobre el análisis de marcadores pragmáticos», se centra en los desafíos teóricos y metodológicos asociados al estudio de estos marcadores. En particular, analiza el caso de *es que* como marcador pragmático polifuncional en el habla madrileña, utilizando una muestra representativa del CORMA (Corpus Oral de Madrid).

A continuación, encontramos el artículo titulado «La transcripción de los elementos prosódicos en corpus de habla coloquial espontánea» presentado por Antonio Hidalgo Navarro, catedrático de Lengua Española en la Universidad de Valencia, y Carlos Castelló Vercher, investigador de la Universidad de Valencia, y ambos miembros del grupo Val.Es.Co. Este estudio aborda la relevancia de la prosodia en las interacciones coloquiales, proponiendo un sistema de transcripción basado en el modelo Val.Es.Co., el cual se apoya en herramientas tecnológicas como ELAN y Praat para identificar fenómenos prosódicos específicos. Otro trabajo basado en el corpus Val.Es.Co. es el propuesto por Sara Badia Climent, profesora de la Universidad de Valencia, y Salvador Pons Bordería, catedrático de la misma institución, ambos integrantes del grupo Val.Es.Co. Con el título «De la transcripción al análisis: desarrollos técnicos del corpus Val.Es.Co. 3.0.», esta investigación revisa los principales corpus orales del español, en especial del corpus Val.Es.Co. 3.0., y profundiza en los avances técnicos para la construcción de corpus orales accesibles digitalmente, poniendo especial énfasis en los retos asociados a la segmentación y el etiquetado automatizado en el procesamiento automatizado de los datos. Por último, Marta Garrote Salazar, profesora de la Universidad Autónoma de Madrid, bajo el título «El etiquetado gramatical automático en el procesamiento del habla coloquial», presenta las dificultades que plantea el etiquetado gramatical, o *part-of-speech tagging*, en los textos de habla coloquial y revisa el estado actual de la investigación en este campo, identificando posibles soluciones o mejorad del etiquetado gramatical en contextos coloquiales.

En relación con los marcadores discursivos, este monográfico reúne diversos estudios que abordan su análisis desde perspectivas

variadas. Dos de las contribuciones se centran específicamente en el estudio de *un poco*. En el artículo titulado «Los corpus orales en la investigación pragmática: el caso de la locución *un poco*», Beatriz Méndez Guerrero, profesora de la Universidad Autónoma de Madrid, examina los usos y las variantes semánticas y pragmáticas de este marcador a partir de datos extraídos de los corpus orales PRESEEA-Palma, Val. Es.Co. y COJEM. Por otro lado, en su trabajo «*Un poco* como recurso pragmático-discursivo en corpus orales de nativos y de aprendientes de español: un estudio comparado», Marta Blanco Domínguez, profesora de la Universidad de Santiago de Compostela, y María Eugenia Conde Noguerol, profesora de la Universidad de La Coruña, realizan un análisis contrastivo del uso de este marcador pragmático entre hablantes nativos y aprendientes de español, basándose en los corpus ESLORA y el Spanish Learner Language Oral Corpora (SPLLOC).

Además, el volumen incluye otras investigaciones relacionadas con los marcadores discursivos. Entre ellas, destaca el estudio de Florencio del Barrio de la Rosa, profesor titular en la Universidad de Venecia Ca' Foscari, titulado «*Ni tan mal*: un operador argumentativo en el español del siglo xxi». Este trabajo caracteriza gramatical y discursivamente dicho operador, apoyándose en corpus electrónicos sincrónicos de lengua hablada, coloquial y subestándar, como el Corpus del español, CORPES XXI y EsTenTen18, para analizar su función y su evolución hacia un marcador conversacional. Asimismo, se presenta el estudio de Mar Capilla Martín, filóloga en el CORPES XXI en el Centro de Estudios de la Real Academia Española, titulado «Entre lo oral y lo escrito: un estudio de corpus sobre los reformuladores *es decir*, *o sea* y *en plan* en español». En este artículo, la autora examina las diferencias contextuales en el uso de dichos reformuladores, destacando nuevos enfoques en el análisis de marcadores discursivos gracias a la integración de datos provenientes de corpus y subcorpus orales.

Seguidamente, se exponen tres artículos que se centran en el análisis de corpus lingüísticos específicos como herramientas para la investigación en el ámbito de la lingüística. El primero, titulado «Un acercamiento al estudio paralingüístico de las emociones en el corpus PRESEEA Valencia», ha sido elaborado por Adrián Cabedo Nebot, catedrático de la Universidad de Valencia, y Noelia Ruano, doctoranda en la misma universidad. Este trabajo examina la influencia de las emociones en el discurso y la prosodia a través del corpus PRESEEA-Valencia, que reúne datos sociolingüísticos del español hablado en Valencia (España).

Por su parte, Carles Navarro-Carrascosa, profesor en el University College Dublin, presenta en su artículo «Descripción y análisis de un corpus de español oral de la comunidad de habla LGTBI» un análisis

del Corpus Oral de la Comunidad de Habla LGTBI. Este corpus recoge muestras de habla propias de dicha comunidad, lo que facilita el estudio de sus códigos lingüísticos y prácticas comunicativas. Además, el autor propone diversas mejoras destinadas a optimizar el corpus para investigaciones más detalladas sobre variación lingüística, influencias socioculturales y la evolución del lenguaje en contextos LGTBI.

La tercera contribución es obra de Carlota de Benito Moreno, profesora de la Universidad Autónoma de Madrid, Antonio Corredor Aveledo, investigador posdoctoral en la Universidad de Neuchâtel, y Elena Padrón Castilla, profesora en la Universidad de Neuchâtel, quienes forman parte del proyecto RurICan en la Universidad de Zúrich (UFSP Sprache und Raum). En su artículo «Heterogeneidad e innovación en la isla de La Palma: aproximación sociolingüística y dialectal», los autores exploran diversos fenómenos gramaticales utilizando los datos del corpus RurICan. Este estudio busca describir el grado de heterogeneidad e innovación en las hablas de la isla de La Palma, atendiendo a parámetros tanto sociolingüísticos como geográficos.

Finalmente, Adriana Raya Palmer, profesora en el University College Dublin, presenta su artículo «El lenguaje juvenil en la ficción traducida: una comparación basada en corpus de una selección de marcadores pragmáticos en catalán y español». La autora pone de relieve las particularidades del habla juvenil en las traducciones literarias, explorando las diferencias en la representación de este registro. Además, el estudio investiga la frecuencia y distribución de dichos marcadores en catalán y español, utilizando como base un corpus paralelo de diálogos traducidos de novelas contemporáneas y dos corpus orales, lo que permite una aproximación tanto lingüística como sociocultural.

En definitiva, este monográfico ofrece una mirada amplia y profunda sobre la utilización de los corpus orales para la investigación del habla coloquial, abordado desde diversas perspectivas gracias a la colaboración de expertos en la materia. A lo largo de los capítulos, los lectores encontrarán un análisis riguroso y enriquecedor que no solo desglosa los aspectos clave del tema, sino que también aporta nuevas reflexiones y enfoques. Con la participación de los distintos autores, cuyas contribuciones han sido fundamentales, este trabajo aspira a convertirse en una referencia imprescindible para quienes deseen profundizar en el tema. Confiamos en que este recorrido cumpla las expectativas y sirva de punto de partida para futuros debates y estudios.

# Artículos

Renata Enghels (ORCID: 0000-0002-7785-0009), Marlies Jansegers (ORCID: 0000-0003-4205-9126) y Nele Van Den Driessche (ORCID: 0000-0001-5508-8343), «Reflexiones metodológicas y teóricas sobre el análisis de marcadores pragmáticos: ilustraciones a través del estudio de *es que*». *Biblioteca de Babel: Revista de Filología Hispánica,* vol. extra. 2 (2024), pp. 19-50.
**DOI**: 10.15366/bibliotecababel2024.extra2.001 · **ISSNe**: 2695-6349 · Recibido: 02/09/2024 · Aceptado: 28/10/2024

# Reflexiones metodológicas y teóricas sobre el análisis de marcadores pragmáticos: ilustraciones a través del estudio de *es que*

Renata Enghels
*Universiteit Gent*
renata.enghels@ugent.be

Marlies Jansegers
*Universiteit Gent*
marlies.jansegers@ugent.be

Nele Van Den Driessche
*Universiteit Gent*
nele.vandendriessche@ugent.be

**Resumen**: Aunque los marcadores pragmáticos fueron considerados una categoría lingüística marginal hasta finales de los años 80, su estudio ha ganado considerable atención en las últimas décadas. No obstante, el análisis de sus funciones pragmáticas conlleva múltiples desafíos. Estos incluyen la elección entre enfatizar macro o microcategorías funcionales, decidir entre un enfoque semasiológico u onomasiológico, abordar su polifuncionalidad en contextos específicos y establecer criterios formales para identificar funciones pragmáticas concretas. Este trabajo tiene como objetivo principal explorar estas opciones teóricas y metodológicas. Se ejemplifica mediante un estudio de caso del marcador *es que*, tal y como se observa en el habla coloquial de Madrid. Utilizando una muestra representativa del corpus CORMA (Corpus Oral de Madrid), se argumenta que *es que* actúa como un marcador pragmático polifuncional con significado procedimental, cuya interpretación es moldeada por el contexto, y cuyo análisis requiere un enfoque multidimensional.

**Palabras clave**: marcador pragmático, español coloquial, polifuncionalidad, *es que*, Corpus Oral de Madrid.

# Methodological and theoretical reflections on the analysis of pragmatic markers: illustrations through the study of *es que* ('it is that')

**Abstract**: Although pragmatic markers were considered a marginal linguistic category until the late 1980s, their study has gained considerable attention in recent decades. However, the analysis of their pragmatic functions involves multiple challenges. These include choosing between emphasizing macro or micro functional categories, deciding between a semasiological or onomasiological approach, addressing their polyfunctionality in specific contexts, and establishing formal criteria for identifying concrete pragmatic functions. The main objective of this study is to explore these theoretical and methodological options. This is illustrated through a case study of the marker *es que* ('it is that'), as observed in the colloquial speech of Madrid. Using a representative sample from the CORMA corpus (Oral Corpus of Madrid), it is argued that *es que* acts as a polyfunctional pragmatic marker with procedural meaning, whose interpretation is shaped by context, and whose analysis requires a multidimensional approach.

**Keywords**: pragmatic marker, colloquial Spanish, polyfunctionality, *es que*, Corpus Oral de Madrid.

## 1. Introducción

El español coloquial informal se distingue por una serie de características fonéticas, léxicas y gramaticales que lo diferencian de variedades más formales. Además, para entender su funcionamiento, es imprescindible destacar su naturaleza pragmática subjetiva e intersubjetiva (Ghezzi 2014). Efectivamente, suele incluir elementos altamente expresivos, que refuerzan el mensaje y dan cuenta de una mayor implicación en el acto comunicativo por parte del emisor, mientras que su alto grado de intersubjetividad da cuenta de los vínculos entre los hablantes y refuerzan la identidad de grupo. En esta compleja construcción de relaciones (inter)personales, los denominados *marcadores pragmáticos* desempeñan un papel fundamental.

La definición y terminología asociada con el concepto de marcador pragmático es extensamente debatida. En el presente estudio, partimos de la definición de Brinton (1996) y Fraser (1999) según la cual se trata de elementos lingüísticos altamente multifuncionales que tienen un significado núcleo procedimental (y no condicionado por la verdad [Brinton 2008]) cuya interpretación específica es negociada por el

contexto[1]. En las últimas tres décadas, se ha observado un creciente interés académico que debería contribuir a una ampliación significativa del conocimiento en esta área lingüística. Una revisión preliminar de la bibliografía revela efectivamente que se han explorado diversos aspectos sobre temas como el desarrollo histórico y el funcionamiento actual de marcadores específicos, la contribución de su uso en los procesos de procesamiento y producción lingüística, y su relevancia en la enseñanza del idioma. Sin embargo, tras generaciones de investigación, los marcadores pragmáticos continúan fascinando a los lingüistas, lo que algunos han considerado paradójico (Crible y Pascual 2020; Degand 2016). Parte de este interés se debe a su gran variedad y ambivalencia formal, pero ante todo funcional.

Considerando la multifuncionalidad y la naturaleza híbrida tanto de los marcadores como de los estudios realizados sobre ellos, el presente trabajo pretende aportar nuevas perspectivas a la discusión sobre la viabilidad de continuar clasificándolos funcionalmente. En efecto, numerosas dificultades teóricas y metodológicas persisten en su análisis. En términos generales, la semántica y la pragmática, disciplinas que abordan el significado y el uso del lenguaje, a menudo desafían su operacionalización debido a la abstracción de los conceptos que manejan. Sin embargo, la orientación hacia el giro empírico ha llevado a estos campos a adoptar metodologías que permiten un análisis basado en datos más objetivos. Frente a metodologías convencionales centradas en la introspección, el enfoque del ciclo empírico (p. ej. Geeraerts 2010) privilegia la observación y la experimentación como medios para articular una definición de significado más estructurada y fundamentada. No obstante, cuando nos centramos en los marcadores pragmáticos, cuyo significado no es referencial sino procedimental, la complejidad aumenta. Ya sabemos que no se refieren directamente a entidades o estados del mundo, sino que guían la interpretación del discurso, estructurando la interacción y gestionando la relación entre interlocutores. Por lo tanto, su significado es intrínsecamente dependiente del contexto y de la interpretación que los usuarios del lenguaje hacen en el discurso concreto. Esto plantea la cuestión de cómo metodologías como el análisis de corpus pueden capturar (incluso cuantificar) adecuadamente la función de tales elementos lingüísticos, que están profundamente arraigados en la dinámica del discurso coloquial.

[1] Sin entrar en detalle, existe una variedad de términos utilizados que reflejan diferentes matices y perspectivas en el estudio de estos elementos. Algunos de los más conocidos incluyen *marcador discursivo* (o *discourse marker*, por ejemplo en Schiffrin 1987), *partícula discursiva* (*discourse particle*, Aijmer 2002), y *partícula pragmática* (*pragmatic particle*, Östman 1995). Foolen (2011) propone el uso de *marcador pragmático* como el término más neutro. A diferencia de *partícula*, no sugiere necesariamente una forma pequeña, como se ejemplifica en expresiones como *you know* en inglés (*sabes*) o *es que*, objeto de análisis del presente estudio. Además, el término *pragmático* es neutro con respecto al medio y cubre aspectos no verbales de la interacción.

El presente estudio se concentra específicamente en abordar las cuatro cuestiones siguientes:

1. ¿Hasta qué nivel de detalle es posible y preferible definir la función de los marcadores pragmáticos? ¿Es más efectivo emplear una extensa lista de microfunciones específicas o convendría limitarse a un conjunto reducido de macrofunciones generales?

2. ¿Resulta más efectivo enfocar el estudio desde una perspectiva semasiológica, que parte de las formas lingüísticas para analizar su uso en el discurso, o desde un enfoque onomasiológico, que identifica primero las funciones en el texto y luego examina las formas que los hablantes utilizan para expresarlas, considerando diferentes rasgos contextuales y discursivos?

3. ¿En qué medida es posible establecer criterios objetivos para identificar la función pragmática de un marcador, dado que las categorías funcionales a menudo presentan límites difusos y continuos? Este rasgo parece plantear un desafío para los métodos cuantitativos que requieren categorías claramente delimitadas.

4. ¿Cuáles son los criterios adecuados para definir *función* en contextos donde esta puede variar significativamente? Específicamente, ¿cómo se puede desarrollar una clasificación funcional de los marcadores pragmáticos que incorpore diversos niveles de análisis, tales como el valor expresivo, intersubjetivo y metadiscursivo, así como consideraciones de cortesía, atenuación e intensificación?

Abordaremos los cuatro dilemas planteados anteriormente mediante el examen detallado del marcador pragmático *es que* en el Corpus Oral de Madrid (CORMA), un corpus que incluye conversaciones espontáneas grabadas en Madrid entre 2016 y 2019. Además, dentro del paradigma de los marcadores pragmáticos, el elemento *es que* constituye un fenómeno lingüístico fascinante pero relativamente inexplorado (excepción hecha de Fuentes Rodríguez 1997, 2015; Remberger 2020; Van Den Driessche y Enghels 2025, en prensa). Como muestran los ejemplos (1a-c)[2], cumple funciones muy diversas, tal y como la función de justificación (1a). En este caso, el uso de *es que* indica que seguirá una razón por la que la hablante IR2F18[3] quiere comprarse nueva ropa.

[2] Nótese que es posible que *es que* exprese varias funciones a la vez. Así, en el ejemplo (1c), al lado de la función atenuadora que comentamos, también se observa un valor justificativo. Abordamos la idea de la polifuncionalidad de los marcadores en la sección 3.4.

[3] El código del hablante refiere a la situación de la conversación o el centro de enseñanza, la edad, el sexo y la intervención del hablante. En el caso del hablante IR2F18, IR refiere al instituto; se trata de un participante que pertenece a la segunda generación (2), es una mujer (F = sexo), y es la participante número 18 con este perfil sociolingüístico que participa en la conversación. Luego la conversación IR_AM2_F_09 es la conversación número 9 (09) grabada en el instituto IR entre amigas (AM, F) de la segunda generación (2). Para obtener una descripción detallada de los códigos se recomienda consultar Enghels, De Latte y Roels (2020).

El ejemplo (1b), por su parte, ilustra el uso de *es que* como partícula de relleno. La hablante aún no tiene claro lo que desea expresar, lo cual se evidencia en las palabras aisladas (*ya – eh – es que*) que preceden su enunciado (*pues la verdad que lo fui* […]). Para ganar tiempo y organizar sus ideas, utiliza *es que* para rellenar el discurso. El marcador *es que* también opera en el campo de la atenuación. En (1c), IR2F1 recurre a *es que* para evitar un posible daño a su imagen y evitar expresar un desacuerdo de manera demasiado explícita.

(1) a. IR2F18: Tío quiero comprarme ropa más bonita, *es que* mi ropa es una puta mierda. En serio yo no entiendo cómo la gente se viste tan bien y yo me visto como el culo. (CORMA: IR_AM2_F_09)
b. IIC2F7: Ya- eh- *es que*- pues la verdad que lo fui como deduciendo cuando estuvo dando la charla (CORMA: IIC_AM2_F_03)
c. IR2F2: Ugh o máh rico es la hamburguesa tía está puto buenísima
IR2F1: *Es que* no me apetece tía hamburguesa, nada. (CORMA: IR_AM2_F_02)

A través del análisis de *es que* pretendemos ilustrar que su estudio puede desembocar en distintas trayectorias analíticas en función de la perspectiva seleccionada. Ello nos permitirá no solamente comprender las diversas funciones y aplicaciones de *es que* en el español coloquial actual, sino también explorar la validez y aplicabilidad de las distintas aproximaciones teóricas y metodológicas frente a las cuatro tipos de ambigüedades inherentes a este tipo de marcadores que se han planteado antes[4].

El estudio se estructura de la siguiente manera: la sección 2 se dedica a la exposición de los datos y a la metodología empleada, proporcionando más informaciones sobre el corpus CORMA y el procedimiento adoptado para la selección de los datos pertinentes de *es que*. La sección 3 profundiza en los desafíos teóricos y metodológicos inherentes al análisis de los marcadores pragmáticos, discutiendo en más detalle las complejidades y perspectivas de su estudio. Al mismo tiempo se presentan los resultados obtenidos en el marco del análisis empírico de *es que*. En la sección 4 se sintetizan los hallazgos principales y se reflexiona sobre su relevancia e implicaciones para la investigación futura en el campo de los marcadores pragmáticos.

[4] Este artículo adopta principalmente un enfoque teórico y metodológico. Para obtener información detallada sobre *es que* y los resultados del análisis, se recomienda consultar Van Den Driessche y Enghels (en prensa) y otros trabajos futuros de las autoras.

## 2. Datos y metodología

### *2.1. Estudiar el español coloquial a través del corpus CORMA*

El Corpus Oral de Madrid (CORMA) busca documentar el español conversacional espontáneo tal como se utiliza en Madrid en la época actual (Enghels, De Latte y Roels 2020). Desarrollado por el equipo de lingüística española de la Universidad de Gante en colaboración con la UNED, CORMA es una colección de interacciones orales espontáneas grabadas de la vida cotidiana. Se define por su diversidad y representa conversaciones lingüísticas en actividades cotidianas con una amplia variación situacional y sociolingüística. La variación incluye participantes de ambos sexos, distintas generaciones y niveles socioculturales, en conversaciones orales y espontáneas que implican diálogos cooperativos y dinámicos. Contiene conversaciones entre amigos, conocidos y familiares y también diálogos en situaciones de atención al cliente. La recogida de datos se realizó en tres campañas de trabajo de campo entre 2016 y 2019 y consta de 106 conversaciones entre 485 hablantes madrileños, dando como resultado 57 horas de grabación y 469 860 palabras transcritas[5].

Por lo tanto, CORMA no solo proporciona una rica fuente de datos para el estudio del español coloquial contemporáneo, sino que también ofrece un marco para analizar los marcadores pragmáticos y su función en un contexto lingüístico auténtico y variado. Con su enfoque en las interacciones cotidianas y la amplia gama de hablantes involucrados, es ideal para estudiar la función y el uso de los marcadores pragmáticos en diferentes contextos situacionales. Además, el detalle de las transcripciones facilita una comprensión profunda de los matices pragmáticos y la variabilidad contextual de los marcadores, incluso para manejar las relaciones interpersonales como la cortesía, la atenuación y otros aspectos del discurso.

### *2.2. Composición del muestreo del marcador* es que

Para el estudio del marcador *es que*, seleccionamos una muestra de 48 conversaciones en CORMA, que dio como resultado un subcorpus de 220 468 palabras. Se trata de datos sociolingüísticos variados: los 239 hablantes masculinos y femeninos pertenecen a diferentes generaciones (adolescentes y adultos). Debido a la dificultad para determinar la clase social de los participantes, no se ha tomado en cuenta este parámetro. La tabla 1 proporciona la información sobre los participantes y su perfil sociolingüístico.

[5] Para acceder al CORMA, los interesados pueden consultar la página www.corma.ugent.be.

| | Jóvenes (12-25) | Adultos (25+) | Total |
|---|---|---|---|
| **Hombre** | 31 | 57 | **88** |
| **Mujer** | 75 | 76 | **151** |
| **Total** | **106** | **133** | **239** |

*Tabla 1. Perfil sociolingüístico de los participantes.*

A partir de una lectura detenida y una búsqueda manual en las conversaciones, se llegó a un muestreo de 1 474 ocurrencias. Asimismo, el análisis se centra en el uso de *es que* como estructura gramaticalizada. Se excluyeron por tanto casos como (2) en que *es que* se acerca al uso de dos segmentos separados, es decir, el verbo *ser* y la conjunción *que*, y ejemplos como (3) en que el verbo *ser* no está en presente de indicativo.

(2) IR2F9: […] Bueno, *el caso es que* yo había queda'o con estas que te dije y también estábamos con Aragón, Gallego […]. (CORMA: IR_AM2_F_05)

(3) CPEL4F12: No, *será que* aquel día que coincidió (()) no me no me llamó mucho la atención. (CORMA: ATpel.01)

Por la extensa tarea de anotación funcional-pragmática[6], seleccionamos de manera aleatoria un muestreo de 200 ejemplos de *es que*. Por un lado, estos 200 casos se han sometido a un análisis que incluye dos variables sociolingüísticas: (a) la generación (jóvenes vs. adultos), (b) el sexo (masculino vs. femenino). Por otro, analizamos una amplia gama de variables lingüísticas, formales y funcionales como: (c) la posición de *es que* en el acto de habla (periferia inicial, posición media, periferia derecha, posición independiente), (d) la posición de *es que* en la intervención (inicial, media, final, independiente), (e) la presencia de una colocación (p. ej. *pero es que*, *yo es que*, etc.), (f) el tipo de elemento con el que *es que* forma una colocación (p. ej. marcador discursivo, pronombre personal, etc.), (g) la función de *es que*, como criterio principal. A lo largo de los apartados siguientes, se ofrecerá una descripción más detallada de estas variables, así como de los desafíos que implican su clasificación y aplicación.

[6] Se excluyeron del análisis funcional casos difícilmente interpretables por la ininteligibilidad de las palabras que preceden o siguen a *es que* (i) o por el hecho de que se interrumpe al hablante (ii).

(i) MAM2F2: Qué hijo de puta tío, de verdad, *es que* (()) (CORMA: M_AM2_01)

(ii) IR2F19: Un viernes Siempre ponen las actividades los viernes, yo no entiendo, ¿qué quieren?
IR2F20: *Es que*
IR2F19: ¿arruinarnos los-los viernes? (CORMA: IR_AM2_F_09)

## 3. Desafíos teóricos y metodológicos para el análisis de los marcadores pragmáticos

### *3.1. El equilibrio entre la especificidad y la generalización: microfunciones frente a macrofunciones*

El primer dilema que atraviesa la bibliografía sobre los marcadores pragmáticos concierne al nivel de precisión con el que se pueden definir sus funciones. Se distinguen dos estrategias en el ámbito de estudio. Por un lado, los lingüistas se enfocaron en comprender su funcionamiento casi desde una perspectiva de funciones primitivas, esto es, con un alto grado de especificidad. Sin embargo, ante la complejidad de delinear estas microfunciones en corpus extensos, se percibe una tendencia hacia la formulación de un número más limitado de macrofunciones que abarcan de manera general los diversos usos de estos marcadores en el discurso. Este dilema refleja una tensión entre la especificidad y la generalización en el estudio de los marcadores pragmáticos.

Como ilustración de este enfoque, Brinton (1996, 2008) distingue entre las macrofunciones textual e interpersonal. En su modelo, la función textual comprende las siguientes microfunciones: reclamar la atención del oyente, iniciar y finalizar un turno (o la conversación), continuar el turno o el discurso, marcar límites y/o cambios de tema y reparar el discurso. La macrofunción interpersonal comprende microfunciones subjetivas tales como expresar respuestas, reacciones, actitudes y comprensión, así como funciones interactivas como expresar intimidad, cooperación, conocimiento compartido o acciones de protección de imagen (cortesía). Este enfoque metodológico refleja la reconocida dificultad de identificar y describir exhaustivamente todas las connotaciones de los marcadores, dado que la cantidad de microfunciones varía tanto como los contextos de uso en los que se aplican (Martín Zorraquino y Portolés 1999).

Es importante observar que los límites que los autores establecen entre las macrofunciones no siempre coinciden. A diferencia de Brinton, López Serena (2011) no identifica dos sino tres macrofunciones, a saber, las funciones interaccional, metadiscursiva y cognitiva. La función interaccional se considera esencial para la comprensión del carácter dialógico del discurso coloquial. Incluye la indicación de los movimientos conversacionales de los interlocutores, tales como la toma, el mantenimiento o la cesión del turno de habla, el control de la recepción, las respuestas reactivas, el desacuerdo y las peticiones de aclaración. Se puede observar que ciertos valores, como la toma o la cesión del turno de habla, se categorizan bajo la función textual según la clasificación de Brinton. Luego, la función metadiscursiva es la que

aporta cohesión al discurso, relacionando adecuadamente las diversas partes del diálogo entre sí y permitiendo a los interlocutores seguir el hilo de la conversación. Finalmente, la función cognitiva se describe como la macrofunción más compleja, que resalta las relaciones lógicas y argumentativas dentro del texto. Esta función no solo conecta los contenidos, conocimientos o presupuestos compartidos por los participantes, sino que también activa distintos mecanismos de deducción e inducción. Adicionalmente, establece un puente entre el contenido textual y la actitud del hablante respecto a lo dicho.

En la misma línea de ideas, en lugar de abordar la amplia gama de valores sugeridos en la literatura para marcadores derivados de verbos de movimiento como *anda*, *vaya* o *venga*, Tanghe (2015) sugiere centrarse en el análisis de tres macrofunciones: apelativa, expresiva y metadiscursiva. Su análisis establece una diferenciación entre funciones que se asocian primordialmente ya sea con el emisor o con el receptor. Así todos los matices emotivos, como sorpresa, incredulidad, énfasis o evaluación, se agrupaban en la categoría expresiva.

Además de las diversas clasificaciones, este método de macrocategorización presenta tanto ventajas como desventajas. La principal ventaja es que permite sistematizar la variedad de valores que puede adoptar una forma lingüística. En este sentido, Ghezzi (2014) sostiene que, aunque no siempre es posible diferenciar claramente los distintos planos discursivos, es esencial para propósitos heurísticos explicar los diferentes valores como un conjunto de macrofunciones, especialmente en el contexto del *giro cuantitativo* en la lingüística. No obstante, al mismo tiempo, surge la pregunta de si esta generalización no lleva a una pérdida de comprensión sobre el funcionamiento específico de los marcadores pragmáticos individuales. Efectivamente, estudios anteriores han identificado las mismas macrofunciones —expresiva/modal, intersubjetiva y metadiscursiva/textual— en varios marcadores, como *sabes* (Azofra Sierra y Enghels 2017; Enghels y Azofra Sierra 2018), *nada* (Azofra Sierra y Enghels 2022; Enghels y Azofra Sierra 2024) y *en plan* (De Smet y Enghels 2020). Así se plantea la cuestión de si realmente estamos ante marcadores con perfiles tan similares.

El dilema radica entonces en decidir entre describir las microfunciones, con el riesgo de no alcanzar una lista exhaustiva debido a la polifuncionalidad de los marcadores y la dificultad de tomar decisiones precisas en un corpus de datos empíricos, o bien optar por trabajar con categorías más amplias que faciliten la clasificación de los marcadores, pero corriendo el riesgo de no identificar suficientemente la particularidad de cada marcador.

La implicación concreta de esta elección metodológica se ilustra mediante la partícula *es que*, cuyo análisis funcional dio, en una

primera fase, una lista extensa de 17 microfunciones: (i) justificación, (ii) explicación, (iii) excusa, (iv) disculpa, (v) contraste, (vi) resultado, (vii) conclusión, (viii) evaluación, (ix) reformulación, (x) continuación, (xi) introducción de cambio de tema, (xii) vuelta a un tema anterior, (xiii) partícula de relleno, (xiv) introducción de discurso directo, (xv) intensificación, (xvi) atenuación orientada al hablante y (xvii) atenuación orientada al interlocutor. Así, en el ejemplo (4), el uso de *es que* permite a la hablante proteger su imagen, puesto que no puede aportar información a las dudas de las interlocutoras. Se analiza, por lo tanto, como un atenuador. El ejemplo (5) ejemplifica su uso como reformulador. La hablante RE2F1 afirma primero que no lo había pensado y luego reformula esta frase mediante *bueno es que*. En el ejemplo (6), la hablante usa *es que* como partícula de relleno mientras está buscando las palabras adecuadas.

(4) RE2F4: Ah y dijeron <Chicas y chicos solos> Y ahí (()) Mariam (()) Que no puede compartirla.
RE2F3: ¿Sabes?
RE2F2: Yo *es que* la verdad no lo sé.
RE2F1: Hostia, es verdah
RE2F3: Pero si por eso lo dije. (CORMA: RE_AM2_F_01)

(5) RE2F1: No lo había pensa'o bueno *es que* no me había enterado de la conversación. (CORMA: RE_AM2_F_01)

(6) AM3F4: Ya es que tú tampoco, entre que el trabajo y luego las tardes y-
AM3F3: *Es queee*, es es muy complica'o, osea ¿dónde conoces gente, en el instituto? No, ya nada. Claaro. (CORMA: AM.GEN3.F.02)

Aunque esta clasificación nos da una visión detallada del uso de *es que*, tiene algunas restricciones, relacionadas con la anotación de datos empíricos. En primer lugar, los límites entre las diferentes microfunciones resultan a veces ser mínimos. En (7), por ejemplo, se puede argumentar que *es que* aparece en el contexto de un contraste, reforzado por el conector contrastivo *pero*. Sin embargo, también se podría sostener que la hablante CBAR4F8 está describiendo la situación y explicando su problema. Por lo tanto, el análisis de *es que* como explicación también se puede argumentar.

(7) BAR3M1: Tenía a lo mejor cobraba el día quince del otro mes y ya me habían pasado todos los recibos todos con suuu
CBAR4F8: Claro yo que mi marido siempre cobraba el día diez y entonces claro cobraba el día diez y íbamos amoldando

> al día diez pagamos esto esto esto y ya está Pero *es que* ahora no es el día diez. Eh mejor el día veinte
> BAR3M1: El día veinte. Ya cuando le quiere pagar ya le debe dos meses. (CORMA: ATbar_01b-(2))

En segundo lugar, la lista de microfunciones depende sustancialmente del corpus usado. Así, Delahunty y Gatzkiewicz (2000) mencionan en su estudio sobre la construcción inferencial *ser que* la presencia de las microfunciones de consecuencia y efecto. Sin embargo, nuestra muestra no parece incluir ejemplos de estos valores. Fuentes Rodríguez (2015), por su parte, dedica un artículo entero a la función de intensificación de *es que*, mientras que las demás funciones mencionadas más arriba apenas se mencionan. Esto se debe al tipo de corpus utilizado; es decir, Fuentes Rodríguez (2015) analiza el discurso parlamentario, el cual exhibe características distintas a las del lenguaje utilizado en las conversaciones informales que se examinan aquí. La presencia o ausencia de una determinada microfunción y su frecuencia se vinculan, por lo tanto, al tipo de corpus analizado y al tamaño de la muestra. En tercer lugar, y en parte vinculado al segundo problema, cada investigador, con base en su corpus, llegará a su propia lista de microfunciones, lo cual no solo complicará la comparación entre marcadores similares entre distintas lenguas, por ejemplo, la comparación entre *es que* y su versión catalana *és que* (cf. Cuenca 2013), sino también las comparaciones entre distintos marcadores pragmáticos, tales como *es que* y *en plan*.

Por estas razones, conviene recurrir a la identificación de sus macrofunciones, siguiendo los modelos teóricos antedichos. Así, es posible reagrupar las microfunciones (i)-(xiv) bajo la macrofunción metadiscursiva y las microfunciones (xv)-(xvii) bajo la macrofunción modal. Se constata que *es que* puede asumir las mismas macrofunciones que han sido mencionadas anteriormente para otros marcadores como *nada*, *sabes* y *en plan*. Así, el ejemplo (4) se clasifica como uso de la macrofunción modal, mientras que los ejemplos (5) y (6) ilustran la macrofunción metadiscursiva. La clasificación en términos de macrofunciones permite más fácilmente establecer relaciones con otros marcadores pragmáticos como *en plan* que también se usa en el nivel metadiscursivo y modal (cf. *infra* § 3.2). Sin embargo, como ya sabemos, es verdad que mediante este segundo método se pierde bastante información específica de cada marcador.

Por lo tanto, resulta útil recurrir a una tercera opción metodológica, la de combinar ambas perspectivas: en un nivel más general se destacan las macrofunciones, aquí llamadas dimensiones[7] o macrocategorías (cf.

[7] En § 3.4 se explica en más detalle por qué se opta por el término de *dimensiones*. En pocas palabras, este término da cuenta de la polifuncionalidad de *es que* y el hecho de que puede intervenir

*infra* § 3.4), que incluyen explícitamente una lista de (micro)-funciones más delimitadas. La tabla 2 ilustra las dos dimensiones de *es que* con sus respectivas microfunciones y las frecuencias encontradas en nuestra muestra.

| **Dimensión metadiscursiva** | | |
|---|---|---|
| **Microfunción** | **N** | **%** |
| Razón | 77 | 38,5 |
| Contraste | 26 | 13,0 |
| Explicación | 60 | 30,0 |
| Evaluación | 10 | 5,0 |
| Cambio de tópico | 1 | 0,5 |
| Introducción de discurso directo | 6 | 3,0 |
| Partícula de relleno | 15 | 7,5 |
| Reformulación | 5 | 2,5 |
| **Total** | **200** | **100** |
| **Dimensión modal** | | |
| **Microfunción** | **N** | **%** |
| Intensificación | 58 | 55,24 |
| Atenuación orientada al locutor | 42 | 40,0 |
| Atenuación orientada al interlocutor | 5 | 4,76 |
| **Total** | **105** | **100** |

*Tabla 2. Dimensiones y funciones de* es que *y sus frecuencias.*

## *3.2 ¿Es preferible adoptar una perspectiva semasiológica u onomasiológica en el estudio de los marcadores pragmáticos?*

Como consecuencia de sus macrofunciones compartidas descritas en el apartado anterior, el estudio de los marcadores pragmáticos enfrenta un segundo dilema metodológico: la elección entre una perspectiva semasiológica y una onomasiológica, o la posible integración de ambas. La tendencia dominante en la literatura especializada ha sido adoptar un enfoque semasiológico, partiendo de los marcadores individuales como unidades de análisis. Esta orientación permite una descripción detallada de su comportamiento en distintos contextos comunicativos. No obstante, la perspectiva onomasiológica ofrece una alternativa complementaria que, en nuestra opinión, requiere más consideración. Este enfoque (que también se defiende en Briz Gómez y Albelda Marco 2013) prioriza la identificación de las funciones pragmáticas que los marcadores desempeñan en el discurso antes de

en diferentes dimensiones o niveles al mismo tiempo. Se observa que la dimensión metadiscursiva alcanza el 100 %, es decir, que está presente en cada uno de los casos observados, aunque con microfunciones diferentes, mientras que solo opera en la dimensión modal en el 52,5 % de los casos observados.

examinar las expresiones lingüísticas específicas utilizadas por los hablantes. Concretamente, en una fase inicial, sería posible identificar en un corpus delimitado las ocurrencias de las macrofunciones metadiscursivas, expresivas/modales e intersubjetivas. Posteriormente, se podría examinar la distribución de las unidades según las funciones. Sin embargo, para comprender por qué se selecciona un marcador específico en lugar de otro, el enfoque onomasiológico podría tomar en cuenta variables como el perfil sociolingüístico de los hablantes y ciertos elementos contextuales o discursivos. Así, el estudio onomasiológico permite una comprensión más profunda de cómo y por qué ciertas formas se utilizan para realizar determinadas funciones comunicativas, revelando la relación dinámica entre el uso del lenguaje y su contexto social y situacional. En esta línea de ideas Romero-Trillo (2006: 640) resalta que las funciones son más estables que las formas:

> [T]he phenomenon of discourse markers shows that spoken interaction needs to have a pragmatic skeleton, consisting of such discourse slots that hold the communicative force of the interaction together. The slots are filled by elements that may vary according to regional, idiolectal, or sociolinguistic features within one and the same language.

Para el análisis de *es que*, adoptamos un enfoque semasiológico, es decir, partimos de *es que* en contextos concretos para así identificar sus funciones (cf. *supra* § 3.1). Sin embargo, los resultados también han revelado que *es que* presenta similitudes funcionales con otros marcadores pragmáticos, por lo que la perspectiva onomasiológica también podría aplicarse. Así, en (8a), se percibe que la hablante IIC2F7 busca ganar tiempo antes de continuar su discurso, lo cual se puede marcar por el *slot* FUNCIÓN METADISCURSIVA. En el caso específico de (8b), la hablante utiliza el marcador *es que* precedido por *eh* como partícula de relleno. No obstante, este podría ser reemplazado por otros marcadores equivalentes como *bueno*, *sabes*, *mira*, *nada* o *en plan*, como se ilustra en (8c). En general, el comportamiento funcional de *es que* se parece mucho al perfil del marcador *en plan*. Ya sabemos que ambos se usan como partícula de relleno (8), pero también para introducir el discurso directo (9a-b) o con fines atenuantes (Camargo-Fernández y Grimalt Crespo 2022; De Smet y Enghels 2020; Méndez Orense 2016). Efectivamente, ambos marcadores comparten la función de atenuación con *nada* por lo cual un hablante puede recurrir a las tres opciones, como el ejemplo (10) muestra.

(8) a. IIC2F8: Bueno Irene, que seguro que te ha salido a ti bien, que yo ni siquiera averigüé que era un cataúd de estos de los cojones, qué frustración
IIC2F7: Ya- [FUNCIÓN METADISCURSIVA] pues la verdad que lo fui como deduciendo cuando estuvo dando la charla

b. Ya- [*eh- es que*-] pues la verdad que lo fui como deduciendo cuando estuvo dando la charla (CORMA: IIC_AM2_F_03)
c. Ya- eh- [*bueno* / *sabes* / *mira* / *nada* / *en plan*] pues la verdad que lo fui como deduciendo cuando estuvo dando la charla

(9) a. VV2F6: es que me llama cada notas. [Introducir discurso directo] <esto está tardando más de la cuenta, quiero un pedido gratis estoy muy enfadado>. Y yo <mira eres un jeta, madre mía madre mía>.
b. VV2F6: es que me llama cada notas. *Es que* <esto está tardando más de la cuenta, quiero un pedido gratis estoy muy enfadado>. Y yo <mira eres un jeta, madre mía madre mía>. (CORMA: VV_AM2_F_04)
c. VV2F6: es que me llama cada notas. *En plan* <esto está tardando más de la cuenta, quiero un pedido gratis estoy muy enfadado>. Y yo <mira eres un jeta, madre mía madre mía>.

(10) a. IIC2F3: Y yo Erika Y lo que te iba a decir Joder, [atenuación] somos alumnos. No hablamos de otra cosa que el instituto, en plan, hola ¿Qué s- ¿Qué se cree la gente, que tenemos vida personal?
b. IIC2F3: Y yo Erika Y lo que te iba a decir Joder, *es que* somos alumnos. No hablamos de otra cosa que el instituto, en plan, hola ¿Qué s- ¿Qué se cree la gente, que tenemos vida personal? (CORMA: IIC_AM2_F_01)
c. IIC2F3: Y yo Erika Y lo que te iba a decir Joder, *en plan/nada*, somos alumnos. No hablamos de otra cosa que el instituto, en plan, hola ¿Qué s- ¿Qué se cree la gente, que tenemos vida personal? (CORMA: IIC_AM2_F_01)

Si bien cada marcador seleccionado puede añadir un matiz diferente, resulta notable que no existe una relación directa y exclusiva entre las (macro)funciones discursivas y los marcadores que desempeñan estas funciones, tal y como afirma también López Serena (2011). Siguiendo el ejemplo de Fischer (2006) y Pons Bordería (2006), entre otros, la autora aboga explícitamente contra el mantenimiento acrítico del enfoque semasiológico y una orientación lexicocentrista en el estudio del lenguaje hablado en general, incluso de los marcadores pragmáticos.

Un elemento clave para entender por qué un hablante ha optado por una forma más que otra —como los originales con *es que* en (8b), (9b) y (10b)— podría situarse, por lo menos parcialmente, en su valor de 'indexicalidad social'. Siguiendo las teorías de Blommaert (2005), Coupland (2007) y Silverstein (2003, 2009), entre otros, es de suponer

que los marcadores pragmáticos no solo cumplen una función lingüística, sino que también pueden señalar o indexar la identidad social de los hablantes, sus relaciones interpersonales y sus actitudes o estados emocionales. Por ejemplo, el uso de ciertos marcadores puede variar significativamente según el grupo etario, el género, la clase social o la región geográfica de los usuarios, ofreciendo así pistas valiosas sobre sus identidades y sus contextos socioculturales. Beeching (2016), por ejemplo, subraya cómo determinados marcadores sirven para indexar tanto la informalidad como la pertenencia a determinados grupos sociales. En este sentido, Aijmer (2018) observa que las formas que llaman la atención, como *look*, *listen* o *excuse me* en inglés, son particularmente frecuentes y características del habla juvenil. Se encuentran resultados similares para el marcador *es que*, aunque menos pronunciados. El uso de *es que* se observa en todas las generaciones, pero llama la atención su alta productividad en el lenguaje juvenil. En concreto, tiene una frecuencia normalizada de 91,69 casos por cada 10 000 palabras en las conversaciones de los jóvenes y una frecuencia normalizada de 54,23 para los adultos. Sin embargo, *es que* aún no ha alcanzado el mismo estatus que el marcador *en plan*, que es casi exclusivo del lenguaje juvenil (Borreguero Zuloaga 2020; ver también Enghels en prensa para un análisis comparativo de *en plan* y *nada*).

Conviene destacar también que el perfil de *es que* no solo se asemeja al funcionamiento de otros marcadores pragmáticos, sino también a otros fenómenos lingüísticos, tales como la insubordinación con *que*. Efectivamente, estas construcciones también pueden intervenir en la dimensión metadiscursiva (con un valor explicativo o de introducción de discurso directo) (Gras 2011; Gras y Sansiñena 2017). En el ejemplo (11), el hablante IJ2F1 justifica y explica por qué le gusta más el Burger. En el caso concreto, la introducción de esta explicación se hace mediante *es que* (11b), pero también hubiera sido posible una insubordinada con *que* (11c). Así pues, la perspectiva onomasiológica revela también similitudes funcionales entre *es que* y otros fenómenos lingüísticos.

(11) a. IJ2F2: Pueh a mí me gusta m- ¿a ti qué te gusta más el Burger o el McDonald's?
IJ2F1: A mí el Burger yo creo.
IJ2F2: A mí igual te lo juro eh.
IJ2F1: [Razón] lah patatas están más buenas.
b. IJ2F1: *Eh que* lah patatas están más buenas. (CORMA: IJ_AM2_F_01)
c. IJ2F1: *que* lah patatas están más buenas.

En conclusión, según nuestro análisis, la elección de adoptar una perspectiva semasiológica, onomasiológica o una combinación de

ambas en el estudio de los marcadores pragmáticos debe guiarse por los objetivos específicos de la investigación. Mientras que la perspectiva semasiológica puede resultar más adecuada para investigaciones centradas en la forma y la variación lingüística, la perspectiva onomasiológica podría ser esencial para descubrir cómo diversas formas lingüísticas desempeñan funciones específicas dentro de prácticas comunicativas concretas. Optar por una combinación de ambas perspectivas podría proporcionar un enfoque más holístico y matizado, permitiendo una comprensión más amplia y detallada de la pragmática lingüística. Sin entrar en detalle, opinamos que el enfoque construccionista ofrece una valiosa herramienta para abordar tal perspectiva complementaria en el estudio de los marcadores pragmáticos, como se sugiere en los trabajos de Brinton (2008), Enghels (2018), Fried y Östman (2005) y Traugott (2018), entre otros. Este marco permite considerar los marcadores como una combinación de forma y función, y como «esquemas muy abiertos» (López Serena 2011). Además, permite dar cuenta tanto de la polifuncionalidad de los marcadores, que abarca diversos aspectos del significado (semánticos, pragmáticos, sociales y culturales) (cf. *infra* § 3.3), como de su comportamiento formal complejo (incluso su posición, colocaciones recurrentes y diferentes rasgos prosódicos) (cf. *infra* § 3.4).

### *3.3. Criterios formales para la identificación de funciones pragmáticas*

De lo que precede resulta claro que una de las áreas de la investigación lingüística más difíciles de estudiar es el significado y, en concreto, la función pragmática de una entidad lingüística. En efecto, preguntas del tipo ¿qué significa un determinado marcador?, ¿cuántas funciones tiene y en qué se diferencian exactamente? y ¿qué hacer con la ambigüedad entre funciones? han resultado durante mucho tiempo extremadamente difíciles de abordar. Esta dificultad incluso llega a reflejarse en análisis muy divergentes para un mismo marcador, donde las distinciones parecen más bien una cuestión de interpretación personal del lingüista sobre la que otros investigadores pueden disentir (cf. *supra* § 3.1).

En las últimas décadas, al igual que la lingüística en general, las áreas de la semántica y la pragmática también han experimentado un cambio significativo desde los enfoques basados en la intuición hacia el uso de corpus y métodos empíricos. Sin embargo, la aplicación de tales métodos empíricos y cuantitativos al estudio de las funciones pragmáticas dista de ser unívoca. En efecto, como ya sabemos, el significado de marcadores pragmáticos es intrínsecamente dependiente del contexto y de la interpretación que los interlocutores hacen en el

discurso concreto. Esto hace surgir la cuestión fundamental de cómo se puede estudiar el significado/la función —un fenómeno intrínsecamente subjetivo y no observable entre hablantes — mediante métodos cuantitativos. De ahí que un reto importante de la investigación semántico-pragmática de corpus consista en la operacionalización de este objeto de estudio altamente subjetivo y escurridizo que es el significado/la función.

A fin de encarar este reto, la lingüística de corpus se funda esencialmente en la hipótesis distribucional: la noción de que las diferencias de significado se reflejan en diferencias de distribución (véase Firth 1957; Harris 1954). Esta idea ha sido formulada de manera explícita en la afirmación bien conocida de Firth (1957:11): «you shall know a word by the company it keeps». Esta noción ha sido aplicada primero en el campo de la lexicografía (cf. por ejemplo Sinclair 1987) y la semántica cognitiva-funcional (cf. por ejemplo los análisis *Behavioral Profiles* de entre otros Divjak y Gries 2010 y Gries 2006) y se ha extendido más recientemente al análisis pragmático. En efecto, también en la pragmática, se ha recurrido al contexto para la identificación de funciones. Sin embargo, la gran diferencia entre los distintos tipos de análisis pragmáticos de índole empírica tiene que ver con la cantidad de información de coocurrencia utilizada y el tipo de elementos contextuales utilizados en el análisis. Así, por ejemplo, muchos estudios surgen de la pregunta de saber cómo las funciones pragmáticas correlacionan con determinados rasgos léxicos o morfosintácticos. Por eso, se fijan en la presencia de ciertos indicios léxicos y morfosintácticos en el contexto tales como la posición, las coocurrencias léxicas, el tipo de oración, el significado semántico original (cf. por ejemplo Azofra Sierra y Enghels 2017; Tanghe 2015) o incluso la configuración prosódica del marcador (Hidalgo Navarro 2010).

En la amplia bibliografía sobre este tema, mucha atención se ha dedicado a la relación entre una determinada función y la posición del marcador en el discurso. En el ámbito hispánico cabe mencionar el modelo del grupo de investigación Val.Es.Co (cf. por ejemplo Briz Gómez y Pons Bordería 2010). De esta manera, se ha podido comprobar que los marcadores en posición periférica, esto es, la periferia izquierda y derecha, favorecen determinados significados. Concretamente, se ha vinculado la periferia izquierda con funciones orientadas hacia el hablante (expresivas/subjetivas) y la periferia derecha con funciones orientadas más bien hacia el interlocutor (intersubjetivas/apelativas o fáticas) (cf. por ejemplo Degand 2014; Ghesquiere *et al*. 2012). Sin embargo, otros estudios han demostrado que la variación posicional de los marcadores no siempre es tan predecible y sistemática (cf. entre otros Traugott 2012 sobre los límites de la asociación entre el uso periférico y los significados subjetivos o intersubjetivos).

El análisis del marcador *es que* se enfrenta a los mismos problemas. En concreto, para operacionalizar las microfunciones de *es que* dentro de la dimensión metadiscursiva, se intentó considerar la posición del marcador tanto en el acto de habla como en la intervención como indicio de su función metadiscursiva. Sin embargo, como la tabla 3 muestra, el análisis revela que la posición y la función no se correlacionan, lo cual se podría explicar por la poca libertad posicional que *es que* experimenta en comparación con otros marcadores pragmáticos. Es decir, *es que* se posiciona mayoritariamente al inicio del acto de habla (12) y se encuentra apenas en posiciones medias o independientes del acto de habla o de la intervención.

(12) MS2M6: Porque no he- no hemos elegido subdelegado como tal.
MS2M5: Ej que nosotros tampoco tenemos subdelega'o.
(CORMA: MS_AM2_M_02)

| **Posición** | **Periferia izquierda** | | **Posición media** | | **Posición independiente** | | **Total** | |
|---|---|---|---|---|---|---|---|---|
| **Función** | **N** | **%** | **N** | **%** | **N** | **%** | **N** | **%** |
| Razón | 73 | 36,5 | 4 | 2,0 | 0 | 0 | 77 | 38,5 |
| Contraste | 26 | 13,0 | 0 | 0 | 0 | 0 | 26 | 13,0 |
| Explicación | 55 | 27,5 | 5 | 2,5 | 0 | 0 | 60 | 30,0 |
| Evaluación | 10 | 5,0 | 0 | 0 | 0 | 0 | 10 | 5,0 |
| Cambio de tópico | 1 | 0,5 | 0 | 0 | 0 | 0 | 1 | 0,5 |
| Introducción de discurso directo | 6 | 3,0 | 0 | 0 | 0 | 0 | 6 | 3,0 |
| Partícula de relleno | 4 | 2,0 | 0 | 0 | 11 | 5,5 | 15 | 7,5 |
| Reformulación | 5 | 2,5 | 0 | 0 | 0 | 0 | 5 | 2,5 |
| Total | 180 | 90,0 | 9 | 4,5 | 11 | 5,5 | 200 | 100 |

*Tabla 3. Correlación entre la posición de* es que *en el acto de habla y su función en la dimensión metadiscursiva*

Las limitaciones de usar exclusivamente la posición como indicador de función han impulsado a varios estudiosos a redefinir el concepto de *contexto* y a enriquecerlo, complementando la variable de posición con otros factores de diversa índole. Por ejemplo, recientemente, Ariel (2022) muestra cómo cada una de las funciones del marcador *harey* ('here is', 'hereby', 'after all') en hebreo está asociada a un perfil discursivo único (*discourse profiles*), es decir, un conjunto específico de características gramaticales y pragmáticas. Además de la posición, toma en cuenta factores tales como el estatuto de accesibilidad de la información que sigue (nueva vs. conocida), el constituyente modificado

(NP, oración…), la coocurrencia con otros elementos, posibles sinónimos, el género discursivo, la frecuencia del marcador, etc.

Avanzando en la misma dirección de los perfiles discursivos, el estudio de Van Olmen y Tantucci (2022) presenta un paso adicional. Estos autores se centran en el marcador *look* ('mira') en chino, neerlandés, inglés e italiano. Debido a este enfoque interlingüístico, uno de los principales objetivos consiste precisamente en desarrollar un marco analítico para la investigación pragmática de corpus que pueda aplicarse fácilmente a varias lenguas de forma coherente y que permita una cuantificación fiable. Más concretamente, el análisis se fundamenta en parámetros que deben ser fácilmente operacionalizables y aplicables de manera uniforme en la medida de lo posible (tales como el género discursivo, la posición en el turno, la posición en el enunciado, el significado original, la presencia de *look* en discurso reportado (*reported speech*), el acto de habla, la presencia de vocativos, etc.). Tal método les permite aplicar una serie de métodos estadísticos más avanzados (como *random forests* o *conditional inference trees*) para descubrir patrones de (di)similitud en los datos. Lo que llama la atención es que el estudio parte de dos premisas teóricas importantes: la primera es que básicamente se equiparan las distintas funciones del marcador con distintos contextos de uso (los *perfiles discursivos* únicos en términos de Ariel 2022). Segundo, en vez de centrarse en *colocaciones léxicas* a nivel textual, aspiran a identificar lo que llaman *concurrencias ilocucionales* (*illocutional concurrences*): se trata de intersecciones de la forma (p. ej., coocurrencias léxicas), la ilocución (p. ej., el acto de habla del enunciado marcado por el marcador) y la situación contextual (p. ej., el turno o la posición de la cláusula, así como el género en el que se realiza el acto lingüístico). En pocas palabras, mientras que las colocaciones pertenecen al texto como tal, las concurrencias encarnan el comportamiento interaccional más holístico y permiten así llegar a una especie de perfil de comportamiento interaccional del marcador.

El análisis de *es que* se inscribe en la misma línea de los estudios de Ariel (2022) y de Van Olmen (2022). En concreto, el perfil funcional de *es que* se fundamenta en diferentes parámetros. Al lado de la posición en el acto de habla y en la intervención, también se ha tomado en cuenta el tipo de intervención en que aparece *es que*. Siguiendo el estudio de la insubordinación de Gras y Sansiñena (2020), distinguimos entre una intervención iniciativa, reactiva o iniciativa-reactiva. Luego, las colocaciones léxicas también constituyen una herramienta útil en las anotaciones de los datos. Más en concreto, las palabras que preceden y que siguen a *es que* sirven como indicio para indicar ciertas microfunciones. Por ejemplo, en el caso de la dimensión metadiscursiva, la combinación *pero es que* suele indicar que seguirá una justificación de un contraste (13), la colocación *o sea es que* o *bueno es*

*que*, una reformulación, y *yo es que*, una explicación de una opinión. La combinación de *es que* seguida por elementos expresivos, como *joder* en (14), sugiere la microfunción de evaluación. Además, las palabras que rodean a *es que* desempeñan un papel importante para indicar una posible función de atenuación. En concreto, la presencia de otros recursos atenuantes como otros marcadores pragmáticos (como *nada* [15] o *en plan*), diminutivos, verbos y adverbios modales, etc. apoya la función de atenuación de *es que*.

(13) VV2M10: Hay hay gente con loj que se puede ir y se saben saben comportarse tío pero *eh que* hay otros que se ponen a hacer el retrasa'o. (CORMA: VV_AM2_M_02)

(14) (Están hablando de prácticas.)
IIC2F6: En el mío que yo sepa siguen sin hacerlo. Pero- sería muy triste que hubiera pasado lo mismo eh-
IIC2F7: Pues sí la verdad.
IIC2F6: Guay.
IIC2F8: Desde luego.
IIC2F9: *Eh que joder*, me dio una rabia. (CORMA: IIC_AM2_F_03)

(15) IR2F1: Los chicken fillets
IR2F2: Ugh o máh rico es la hamburguesa tía está puto buenísima
IR2F1: *Es que* no me apetece tía hamburguesa, *nada*. (CORMA: IR_AM2_F_02)

Al lado de estas características gramaticales y contextuales, otro factor que ha sido integrado para determinar su perfil discursivo es una posible reformulación o sustitución del enunciado con *es que*. Este método, también aplicado por Fuentes Rodríguez (1997, 2015), ha resultado ser muy fructífero para indicar las microfunciones metadiscursivas. En concreto, en los casos en que *es que* asume la microfunción de razón (16), se puede reformular el enunciado con *es que* por una de las siguientes expresiones: *digo esto porque…*, *pregunto o dudo porque…*, *también se debe a…*, *es porque…*, *otro argumento que lo justifica es…*, *la razón por la que digo esto es que…*, etc. Estas reformulaciones no son posibles con casos de *es que* con la microfunción de explicación que, al revés, admite reformulaciones como *en cuanto a este tópico*, *con respecto a este tema*, *por lo que se refiere a esta situación*, *te explico lo que pasó en esta situación* (17).

(16) a. ROPAj3F2: Pero este este. Este es que va a ser pequeño. Bueno a lo mejor sí le vale. *Es que* dan mucha talla estos. Estos dan muuucha talla. (CORMA: ATropa02_(1))
b. *La razón por la que digo que a lo mejor te vale es porque* dan mucha talla estos.

(17) a. MS2M7: Es como quien no ha visto Código Lyoko. Pues ¿quién no ha- quién no ha visto La Niña Repelente?
MS2F4: Que no
MS2F3: Te voy a poner un capítulo.
MS2M7: *Es que* duran Hay canciones que duran muchísimo más. (CORMA: MS_AM2_03)
b. *Por lo que refiere a estos capítulos*, duran […]

Todos los parámetros mencionados arriba nos permiten construir el contexto discursivo amplio en el que aparece *es que*. Está claro, pues, cómo la concepción de lo que es el contexto se ha ampliado con el paso del tiempo: desde los estudios centrados en los correlatos morfosintácticos y la posición, pasando por las colocaciones textuales para llegar finalmente a perfiles discursivos más amplios que son susceptibles de una cuantificación y operacionalización cada vez más sistematizada.

### *3.4. La polifuncionalidad y reconciliación de diferentes niveles de análisis*

La cuarta y última reflexión metodológica concierne la polifuncionalidad de los marcadores pragmáticos, como rasgo inherente de su definición (cf. *supra* § 1). Antes que nada, en la literatura especializada ni siquiera hay uniformidad en cuanto a la terminología utilizada para referirse a este fenómeno, puesto que aparece tanto el término *multifuncionalidad* como *polifuncionalidad* de manera indistinta. Así, por ejemplo, Degand (2019) habla de multifuncionalidad de la clase entera de los marcadores y polifuncionalidad dentro de los distintos tipos de marcadores discursivos, pero Crible (2017) utiliza *multifuncionalidad* como término paraguas y distingue tres situaciones a las que puede referir el término: (1) la categoría abarca elementos que desempeñan muchas funciones diferentes; (2) un único miembro puede desempeñar diferentes funciones en diferentes contextos; y (3) un solo miembro puede desempeñar diferentes funciones simultáneamente en el mismo contexto, dada su gran polisemia. Los fenómenos bajo (2) y (3) también han sido denominados casos de *polifuncionalidad paradigmática* y *polifuncionalidad sintagmática* respectivamente (Ghezzi y Molinelli 2014; López Serena y Borreguero Zuloaga 2010).

Aunque hay, pues, ciertas divergencias terminológicas, los autores suelen coincidir en reconocer la polifuncionalidad intrínseca de la categoría. Más recientemente, han surgido nuevas taxonomías para abordar la polifuncionalidad de los marcadores de una manera más sistematizada. Concretamente, Crible y Degand (2019) abogan a favor de un acercamiento bidimensional a la polifuncionalidad de los marcadores: su principal característica innovadora consiste en distinguir entre dos capas independientes de información semántico-pragmática, llamadas *dominios*, por un lado, y *funciones*, por otro. Más concretamente, distinguen entre cuatro dominios (en concreto, secuencial, retórico, ideacional e interpersonal) y quince funciones (por ejemplo temporal, causal, desacuerdo, concesión o hedging) que son independientes. Esto significa, pues, que cualquier función puede combinarse con cualquier dominio, llegando de esta forma a un espacio bidimensional con 60 (4 x 15) combinaciones posibles de dominio-función. En el ámbito hispánico encontramos acercamientos muy similares. En este sentido, cabe destacar el estudio de Azofra Sierra y Enghels (2022) donde se propone una representación radial de los distintos valores del marcador *nada* a partir de dos macrofunciones (discursiva e [inter]subjetiva) para conceptualizar su polifuncionalidad sintagmática.

Una segunda dificultad que surge de la polifuncionalidad se relaciona con el alcance de inclusión de los niveles de análisis. En efecto, la clasificación funcional puede situarse en diferentes niveles de análisis y lo que se entiende como *función* puede ser de índole muy diversa. De ahí que surja la pregunta de saber cuál es el grado de inclusión más apropiado y cómo se pueden reconciliar los distintos niveles de análisis (cf. Azofra Sierra y Enghels 2017, 2022; Martín Zorraquino y Portolés 1999; Shiffrin 1987 entre muchos otros). Aunque la mayoría de los estudios toman como punto de partida las consabidas (macro) funciones semántico-pragmáticas, también se nota una ampliación cada vez mayor del concepto de la *función*.

En esta línea de ideas, varios estudios sobre marcadores toman en cuenta también la imagen social de los participantes en el discurso, vinculada a la macrofunción modal, que se puede subdividir en la función subjetiva (en el caso de la imagen del hablante) o intersubjetiva (en el caso de la imagen del oyente). Este concepto de la imagen ayuda, por ejemplo, a discernir entre distintos tipos de atenuación y refinar así el análisis funcional de los marcadores. Como afirman Albelda Marco y Cestero Mancera (2011: 15), la atenuación «[p]uede afectar a diversos elementos del proceso comunicativo: al mensaje, al hablante, al oyente o a la relación entre ambos». Esta necesidad de protección de la imagen (propia o ajena) se nota, por ejemplo, en el marcador *nada* que manifiesta tanto lo que se denomina *atenuación pragmática orientada al hablante* (inherentemente subjetiva, donde se utiliza para minimizar

el grado de responsabilidad del propio hablante con el contenido de su intervención o para proteger su propia imagen) como *atenuación pragmática orientada al oyente* (donde el hablante descarga de responsabilidad o preocupación al oyente, constituyendo así una forma de cortesía) (cf. Azofra Sierra y Enghels 2022).

La polifuncionalidad también atraviesa el análisis de *es que*. La mayoría de los pocos estudios sobre *es que* se centran en las funciones pragmático-discursivas de *es que* (cf. estudios de Fuentes Rodríguez 1997, 2015 y Remberger 2020), tales como los valores explicativo, justificativo, etc. (cf. los valores [i]-[xiv] anteriormente mencionados, cf. *supra* § 3.1). En una primera fase del análisis, se limitó la anotación de los casos a una sola función, su función que consideramos como dominante. En este sentido, en el ejemplo (18), *es que* se usa predominantemente para introducir una disculpa, indicada por el uso de *perdona*. En el ejemplo (19) (citado *supra* como ejemplo [15]), por su parte, *es que* aparece en un contexto de contraste donde la hablante explica por qué no está de acuerdo con la opinión de su interlocutora.

(18) IR2F13: ¿Estás imitando mi acento gallego?
IR2F15: Perdona, *es que* te ha salido y ha sido como-.
(CORMA: IR_AM2_F_07)

(19) IR2F1: Los chicken fillets
IR2F2: Ugh o máh rico es la hamburguesa tía está puto buenísima
IR2F1: *Es que* no me apetece tía hamburguesa, *nada*. (CORMA: IR_AM2_F_02)

Sin embargo, no se puede negar que *es que* se usa en ambos ejemplos con fines atenuantes. En (18), al utilizar este marcador, la hablante IR2F15 afirma que se trata de su opinión e intenta así proteger su propia imagen, así como la imagen del interlocutor al que puede haber ofendido. En (19), al oponerse a la interlocutora, existe el riesgo de ofender la imagen de ella. El uso de *es que*, por tanto, sirve para evitar un posible conflicto. Por lo tanto, limitar el análisis de *es que* a una función dominante no nos daría la imagen completa de su funcionamiento. Así, dando cuenta de la polifuncionalidad de los marcadores pragmáticos, se ha llegado a un acercamiento dimensional de las funciones de *es que*, considerando que *es que* puede intervenir en dos niveles o dimensiones al mismo tiempo: la dimensión (i) metadiscursiva y (ii) modal. Dentro de cada dimensión, *es que* puede asumir diferentes funciones, como hemos explicado en § 3.1.

Está claro, pues, que paralelamente a la ampliación del concepto de lo que es el contexto, discutido en el apartado anterior (cf. *supra* § 3.3),

se nota también una interpretación cada vez más amplia de lo que se entiende por función y cómo se aborda la polifuncionalidad intrínseca de los marcadores: la función va más allá del nivel semántico-pragmático para incorporar también fenómenos relacionados con la imagen (cortesía verbal) e incluso la indexicalidad social y de identidad (cf. *supra* § 3.2).

## 4. Conclusión

El estudio de los marcadores pragmáticos, como se ha evidenciado en la literatura lingüística reciente, continúa siendo un campo fértil para la investigación, a pesar de las complejidades inherentes a su análisis. Estos elementos del lenguaje, que juegan un papel crucial en la organización del discurso y en la gestión de la interacción entre los interlocutores, presentan una polifuncionalidad y una naturaleza híbrida que han generado un debate continuo sobre la mejor manera de clasificarlos y analizarlos.

Este trabajo se ha enfocado en discutir la viabilidad de continuar clasificando funcionalmente los marcadores pragmáticos, enfrentando numerosas dificultades teóricas y metodológicas. El punto de partida lo constituye la idea de que la semántica y la pragmática, disciplinas que abordan el significado, presentan desafíos particulares para la operacionalización de conceptos abstractos, especialmente cuando se trata de elementos lingüísticos cuyo significado es procedimental y no referencial. Esto plantea la cuestión de cómo metodologías como el análisis de corpus pueden capturar adecuadamente la función de estos elementos, que están profundamente arraigados en la dinámica del discurso coloquial.

A través del análisis del marcador pragmático *es que* en el Corpus Oral de Madrid (CORMA), hemos abordado cuatro cuestiones centrales: el nivel de detalle (macro o micro) en la definición y clasificación de funciones, la elección entre una perspectiva semasiológica u onomasiológica, la posibilidad de establecer criterios objetivos para identificar funciones pragmáticas y los criterios adecuados para dar cuenta de su polifuncionalidad. Estos dilemas han sido examinados mediante un análisis detallado que incluye una amplia gama de variables sociolingüísticas, formales y funcionales.

En primer lugar, uno de los principales desafíos que se ha destacado es el equilibrio entre la especificidad y la generalización en la clasificación de los marcadores pragmáticos. Mientras que una clasificación basada en microfunciones permite un análisis detallado, no siempre resulta fácil observar las diferencias concretas. Por el otro lado, si se generaliza en macrofunciones más amplias, también puede llevar a una

pérdida de comprensión sobre el funcionamiento específico de cada marcador. Este trabajo ilustra cómo *es que*, por un lado, puede desempeñar múltiples funciones, desde la justificación y la explicación hasta la atenuación y la intensificación, dependiendo del contexto discursivo, y, por otro lado, opera en los dos niveles de las macrofunciones metadiscursiva y modal, rasgo que comparte con otros marcadores como *nada* y *en plan*.

En segundo lugar, la elección entre una perspectiva semasiológica, que parte de las formas lingüísticas para analizar su uso, y una perspectiva onomasiológica, que se enfoca primero en las funciones en el discurso, es otro tema central en este estudio. Mientras que la tendencia dominante ha sido adoptar un enfoque semasiológico, este trabajo argumenta que una perspectiva onomasiológica podría ofrecer una comprensión más profunda de cómo y por qué ciertas formas se utilizan para realizar funciones específicas, revelando la relación dinámica entre el uso del lenguaje y su contexto social y situacional.

En tercer lugar, el estudio reconoce la dificultad de utilizar métodos cuantitativos para analizar funciones pragmáticas, dado que el significado de los marcadores pragmáticos es intrínsecamente dependiente del contexto. La investigación sobre *es que* muestra que, si bien la posición del marcador en el discurso puede proporcionar pistas sobre su función, esta variable por sí sola no es suficiente para una clasificación completa, lo que subraya la necesidad de considerar un conjunto más amplio de factores contextuales y discursivos, como sus colocaciones o el tipo de intervención en que aparece.

Finalmente, es generalmente sabido que la polifuncionalidad de los marcadores pragmáticos, un rasgo inherente a su definición, presenta un desafío significativo para su análisis. Este estudio ha abogado por un enfoque que combine tanto la clasificación en dos dimensiones como el análisis de microfunciones más específicas, permitiendo una comprensión más holística y matizada de estos elementos lingüísticos. Además, se ha sugerido que la función de los marcadores va más allá del nivel semántico-pragmático, incorporando también aspectos relacionados con la imagen social y la identidad de los hablantes, y ampliando el concepto de *significado*.

En conclusión, nuestra investigación sobre marcadores pragmáticos como *es que* ha revelado la necesidad de continuar explorando metodologías que puedan capturar la complejidad de estos elementos lingüísticos. Un enfoque que combine diferentes perspectivas y niveles de análisis no solo enriquecerá nuestra comprensión teórica, sino que también permitirá una aplicación más efectiva de métodos verificables y objetivos. Este estudio, por lo tanto, aboga a favor de la consolidación

de un marco teórico y metodológico que pueda abordar la polifacética naturaleza de los marcadores pragmáticos en el discurso.

## Bibliografía

Aijmer, Karin (2018), «Positioning of self in interaction adolescents' use of attention-getters», en Kate Beeching, Chiara Ghezzi y Piera Molinelli (eds.), *Positioning the self and others: linguistic perspectives*, Amsterdam (Phil.), John Benjamins Publishing Company: 177-195. DOI: 10.175/pbns.292.08aij.

Albelda Marco, Marta, y Ana María Cestero Mancera (2011), «De nuevo, sobre los procedimientos de atenuación lingüística», *Español Actual*, 96: 9-40.

Ariel, Mira (2022), «Processing polyfunctional discourse markers: Making sense of Hebrew *harey*», en Maria-Josep Cuenca y Liesbeth Degand (eds.), *Discourse markers in interaction*, Berlín/Nueva York, De Gruyter Mouton: 247-276. DOI: 10.1515/9783110790351-010.

Azofra Sierra, María Elena, y Renata Enghels (2017), «El proceso de gramaticalización del marcador epistémico deverbal *sabes*», *Iberoromania*, 85: 1-25. DOI: 10.1515/iber-2017-0008.

Azofra Sierra, María Elena, y Renata Enghels (2022), «La polifuncionalidad del marcador conversacional *nada*: metadiscurso e intersubjetividad», en Javier Herrero Ruiz de Loizaga, María Elena Azofra Sierra y Rosario González Pérez (eds.), *La configuración histórica del discurso: nuevas perspectivas en los procesos de gramaticalización, lexicalización y pragmaticalización*, Madrid, Vervuert: 13-46. DOI: 10.31819/9783968692944-002.

Beeching, Kate (2016), *Pragmatic markers in British English: meaning in social interaction*, Cambridge, Cambridge University Press.

Blommaert, Jan (2005), *Discourse: a critical introduction*, Cambridge, Cambridge University Press.

Brinton, Laurel J. (1996), *Pragmatic markers in English: grammaticalization and discourse functions*, Berlín/Nueva York, Mouton de Gruyter. DOI: 10.1515/9783110907582.

Brinton, Laurel J. (2008), *The comment clause in English: syntactic origins and pragmatic development*, Cambridge, Cambridge University Press. DOI: 10.1017/CBO9780511551789.

Briz Gómez, Antonio, y Salvador Pons Bordería (2010), «Unidades, marcadores discursivos y posición», en Óscar Loureda Lamas y Esperanza Acín Villa (eds.), *Los estudios sobre marcadores del discurso en español, hoy*, Madrid, Arco Libros: 327-358.

Briz Gómez, Antonio, y Marta Albelda Marco (2013), «Una propuesta teórica y metodológica para el análisis de la atenuación lingüística en español y portugués: la base de un proyecto en común (ES. POR.ATENUACIÓN)», *Onomazein*, 28: 288-319. DOI: 10.7764/onomazein.28.16.

Borreguero Zuloaga, Margarita (2020), «Los marcadores de aproximación (en el lenguaje juvenil): esp. *en plan* vs. it. *tipo*», en Miguel Ángel Cuevas Gómez, Fernando Molina Castillo y Paolo Silvestri (coords.), *España e Italia. Un viaje de ida y vuelta: studia in honorem Manuel Carrera Díaz*, Sevilla, Editorial Universidad de Sevilla: 53-78.

Camargo Fernández, Laura, y Ana María Grimalt Crespo (2022), «Nuevas y viejas funciones de *en plan*: estudio microdiacrónico en corpus orales y digitales del castellano de Mallorca en el siglo xxi», *Revista de Investigación Lingüística*, 25: 15-42. DOI: 10.6018/ril.537931.

Coupland, Nikolas (2007), *Style: language variation and identity*, Cambridge, Cambridge University Press. DOI: 10.1017/CBO9780511755064.

Crible, Ludivine (2017), «Towards an operational category of discourse markers: a definition and its model», en Chiara Fedriani y Andrea Sanso (eds.), *Discourse markers, pragmatics markers and modal particles: new perspectives*, Amsterdam (Phil.), John Benjamins: 101-126. DOI: 10.1075/slcs.186.04cri.

Crible, Ludivine, y Liesbeth Degand (2019), «Domains and functions: a two-dimensional account of discourse markers», *Discours-revue de Linguistique Psycholinguistique et Informatique*, 24. DOI:10.4000/discours.9997.

Crible, Ludivine, y Elena Pascual (2020), «Combinations of discourse markers with repairs and repetitions in English, French and Spanish», *Journal of Pragmatics*, 156: 54-67. DOI: 10.1016/j.pragma.2019.05.002.

Cuenca, Maria Josep (2013), «The fuzzy boundaries between discourse marking and modal marking», en Liesbeth Degand, Bert Cornillie y Paola Pietrandrea (eds.), *Discourse markers and modal particles:*

*categorization and description*, Amsterdam (Phil.), John Benjamins: 191-216. DOI: 10.1075/pbns.234.08cue.

Degand, Liesbeth (2014), «*So Very Fast Then*: discourse markers at left and right periphery in spoken French», en Kate Beeching y Ulrich Detges (eds.), *Discourse functions at the left and right periphery: crosslinguistic investigations of language use and language change*, Leiden/Boston, Brill: 151-178. DOI: 10.1163/9789004274822_008.

Degand, Liesbeth (2016), *Spoken discourse segmentation and the paradox of discourse markers*, L&C talk, Nijmegen.

Degand, Liesbeth (2019), «The paradox of discourse markers: evidence from production under cognitive load», *6th International Conference on Discourse Markers in Romance Languages*, Bergamo.

Delahunty, Gerald P., y Laura Gatzkiewicz (2000), «On the Spanish inferential construction *ser que*», *Pragmatics*, 10: 301-322. DOI: 10.1075/prag.10.3.01del.

De Smet, Emma, y Renata Enghels (2020), «Los datos en Twitter como fuente del discurso oral coloquial: estudio de caso del marcador discursivo *en plan*», *Oralia*, 23 (2): 199-218. DOI: 10.25115/oralia.v23i2.6379.

Divjak, Dagmar, y Stefan Th. Gries (2006), «Ways of trying in Russian: clustering behavioral profiles», *Corpus Linguistics and Linguistic Theory*, 2 (1): 23-60. DOI: 10.1515/CLLT.2006.002.

Enghels, Renata (2018), «Towards a constructional approach to discourse-level phenomena: the case of the Spanish interpersonal epistemic stance construction», *Folia Linguistica*, 52 (1): 107-38. DOI: 10.1515/flin-2018-0002.

Enghels, Renata (en prensa), «Pragmatic markers as social identity signals in contemporary colloquial Spanish», en Marcia Machado Vieira y Vanessa Meireles (eds.), *Variação nas linguas românicas: rumo a uma descrição gramatical pluricêntrica na era das Humanidades Digitais*, Montpellier, Université Paul Valéry,

Enghels, Renata, y María Elena Azofra Sierra (2018), «On the nature of the corpus and the comparability of results in historical linguistics: case study of the pragmatic marker *you know*», *Spanish in Context*, 15 (3): 465-489. DOI: 10.1075/sic.00023.eng.

Enghels, Renata, Fien De Latte y Linde Roels (2020), «El Corpus Oral de Madrid (CORMA): materiales para el estudio (socio)lingüístico del español coloquial actual», *Zeitschrift für Katalanistik*, 33: 45-76. DOI: 10.46586/ZFK.2020.45-76.

Enghels, Renata, y María Elena Azofra Sierra (2024), «El marcador *nada* en el corpus CORMA: un enfoque integrador», en Francisco Javier Herrero Ruiz de Loizaga, Renata Enghels y Rosario González Pérez (eds.), *Cambio y variación en el discurso en español: estudios sobre gramaticalización y lexicalización*, Madrid, Iberoamericana Vervuert: 129-168.

Firth, John Rupert (1957), *Papers in Linguistics, 1934-1951*, Londres/Nueva York: Oxford University Press.

Fischer, Kerstin (2006), «Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume», en Kirsten Fischer (ed.), *Approaches to discourse particles*, Amsterdam, Elsevier: 1-20. DOI: 10.1163/9780080461588_002.

Foolen, Ad (2011), «Pragmatic markers in a sociopragmatic perspective», en Gisle Andersen y Karin Aijmer (eds.), *Pragmatics of society*, Berlín/Nueva York, De Gruyter Mouton: 217-282. DOI: 10.1515/9783110214420.217.

Fraser, Bruce (1999), «What are discourse markers?», *Journal of Pragmatics*, 31: 931-952. DOI: 10.1016/S0378-2166(98)00101-5.

Fried, Mirjam, y Jan-Ola Östman (2005), «Construction grammar and spoken language: the case of pragmatic particles», *Journal of Pragmatics*, 37 (11): 1752-1778. DOI: 10.1016/j.pragma.2005.03.013.

Fuentes Rodríguez, Catalina (1997), «Los conectores en la lengua oral: *es que* como introductor de enunciado», *Verba*, 24: 237-263.

Fuentes Rodríguez, Catalina (2015), «Pragmagramática de *es que*: el operador de intensificación», *Estudios filológicos*, 55: 53-76. DOI: 10.4067/S0071-17132015000100004.

Geeraerts, Dirk (2010), «The doctor and the semantician», en Dylan Glynn y Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics: corpus-driven approaches*, Berlín/Nueva York, De Gruyter Mouton: 63-78. DOI: 10.1515/9783110226423.61.

Ghesquière, Lobke, Lieselotte Brems, y Freek Van de Velde (2012), «Intersubjectivity and intersubjectification: typology and operationalization», *English Text Construction*, 55 (1): 128-152.

Ghezzi, Chiara (2014), «The development of discourse and pragmatic markers», en Chiara Ghezzi y Piera Molinelli (eds.), *Discourse and pragmatic markers from Latin to the romance languages*, Oxford, Oxford University Press: 10-26. DOI: 10.1093/acprof:oso/9780199681600.003.0002.

Ghezzi, Chiara, y Piera Molinelli (eds.) (2014), «Discourse and pragmatic markers from Latin to the romance languages: new insights», en Chiara Ghezzi y Piera Molinelli (eds.), *Discourse and pragmatic: markers from Latin to the Romance languages*, Oxford, Oxford University Press: 1-9. DOI: 10.1093/acprof:oso/9780199681600.001.0001.

Gras, Pedro (2011), *Gramática de construcciones en interacción: propuesta de un modelo y aplicación al análisis de estructuras independientes con marcas de subordinación en español*, tesis doctoral, Universitat de Barcelona.

Gras, Pedro, y María Sol Sansiñena (2017), «Exclamatives in the functional typology of insubordination: evidence from complement insubordinate constructions in Spanish», *Journal of Pragmatics*, 115: 21-36. DOI: 10.1016/j.pragma.2017.04.005.

Gras, Pedro, y María Sol Sansiñena (2020), «Un caso de variación pragmático-discursiva: *que* inicial en tres variedades dialectales del español», *Romanistisches Jahrbuch*, 71 (1): 271-304. DOI: 10.1515/roja-2020-0012.

Gries, Stefan Th. (2006), «Corpus-based methods and cognitive semantics: the many senses of to run», en Stefan Th. Gries y Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, Berlín/Nueva York, Mouton de Gruyter: 57-99. DOI: doi.org/10.1515/9783110197709.57.

Harris, Zellig S. (1954), «Distributional structure», *Word*, 10: 146-162.

Hidalgo Navarro, Antonio (2010), «Los marcadores del discurso y su significante: en torno a la interfaz marcadores-prosodia en español», en Óscar Loureda Lamas y Esperanza Acín Villa (eds.), *Los estudios de los marcadores del discurso en español, hoy*, Madrid, Arco/libros: 61-92.

López Serena, Araceli (2011), «Más allá de los marcadores del discurso», *Sintaxis y análisis del discurso hablado en español: homenaje a Antonio Narbona*, 1: 275-294.

López Serena, Araceli, y Margarita Borreguero Zuloaga (2010), «Los marcadores del discurso y la variación lengua hablada vs. lengua escrita», en Óscar Loureda Lamas y Esperanza Acín Villa (eds.), *Los estudios sobre los marcadores del discurso en español, hoy*, Madrid, Arco/Libros: 415-493.

Martín Zorraquino, María Antonia, y José Portolés Lázaro (1999), «Los marcadores del discurso», en Ignacio Bosque y Violeta Demonte

(eds.), *Gramática descriptiva de la lengua española*, Madrid, Espasa Calpe: 4051-4214.

Méndez Orense, María (2016), «Valores pragmático-discursivos de la construcción lingüística *en plan*: ¿formación de un nuevo marcador?», *Philología Hispalensis*, 30 (1/2): 123-144. DOI: 10.12795/PH.2016.i30.07.

Östman, Jan-Ola (1995), «Pragmatic particles twenty years after», *Organization in discourse*, 14: 95-108.

Pons Bordería, Salvador (2006), «A functional approach to discourse markers», en Kirsten Fischer (ed.), *Approaches to discourse particles*, Amsterdam, Elsevier: 77-99.

Remberger, Eva-María (2020), «Information-structural properties of IS THAT-clauses», en Pierre-Yves Modicom y Olivier Dupâtre (eds.), *Information-structural perspectives on discourse particles*, Amsterdam (Phil.), John Benjamins: 47-70. DOI: 10.1075/slcs.213.02rem.

Romero-Trillo, Jesus (2006), «Discourse markers», en Keith Brown (ed.), *Encyclopedia of language and linguistics*, Oxford, Elsevier: 639-641.

Schiffrin, Deborah (1987), *Discourse markers*, Cambridge, Cambridge University Press. DOI: 10.1017/CBO9780511611841.

Silverstein, Michael (2003), «Indexical order and the dialectics of sociolinguistic life», *Language & Communication*, 23 (3-4): 193-229. DOI: 10.1016/S0271-5309(03)00013-2.

Silverstein, Michael (2009), «Pragmatic indexing», en Jacob L. May (ed.), *Concise encyclopedia of pragmatics*, Amsterdam, Elsevier: 756-759. DOI: 10.1016/B0-08-044854-2/00381-3.

Sinclair, John (ed.) (1987), *Collins cobuild English language dictionary*, London, HarperCollins.

Tanghe, Sanne (2015), *Marcadores derivados de verbos de movimiento: una aproximación cognitiva a su polifuncionalidad*, tesis doctoral, Universidad de Gante.

Traugott, Elizabeth Closs (2012), «Pragmatics and language change», en Keith Allan y Kasia M. Jaszczolt (eds.), *The Cambridge handbook of pragmatics*, Cambridge, Cambridge University Press: 549-566. DOI: 10.1017/CBO9781139022453.030.

Traugott, Elizabeth Closs (2018), «Modeling language change with constructional networks», en Salvador Pons Bordería y Óscar

Loureda Lamas (eds.), *Beyond grammaticalization and discourse markers*, Leiden, Brill: 17-50.

Van Den Driessche, Nele, y Renata Enghels (en prensa), «El marcador pragmático *es que* en el lenguaje juvenil madrileño: productividad lingüística y descripción formal-funcional», *Revue Romane*.

Van Olmen, Daniël, y Vittorio Tantucci (2022), «Getting attention in different languages: a usage-based approach to parenthetical look in Chinese, Dutch, English, and Italian», *Intercultural Pragmatics*, 19 (2): 141-181. DOI: 10.1515/ip-2022-2001.

Antonio Hidalgo Navarro (ORCID: 0000-0002-6534-4168) y Carlos Castelló Vercher (ORCID: 0000-0002-5459-7178),
«La transcripción de los elementos prosódicos en corpus de habla coloquial espontánea».
*Biblioteca de Babel: Revista de Filología Hispánica*, vol. extra. 2 (2024), pp. 51-78.
**DOI**: 10.15366/bibliotecababel2024.extra2.002 · **ISSNe**: 2695-6349 · Recibido: 02/09/2024 · Aceptado: 07/11/2024

# La transcripción de los elementos prosódicos en corpus de habla coloquial espontánea[1]

Antonio Hidalgo Navarro
*Universitat de València*
antonio.hidalgo@uv.es

Carlos Castelló Vercher
*Universitat de València*
carlos.castello@uv.es

**Resumen**: La lingüística de corpus tiene como objetivo la generación de bases de datos procedentes de la lengua real dispuestos para su análisis. En este artículo sintetizamos los conceptos básicos sobre transcripción y elaboración de corpus textuales y orales, centrándonos en estos últimos, así como en los sistemas de transcripción de los elementos prosódicos en la conversación coloquial. Tras esta revisión, proponemos un sistema de transcripción prosódica a partir del modelo del Grupo Val.Es.Co. con una serie de criterios e instrucciones objetivas, posibles gracias al empleo de los programas informáticos ELAN y Praat. El sistema propuesto se centra en la identificación y descripción de límites y contornos prosódicos tras el análisis de los fenómenos *pausa* o *alargamiento* y *reajuste tonal*, *inflexión final*, *tonema* o *F0 inicial* y *final*, enmarcados en las cualidades del sonido *duración* y *F0*, respectivamente.

**Palabras clave**: prosodia, corpus orales, transcripción, conversación coloquial.

## Transcription of prosodic elements in spontaneous colloquial speech corpora

**Abstract**: The aim of corpus linguistics is the elaboration and analysis of real language texts. In this article we synthesise the basic concepts

[1] Esta investigación ha sido posible gracias a la ayuda recibida por el Ministerio de Ciencia, Innovación y Universidades para el proyecto ECOS-C/N, Estudio de los condicionantes sociales del español actual en el centro y norte de España: nuevas identidades, nuevos retos, nuevas soluciones, (ref. PID2023-148371NB-C42).

of transcription and elaboration of textual and oral corpora, focusing on the latter, as well as on the systems of transcription of prosodic elements in colloquial conversation. Following this review, we propose a prosodic transcription system based on the Val.Es.Co. system with a series of objective criteria and instructions, made possible using the ELAN and Praat software. This system focuses on the identification and description of prosodic boundaries and contours after the analysis of the phenomena *pause* or *lengthening* and *tonal readjustment*, *final inflection*, *toneme* or *initial* and *final F0*, framed in the sound qualities *duration* and *F0*, respectively.

**Keywords**: prosody, oral corpus, transcription, colloquial conversation.

## 1. Lingüística de corpus y transcripción de la lengua oral

La moderna lingüística de corpus se ha interesado entre otras cuestiones por la elaboración de métodos para la recopilación y organización de muestras de habla (orales o escritas) en soporte digital. Su propósito es obtener datos objetivos, empíricos, de modo que cada corpus lingüístico pueda llegar a representar a escala el funcionamiento real del lenguaje natural.

Tales corpus han permitido disponer de bases de datos amplias para proceder al estudio de las lenguas convencionales y de sus características intrínsecas (prosodia, léxico, morfología, sintaxis, aspectos históricos, etc.) y valorar otros factores como, por ejemplo, la variación lingüística; por lo demás, la disposición de corpus digitalizados permite someter los datos a diversas formas de procesamiento cualitativo y/o cuantitativo.

### *1.1. Corpus orales y corpus textuales*

Existen dos tipos de corpus lingüísticos: los *textuales* y los *orales*. Los primeros contienen muestras de la lengua escrita, los segundos, muestras de la lengua oral (transcripción ortográfica o grabaciones acompañadas de transcripción). Por lo demás, en líneas generales, la necesidad de obtener modelos estadísticos de la lengua para su aplicación en el desarrollo de sistemas de reconocimiento (aplicables a trabajos como el dictado automático) ha llevado a un uso cada vez más frecuente de los corpus textuales y de las transcripciones del registro oral espontáneo. Además, el creciente interés por los aspectos prosódicos y suprasegmentales de la conversación y del discurso oral ha obligado a disponer del texto transcrito y de la grabación.

Por razones de espacio no tratamos aquí sobre los problemas generales relativos al diseño, confección o aprovechamiento de un corpus, ni tampoco describimos sus tipos o sus criterios de construcción (véanse al respecto Llisterri 1997; Torruella y Llisterri 1999; Alvar Ezquerra y Villena Ponsoda 1994; Ávila Muñoz 1996; Pino y Sánchez 1999). En lo que sigue nos ceñimos más precisamente a tratar sobre los corpus orales (transcripciones ortográficas de lengua hablada[2]) y, dentro de estos, un subgrupo más específico: el de corpus orales de conversación espontánea[3] (Briz y Grupo Val.Es. Co. 1995 y 2002). En este marco conversacional trataremos de manera específica sobre algunos problemas habituales que afectan a la interpretación de los datos transcritos, particularmente en lo que respecta a los fenómenos suprasegmentales.

## *1.2. Sistemas de transcripción de corpus orales*

La transcripción de conversaciones se ha convertido en una práctica común en diferentes disciplinas: antropología, lingüística, psicología, derecho... No extraña, pues, que a objetos de estudio distintos correspondan sistemas de transcripción también diferentes.

Al respecto, Payrató (1995) hacía referencia a distintos sistemas de transcripción y codificación: los procedentes de la etnometodología, interesados en reflejar la interacción verbal (Atkinson y Heritage, Eds. 1984; Button y Lee, Eds. 1987), los desarrollados por la etnografía de la comunicación o la sociolingüística interaccional (Ochs 1979; Tannen 1987; DuBois 1991; DuBois et al. 1993; Gumperz y Berenz 1993) o la propuesta de CHILDES (MacWhynney 1991) utilizada para el estudio

[2] De acuerdo con Torruella y Llisterri (1999:53-59) podemos catalogar un corpus de muy diversas formas, tanto mediante criterios generales:
- según el porcentaje y la distribución de los diferentes tipos de textos,
- según la especificidad de los textos,
- según la cantidad de texto recogida en cada documento,
- según la codificación y la anotación,
- según la documentación que acompaña a los textos,
como mediante criterios específicos, de acuerdo con los objetivos específicos de cada corpus:
- corpus para la descripción fonética de la lengua,
- corpus para el desarrollo de sistemas en el ámbito de las tecnologías del habla,
- transcripciones ortográficas de lengua hablada.

[3] Para el Grupo Val.Es.Co. (Valencia, Español Coloquial) la conversación es un tipo de discurso que se caracteriza por los siguientes rasgos: es oral, es decir, se articula a través del canal fónico; dialogal, lo que implica, frente al monólogo, sucesión de intercambios; inmediato, puesto que, a diferencia de un informativo o un mensaje pregrabado, se desarrolla en la coordenada espacio-temporal aquí-ahora-ante ti; retroalimentado y cooperativo, puesto que se obra juntamente con otro y su intervención; dinámico, como demuestra la alternancia de turnos, que además es no predeterminada, a diferencia de otros discursos dialogales tales como el debate, la entrevista, etc. La conversación es coloquial cuando presenta además los rasgos no planificado, lo que implica un escaso control de la producción de habla, que favorece la presencia de reinicios, vacilaciones y vueltas atrás; no transaccional, es decir, orientada a un fin interpersonal, de comunión fática, frente a la conversación transaccional, constituida como medio para obtener un fin específico. Consecuentemente, el tono de dicha conversación es informal. En suma, en la llamada conversación coloquial se reconocen, por un lado, rasgos conversacionales, relativos al tipo de discurso y, por el otro, rasgos coloquiales, propios del registro de uso.

del lenguaje infantil. Otros sistemas extendidos han sido los presentados por Blanche-Benveniste y Colette (1987) para la transcripción de un corpus sobre el francés hablado, o por Stenström (1994), Cestero (1994) y Tusón (1995) orientados al análisis de la conversación, e incluso la propuesta del propio Payrató (1995) los resume así:

— Neutralidad o fidelidad: la transcripción no debe ser interpretativa.

— Globalidad o complejidad: deben incluirse todos los fenómenos que aparecen en el discurso oral.

— Omnifuncionalidad: la transcripción debe permitir diversos usos y aplicaciones.

— Claridad: en cuanto al aprendizaje del sistema y a la legibilidad de la representación.

— Universalidad: compatibilidad entre sistemas informáticos.

En definitiva, un sistema de transcripción debe ser interpretativo de los datos, selectivo en cuanto a los fenómenos que se transcriben, pertinente para el objeto de investigación, coherente con la base teórica adoptada por el investigador, fiel en cuanto a la representación de los datos y flexible para que sea posible su utilización en estudios diversos. Por lo demás, la simbología utilizada debería ser clara, económica, sencilla, exenta de ambigüedad y compatible con sistemas estandarizados internacionalmente. En cualquiera de los casos, un sistema de transcripción será adecuado siempre que se ajuste al objeto de estudio y a la finalidad para la que se emplee, y cumpla los principios de exhaustividad y pertinencia de los signos.

En cuanto a la consideración de métodos de transcripción de uso extendido, cabe destacar el sistema EAGLES (Expert Advisory Group on Language Engineering Standards), donde se revisan diferentes elementos propios de la lengua oral, como puede observarse en la Tabla 1:

| **Nivel de análisis** | **Elementos transcritos, marcados o codificados** |
|---|---|
| Nivel silábico | Alargamiento, timbre, acento, reconstrucción de segmentos elididos. |
| Nivel léxico | Fronteras de palabras, palabras truncadas, formas no estándar, formas onomato-péyicas, formas deletreadas, acrónimos, abreviaturas, cambios entonativos en la palabra, acento léxico, pausas percibidas entre palabras o en el interior de una palabra. |
| Nivel sintáctico | Fronteras entre enunciados, modalidad, interrupciones en el enunciado con o sin presencia de pausas. |

| Nivel de análisis | Elementos transcritos, marcados o codificados |
|---|---|
| Nivel suprasegmental | **Unidades entonativas**: Fronteras entre unidades entonativas o entre unidades menores, unidades tonales incompletas o truncadas, reajustes (resets) tonales, junturas, índices de cohesión, contornos tonales terminales. |
| | **Tono**: Cambios melódicos en el enunciado o en parte del enunciado, nivel tonal, rango tonal, registro, movimiento tonal en la palabra o en el enunciado. |
| | **Acento**: Acento de palabra, acento de frase, acento tonal, niveles de acento, prominencia, énfasis, acento contrastivo, tensión, propiedades rítmicas, |
| | **Intensidad**: Intensidad absoluta o relativa de partes del enunciado |
| | **Velocidad de elocución**: Cambios en la velocidad de elocución, velocidad de elocución relativa o absoluta. |
| | **Pausas**: Pausas silenciosas, pausas vocalizadas, duración absoluta o relativa de las pausas. |
| Nivel paralingüístico | Vocalizaciones semi-léxicas, vocalizaciones no léxicas, timbre de la voz, otros elementos vocalizados (canto, gritos, etc.). |
| Nivel discursivo | Turnos de palabra, tipo de transición entre turnos, superposición de turnos. |
| Nivel contextual | Fenómenos no comunicativos no léxicos y no vocales, información kinésica. |

*Tabla 1. Elementos transcritos, codificados o marcados en el estudio de la lengua oral (EAGLES 1996).*

Por lo que atañe al ámbito hispánico, al margen del corpus de conversaciones coloquiales de Val.Es.Co., cuya versión actualizada puede consultarse en https://www.valesco.es/#/pages/cod_hj3y7hwvuuajtlkq0ik/cod_fa393ih5l4jx9zssv7, cabe destacar el Corpus de Variedades Vernáculas Malagueñas (Alvar y Villena 1994), constituido con un objetivo predominantemente sociolingüístico, el Corpus de Referencia del Español Actual (CREA) o el Corpus del Español del Siglo XXI (CORPES XXI).

El Corpus de Variedades Vernáculas Malagueñas se basa en la ortografía convencional, pero incorpora convenciones y rasgos específicos que facilitan la reconstrucción por el lector y la inclusión de características fónicas, discursivas y estilísticas (Ávila 1996:103). Su codificación sigue los estándares de la TEI (*Text Encoding and Interchange*), utilizando para ello SGML (*Standard Generalized Markup Language*).

Las etiquetas utilizadas en este corpus codifican información sobre hablantes y turnos de palabra (simultaneidad o interrupción), rasgos prosódicos (tono, intensidad, entonación, tempo, diversos tipos de pausa en función de su duración relativa y énfasis), acciones no verbales y fenómenos no vocales, actuación lingüística e incidencias que tienen lugar en la grabación, e incluso se han introducido etiquetas

para indicar características fonéticas propias de las variedades de Málaga (Ávila 1996:106).

En cuanto al Corpus de Referencia del Español Actual (CREA)[4], desarrollado por el Instituto de Lexicografía de Real Academia Española, introduce una serie de convenciones para el tratamiento de los problemas relacionados con formas reducidas de palabras, abreviaturas y acrónimos, palabras deletreadas, secuencias numéricas, interjecciones, fenómenos comunicativos no vocales, fenómenos no comunicativos no vocales, errores de producción, repeticiones, rectificaciones e interrupciones en el discurso, titubeos, etc. Se trata fundamentalmente de un corpus textual que incluye, en todo caso, transcripciones de lengua oral. Emplea los estándares de la TEI y, una vez transcrito y codificado, el texto se almacena en formato SGML. Se utilizan asimismo signos ortográficos habituales de acuerdo con la normativa de puntuación en español, excepto en el caso del punto y coma, que no se emplea en la transcripción, y de las comillas, cursivas y mayúsculas que se usan como medio tipográfico de realce.

En esta línea, el Corpus del Español del Siglo XXI (CORPES XXI)[5] divide las transcripciones de lengua oral según se ofrezcan o no alineadas con el audio. En el caso de las alineadas, el corpus codifica la trascripción con XML y tiene presente elementos de la lengua oral tales como la pausa y su duración, las palabras cortadas, y elementos paralingüísticos como la risa, el habla solapada y la vacilación.

Por su parte, el corpus de conversaciones coloquiales del Grupo Val. Es.Co. (Briz y Val.Es.Co. 1995 y 2002) se ajusta a los principios que debe cumplir todo sistema de transcripción ya señalados previamente, particularmente a los de exhaustividad (cada signo representa un único fenómeno) y de pertinencia (cada fenómeno aparece codificado mediante una única convención). Este corpus intenta reproducir fielmente la conversación, por lo que incorpora aspectos como la alternancia de turnos, la sucesión inmediata de emisiones, los solapamientos, reinicios y autointerrupciones, las escisiones conversacionales, las pausas y silencios, la entonación (inflexiones finales que influyen en el curso de la conversación con cambios respecto a la prosodia convencional), los fenómenos de énfasis, problemas relacionados con emisiones dudosas o indescifrables, aspectos de fonosintaxis, alargamientos fonéticos, etc. Se combina el método ortográfico con el propuesto por el Análisis de la Conversación, y resulta suficientemente estrecho para conseguir que el lector pueda reproducir con bastante aproximación la conversación

[4] Para más detalles sobre los criterios de diseño y confección de este corpus véase Pino y Sánchez (1999).

[5] Para más información sobre su codificación véase el enlace https://www.rae.es/corpes/assets/rae/files/corpes/codOral.pdf.

original, y suficientemente ancho para permitir una lectura fluida de la misma[6].

## 2. La trascripción de datos prosódicos

La incorporación de información sobre elementos suprasegmentales en un corpus oral plantea diversos problemas, que se concretan en las variaciones continuas de F0, de intensidad o de duración. Por tanto, es necesario un proceso de abstracción para determinar cuáles de estas variaciones son lingüísticamente significativas y cómo se relacionan con categorías discretas. Finalmente, tales categorías deberían ser representadas en un sistema de notación.

En conjunto, pues, un sistema ideal de transcripción prosódica debería permitirnos la representación en varios niveles, ser compatible con el intercambio electrónico de datos y cubrir las necesidades del mayor número de lenguas posible. Ahora bien, en ausencia de un sistema que reúna tales características, parece adecuado elaborar mecanismos de compatibilidad entre los ya existentes a fin de facilitar la reutilización de los datos.

Al respecto, en el marco del proyecto ESPRIT SAM se han desarrollado sistemas de transcripción prosódica compatibles con las necesidades de anotación de bases de datos en soporte electrónico (Llisterri 1997), resultando de ello propuestas como PROSPA, SAMSINT o SAMPROSA (Gibbon 1989; Wells *et al*. 1992). De todos ellos, un conjunto de símbolos que recibió bastante aceptación para la transcripción prosódica fue el de SAMPROSA (*SAM Prosodic Alphabet*), propuesto por Gibbon (1989) y desarrollado por Wells *et al*. (1992) hasta llegar al formato que se presenta en la Tabla 2, tomada de Wells (1995):

| SAM-PROSA | ASCII | Definition | SAM-PROSA | ASCII | Definition |
|---|---|---|---|---|---|
| **Local tone** | | | | | |
| H | 72 | High pitch | M | 77 | Mid pitch |
| L | 76 | Low pitch | + | 43 | High pitch |
| T | 84 | Top pitch (extreme H) | ++ | 43,44 | Much higher pitch |
| B | 66 | Bottom pitch (extreme H) | +- | 43,45 | Peak (upward-downward) |

[6] Las conversaciones transcritas por Val.Es.Co. constituyen de hecho una forma de transcripción semiestrecha. En función del objetivo perseguido, la transcripción puede estrecharse (integrando, p. e., aspectos prosódicos ausentes antes) o ensancharse (prescindiendo, p. e., de reinicios y solapamientos, si no son pertinentes para el estudio). Por ejemplo, Hidalgo (1997), aplicando el sistema de base de Val.es.Co. (véase en Anexo las convenciones de transcripción), desarrolla una versión estrecha de este sistema, que incluye, además de los rasgos generales otros como la frecuencia fundamental (F0) de cada uno de los grupos entonativos segmentados.

| SAM-PROSA | ASCII | Definition | SAM-PROSA | ASCII | Definition |
|---|---|---|---|---|---|
| - | 45 | Lower pitch | ^^ | 94,94 | Wide upstep |
| -- | 45,45 | Much lower pitch | ! | 33 | Downstep |
| -+ | 45,43 | Trough (downward-upward) | !! | 33,33 | Wide downstep |
| ^ | 94 | Upstep | = or > or S | 61,62,83 | Level or same tone |
| **Global tone: from local and nuclear tone repertoire** | | | | | |
| **Terminal tone: from local and nuclear tone repertoire** | | | | | |
| **Nuclear tone** | | | **Pause** | | |
| - | 45 | Level tone (before tone group boundary) | ... | 46,46,46 | Silence |
| ' or / or R | 39,47,82 | Rising tone | **Boundary** | | |
| ` or \ or F | 96,92,70 | Falling tone | $ | 36 | Syllable boundary |
| `' (etc.) | 96,39 (etc.) | Fall-rise | # | 35 | Word boundary |
| '` (etc.) | 39,96 (etc.) | Rise-fall | \| | 124 | Tone group boundary / non-directional) |
| **Length** | | | [ | 91 | Tone group boundary (left) |
| : | 58 | Segmental length mark | ] | 93 | Tone group boundary (right) |
| **Stress** | | | **Metasymbols** | | |
| " | 34 | Primary stress | - | 45 | Separator (the underscore, _, ASCII 95, may replace this owing to ambiguity with level tone) |
| % | 37 | Secondary stress | * | 42 | Conjunctor |

*Tabla 2. Transcripción de datos prosódicos en SAMPROSA (Wells, 1995).*

En todo caso, el sistema de notación prosódica más seguido en la actualidad es, sin duda, ToBI (*Tone and Break Index*). Este método incluye dos tipos de símbolos, los que representan la estructura tonal subyacente (*tones*) y los que marcan los límites entre unidades prosódicas (*break indices*)[7]. ToBI recoge el inventario de unidades entonativas propuestas

[7] Sin embargo, se han dirigido ciertas críticas hacia este sistema ToBI entre las cuales figura, por una parte, su estrecha dependencia del modelo fonológico desarrollado por Pierrehumbert (1980)

por Pierrehumbert (1980) e introduce algunas especificaciones respecto del comportamiento tonal que se traducen en diferentes componentes (Hualde 2003: 157-173):

1. **Tono**: H si es alto (*high*), L si es bajo (*low*) y M si es medio (*mid*).

2. **Acento tonal**: tono o secuencia de tonos fonológicamente asociado con una sílaba acentuada, que viene señalado con un asterisco (*).

3. **Tono de frontera**: tono asociado con el límite de una frase y que va marcado con el símbolo %.

4. **Escalonamiento**: los picos pueden estar distribuidos de manera descendente (*downstep*) de manera predecible o por algún contraste pragmático (!) o de manera ascendente (*upstep*), si marcan un pico significativamente superior al previo (¡).

5. **Frases prosódicas**: pueden ser entonativas o intermedias. La frase entonativa está constituida por frases intermedias con una separación menor entre sí y constituidas por elementos tonales llamados acentos de frase (L-/H-). Estos acentos combinados con los tonos de frontera dan lugar a cuatro posibilidades: descenso (L-L%), descenso con ascenso final (L-H%), suspensión (H-L%), ascenso (H-H%) (Beckman *et al*. 2002).

El sistema ToBI ha tratado de superar la disparidad de soluciones a la hora de transcribir la entonación mediante un inventario cerrado de acentos tonales. Para el español (Sp_ToBI) trabajos como los de Sosa (2003) o Estebas-Vilaplana y Prieto (2008) tratan de precisar la propuesta de Beckman *et al*. (2002), e incluso se ha llegado a proponer una herramienta de transcripción automática en Praat (Elvira-García *et al*. 2015). Asimismo, una de las principales virtudes de ToBI es permitir la representación por separado de las unidades prosódicas en que se organizan los enunciados (*jerarquía entonativa*) y los fenómenos entonativos (en términos de tonos).

No obstante lo anterior, uno de los problemas que plantea el sistema ToBI es que se trata de un procedimiento de anotación fonológica y no fonética que, además, resulta compleja en su proceso de aprendizaje y no recoge de modo preciso las diferencias de rango tonal (fenómenos

y, por otra, su mejor adaptación al inglés que a otras lenguas, lo que explicaría su amplia utilización en Estados Unidos. No obstante, se han realizado trabajos en italiano, alemán, húngaro y español, entre otras lenguas, usando ToBI como sistema de representación. Un inconveniente más importante es que la anotación mediante ToBI requiere un cierto conocimiento previo de los patrones entonativos de la lengua. Aun así, es un sistema que ofrece indudables ventajas, tales como la presentación de una estructura jerárquica, que permite seleccionar de entre subconjuntos o conjuntos mayores de símbolos, la posibilidad de representar problemas que aparecen en la transcripción o el hecho de que existan experimentos que muestran un elevado grado de acuerdo entre transcriptores diferentes (Pitrelli *et al*., 1994).

de altura, fenómenos de reajuste tonal, etc.) ni otros aspectos como los alargamientos, etc.

Otro sistema de transcripción prosódica vinculado a la lengua oral espontánea corresponde a Chafe (1993), con un modelo de unidades basado en su segmentación prosódica, donde se utiliza una serie de símbolos para representar fenómenos como los alargamientos (=), los acentos primario (^^) y secundario (^) y los límites entre unidades de acento (|), además de la coma y el punto que diferencian tipos de contornos melódicos de continuación o cierre.

Por su parte, Mertens (2004) propone un modelo de transcripción prosódica denominado prosograma. Este sistema de transcripción intenta cumplir varios criterios:

1. Es objetivo, robusto y fácil de interpretar, además de la representar la entonación perceptiva.

2. Fija la evolución de largos fragmentos de habla.

3. Esta fijación es cuantificada para ser estimada por intervalos melódicos.

4. Se preserva la organización temporal con el fin de observar pausas e interrupciones.

5. La transcripción es semi-automática.

6. Es neutral, es decir, independiente de cualquier modelo con el que pueda ser usada por investigadores de distintos marcos teóricos.

7. Hay una alineación entre sonido y texto.

8. Permite la manipulación y la evaluación de la transcripción.

El sistema de Mertens permite, además, cuatro variantes según se quiera (1) representar la curva entonativa perceptiva, (2) añadir la curva de intensidad y de F0, (3) añadir la cuantificación temporal y de F0 o (4) incluir toda la información prosódica (curva entonativa perceptiva, curva de intensidad y de F0 y cuantificación temporal y de F0).

Finalmente, Cantero (2002), Cantero y Font (2007 y 2009), Cantero y Mateo (2011) y Mateo (2010) han desarrollado el modelo de Análisis Melódico de Habla, que conceptualiza la entonación como complejo prelingüístico, lingüístico y paralingüístico. El modelo de transcripción prosódico de Cantero y Font (2007) distingue rasgos melódicos y fonológicos. Los rasgos melódicos se analizan a partir de la estilización de la curva entonativa entendida como un conjunto de elementos estructurales. Como modelo de representación, el protocolo del AMH plantea dos pasos. El primer paso consiste en la extracción del valor de

F0 de cada vocal. El segundo paso consiste en estandarizar los datos de F0, de manera que la primera vocal de un segmento dado toma un valor igual a 100 y se recalculan mediante reglas de tres las subidas y bajadas tonales de manera porcentual; los valores positivos implican una subida tonal y los negativos, bajadas (Cantero y Font 2009). Este modelo ha sido complementado recientemente con el denominado Análisis Prosódico del Habla (APH) (Cantero 2019), que integra simultáneamente el tono, la intensidad y la duración.

## 3. Propuesta de implementación de datos prosódicos a partir del sistema de transcripción de Val.Es.Co.

Al margen de lo dicho hasta aquí, la transcripción de un corpus de lengua oral plantea muy diversos problemas, tales como la existencia de variaciones fonéticas no recogidas en diccionarios normativos, el uso de signos de puntuación o la representación de siglas, las palabras deletreadas, las secuencias numéricas, etc. A ello se añaden otros problemas derivados de la necesidad de delimitar enunciados y unidades tonales, la variación entre elementos suprasegmentales, el establecimiento oportuno de pausas (silencios), las pausas oralizadas o llenas, la representación de elementos paralingüísticos o cuasi-léxicos (risas, toses, etc.), los cambios de turno, las intervenciones simultáneas de dos o más hablantes, las palabras o construcciones truncadas, las repeticiones o errores de producción, etc. Sin duda, la codificación de todos estos elementos obliga a enriquecer la transcripción ortográfica, lo que resulta imprescindible para proceder adecuadamente al uso y análisis de lo transcrito, esto es, a su interpretación.

Quedan, pues, sin resolver algunos problemas vinculados a la falta de objetividad en cuanto a la información y al análisis, y a la dificultad de la representación y de lectura de datos prosódicos. Presentamos por ello a continuación una propuesta de transcripción prosódica que atiende a los principios siguientes:

a. Criterios claros y objetivos basados en datos acústicos.

b. Objetivación de los resultados mediante el empleo de herramientas informáticas adecuadas.

c. Facilidad en la lectura.

d. Posibilidad de elaboración del corpus de manera individual y autónoma.

Con el fin de satisfacer estos principios, nuestra propuesta de transcripción prosódica (3.4.) va precedida de tres apartados que afectan

a la segmentación en grupos entonativos (3.1.), a la identificación y representación de la curva entonativa (3.2.) y a las herramientas de transcripción y análisis (3.3.).

## *3.1. Criterios para la segmentación en grupos entonativos*

El concepto de grupo entonativo y su identificación ha sido ampliamente por Quilis, Cantarero y Esgueva (1993), Cabedo (2011a y 2011b) e Hidalgo (2018 y 2019).

En primer lugar, Quilis, Cantarero y Esgueva (1993) proponen que los fenómenos pausa y/o inflexión de F0 son criterios claros para la identificación de límites prosódicos. En segundo lugar, Cabedo (2011a) delimita la rentabilidad de los fenómenos pausa, reajuste tonal, inflexión y alargamiento en su fórmula estadística (MESTEL) para la identificación de límite prosódico, tal como se indica en la Figura 1:

| **Fórmula** | | **Resultado** |
|---|---|---|
| $\frac{1}{1 + 2.71^{(4.97 - (0.01 \times \mathbf{60}) - (1.88 \times \mathbf{1.4}) - (0.004 \times \mathbf{70}) - (17.7 \times \mathbf{0.2}))}}$ | = | 0.89 (89%) |

*Figura 1. Ejemplo de aplicación de la fórmula MESTEL*[8]

En tercer lugar, Hidalgo (2019: 98) propone varios criterios operativos para la identificación de límites prosódicos[9]:

— Pausa mayor o igual de 300 ms[10].

— Reajuste tonal positivo o negativo mayor o igual de 3 st.

[8] MESTEL contiene una serie de constantes asociadas a parámetros que miden la importancia de cada variable para el reconocimiento de fronteras demarcativas:
— constante matemática *e* (2.71);
— constante establecida por la prueba estadística (4.97);
— constante de duración (1.88);
— constante de reajuste (0.004);
— constante de inflexión (0.01);
— constante de pausa (17.7).
Todos estos valores se relacionan en última instancia con los de cada punto concreto del discurso sobre el que queramos determinar su carácter de frontera/no frontera. Con ello se obtiene un valor probabilístico de importancia para cada factor: puede calcularse así la probabilidad de que una frontera melódica se constituya como frontera efectiva de grupo entonativo. Por ejemplo, para un determinado segmento de habla, si existe una pausa posterior de 0,2 segundos, una duración del segmento entonativo de 1.4 segundos, una inflexión melódica del 60 % y un reajuste tonal del 70 %, la probabilidad de que dicho segmento sea grupo entonativo sería del 89 %. Este es el caso del ejemplo formulado en la Figura 1.

[9] El Corpus Val.Es.Co. 3.0 emplea de hecho los criterios de la pausa (≥ 300 ms) y el reajuste (3 st) para la identificación de límites prosódicos (Pons 2022).

[10] Se mantienen así los criterios aplicados en el marco del sistema de transcripción del Corpus Val.Es.Co. 3.0: pausa (≥ 300 ms) y reajuste (3 st) para la identificación de límites prosódicos (Pons 2022).

— Inflexión tonal mayor o igual de 3 st.

— Duración de la sílaba final del GE (grupo entonativo) cuando sea al menos el doble respecto de la media del GE.

Así pues, los fenómenos que indican la presencia de un límite y que delimitan un GE se pueden agrupar en dos bloques según su naturaleza: duración (duración de la pausa y alargamiento) y frecuencia fundamental (reajuste tonal e inflexión).

Por lo que respecta a la representación del límite prosódico, el sistema de transcripción Val.Es.Co. (Briz y Val.Es.Co. 2002) ofrece una serie de convenciones que permiten señalar la presencia de estos fenómenos demarcativos, a excepción del fenómeno de reajuste tonal, que no cuenta con una forma de transcripción, pero sí ha sido incluido en nuestra propuesta como se verá en 3.4.:

a. En primer lugar, la pausa se puede señalar con / (300-500 ms), // (500-1000 ms) y /// (>1000 ms). En este último caso, se señala entre paréntesis la duración de la pausa en segundos[11].

b. En segundo lugar, el alargamiento de la vocal se señala con la duplicación de la vocal alargada (aa, ee, ii, oo, uu).

c. En tercer lugar, la inflexión tonal se marca con flechas ↑ (entonación ascendente), ↓ (entonación descendente), → (entonación mantenida o suspendida), y el acento circunflejo ^ (entonación circunfleja).

## *3.2. Criterios para la identificación y la representación de la curva entonativa*

Como hemos visto previamente en § 2, para la representación de la curva entonativa es especialmente productivo el sistema de transcripción Tone and Break Index (ToBI), que permite representar, con un número reducido de símbolos, la diversidad entonativa. No obstante, su lectura resulta poco transparente para usuarios no expertos. Por su parte, el modelo de Mertens o el Análisis Melódico del Habla permiten visualizar de manera más clara la curva entonativa. Sin embargo, tanto estos modelos como el sistema ToBI, necesitan de una segunda línea de transcripción para su clara visualización, lo que va en detrimento de la claridad en la transcripción y de la facilidad de lectura.

Por su parte, el sistema de Val.Es.Co. permite transcribir las inflexiones finales mediante las flechas ↑, ↓ y → y el acento circunflejo ^, pero no ofrece información sobre el inicio o el cuerpo del GE. Por ello,

[11] Las razones que justifican esta decisión metodológica se derivan de la experiencia de transcripción de años llevada a cabo en el marco del grupo Val.Es.Co, tal como se refleja en la introducción de Briz y Val.Es.Co. (2002).

Hidalgo (1997, 2002) propone transcribir la F0 inicial y final de cada GE en hercios. Esta información permite indicar la dirección tonal del GE y ofrece información acerca del reajuste tonal entre GE sucesivos.

La cuantificación de la F0 en hercios, sin embargo, supone un problema a nivel técnico, puesto que los hombres presentan una media de 120 hercios frente a los 200 que suelen presentar las mujeres (González et al. 2002), por lo que habría que identificar el sexo de los hablantes en la transcripción de la conversación. Así pues, el empleo de cifras numéricas en hercios dificulta la labor de transcripción frente a otras propuestas para la identificación del límite prosódico que emplean semitonos, por dos motivos: por una parte, la extracción de valores en hercios requiere de una estandarización posterior para evitar las diferencias de sexo que ya viene dada con la unidad semitono y, por otra parte, la extracción de valores en semitonos facilita la identificación de límites prosódicos y el empleo de la misma unidad resulta más operativa para la obtención de todos los datos prosódicos, evitando así la variedad de unidades de medida para un mismo fenómeno.

### *3.3. Herramientas de transcripción y análisis*

Los criterios previos necesitan de herramientas informáticas que objetiven la transcripción. En este sentido, el programa ELAN[12] resulta operativo para la transcripción prosódica, debido a que permite, por una parte, sincronizar el audio y el texto y, por otra, interactuar con el programa Praat (Boersma y Weenink 2022), que facilita la información prosódica relativa a la duración, la F0 y la intensidad. ELAN es una herramienta de transcripción offline que emplea el lenguaje XML y que permite añadir capas o niveles de anotación con los que añadir texto o comentarios de una manera jerarquizada, alineada con el audio y, además, fácilmente exportable a otros formatos. Por su parte, Praat puede automatizar la extracción de datos acústicos en análisis más precisos mediante el uso de archivos Textgrid y el empleo de scripts como Analyze_tier_modify[13].

Con todo ello, la propuesta de transcripción prosódica que presentamos aquí obedece al siguiente protocolo:

1. Grabación de audio según el objetivo de la investigación.

2. Importación del archivo de audio en formato .wav al programa ELAN.

[12] El siguiente enlace ofrece información detallada sobre el funcionamiento de este programa: https://archive.mpi.nl/tla/elan

[13] Disponible en: http://uk.groups.yahoo.com/group/praat-users/files/Daniel_Hirst/analyse_tier.praat

3. Creación de líneas de transcripción: una por hablante y otras que se adapten al objetivo de la transcripción.

4. Transcripción del texto por cada grupo fónico o espiratorio.

5. Observación y cuantificación de la duración de los silencios dentro de cada grupo fónico. De esta manera, los silencios iguales o superiores a 300 ms constituyen una pausa, hecho que señala un límite prosódico, por lo que la caja de transcripción deberá ser dividida, como se observa en la Figura 2:

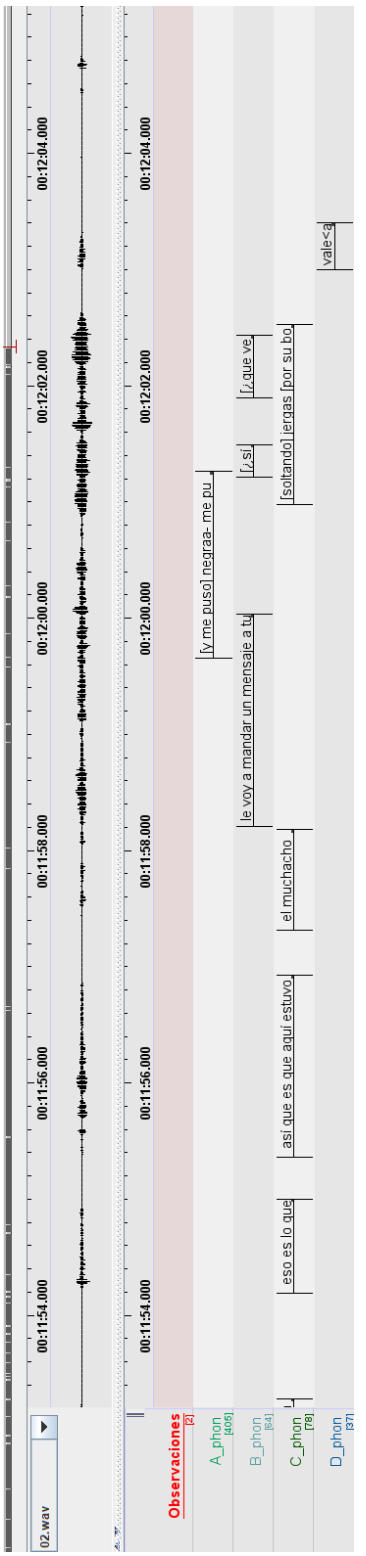



*Figura 2. Fragmento de transcripción en ELAN.*

Llegados a este punto, dispondremos de una transcripción en GE que solo tienen presente la pausa como límite prosódico, por lo que la identificación de los demás fenómenos deberá llevarse a cabo de manera o bien *manual*, o bien *automatizada*.

En primer lugar, para la identificación manual proponemos los siguientes pasos:

Abrir cada caja de ELAN en Praat, observar los posibles saltos tonales existentes entre palabras y dividir las cajas en los casos en los que se den saltos iguales o superiores a 3 st, hecho que señala un reajuste tonal. Este salto se mide tomando el valor en semitonos del final de una palabra y restando el valor inicial en semitonos de la palabra siguiente. Véase al respecto la Figura 3:

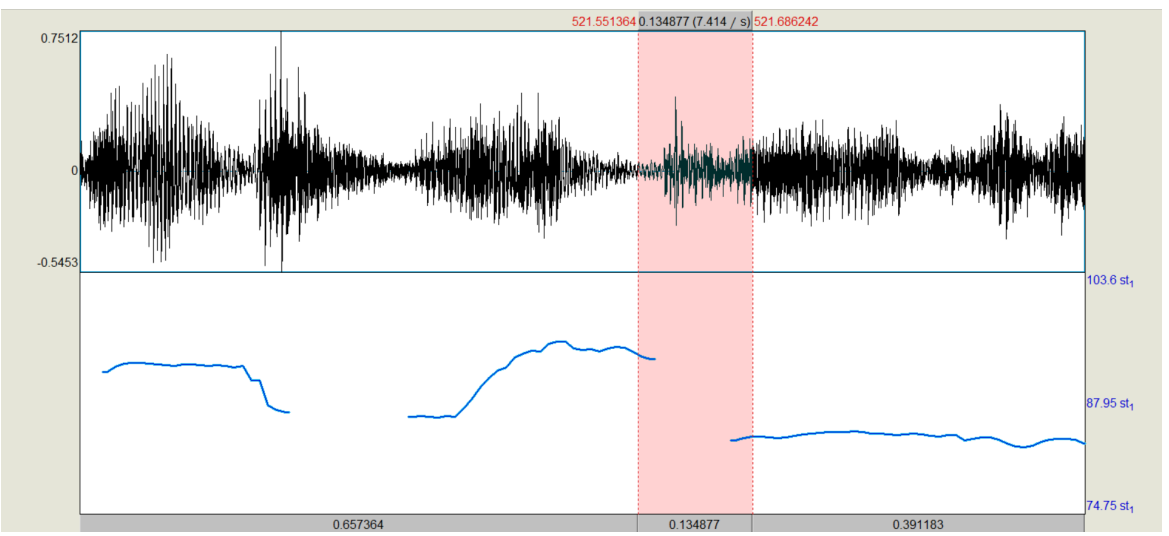



*Figura 3. Representación sombreada del reajuste tonal en Praat.*

Observar la presencia de inflexión tonal final de GE, esto es, cuando la diferencia entre la sílaba tónica y el final de la palabra del GE es igual o superior a 3 semitonos. En este caso, podemos señalar la presencia de un tonema ascendente o descendente según la dirección de la curva. Véase al respecto la Figura 4:

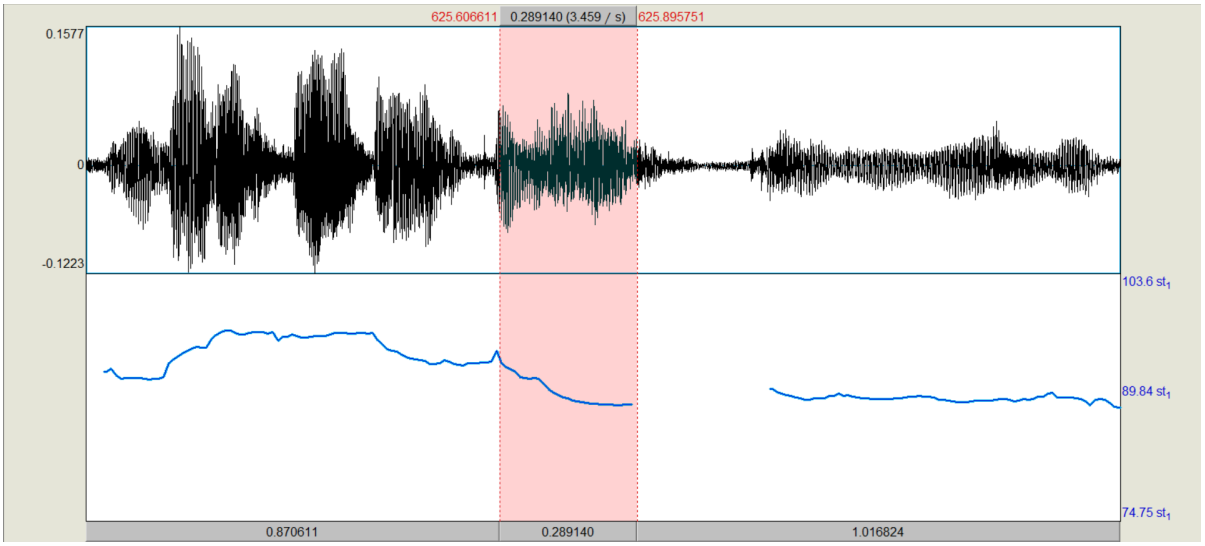



*Figura 4. Representación sombreada de la inflexión tonal en Praat.*

La identificación del alargamiento puede hacerse de modo automatizado mediante el uso de Textgrid y el script mencionado anteriormente (Analyze_tier_modify), que permiten obtener la información de manera más objetiva y completa. Así pues, el uso de Textgrid exige la selección de «películas»[14] en Praat con las que se crea un archivo de audio de manera automática que permite la edición y la anotación. Para crear el Textgrid en Praat debemos clicar en la opción Annotate > To TextGrid y crear las líneas Palabra y Sílaba.

Después de crear los TextGrid de todas las cajas de ELAN, hay que guardar los archivos de audio (*sound*) en una carpeta y los de texto (*TextGrid*) en otra. Posteriormente, se instala el script Analyze_tier_modify y se abren las carpetas con todos los fragmentos anotados. El programa nos ofrecerá un archivo .txt que puede ser importado a una hoja Excel con los valores totales de duración, medias máximas y mínimas de F0 e intensidad de cada fragmento, tal como puede observarse en la Figura 5:

[14] ELAN permite la extracción de fragmentos de audio ya editables gracias a la opción Selección de «películas» con Praat.

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | file | label | duration | f0_min | f0_mean | f0_max | int_min | int_mean | int_max |
| 2 | # file : 02_1 | | | | | | | | | |
| 3 | # min_pitch: 135; median_pitch: 203 max_pitch 336; | | | | | | | | | |
| 4 | 1 | 02_1 | di | 115.741 | 88.579 | 90.129 | 92.291 | 53.102 | 60.016 | 62.750 |
| 5 | 2 | 02_1 | solapado | 1.427.201 | 83.976 | 92.626 | 97.242 | 53.978 | 70.023 | 75.201 |
| 6 | 3 | 02_1 | al | 174.516 | 89.643 | 90.767 | 92.929 | 46.640 | 58.499 | 62.755 |
| 7 | 4 | 02_1 | sue | 162.229 | 92.394 | 92.952 | 94.094 | 46.772 | 57.698 | 63.262 |
| 8 | 5 | 02_1 | gro | 186.563 | 93.037 | 93.603 | 94.753 | 45.095 | 60.686 | 64.402 |
| 9 | 6 | 02_1 | por | 75.031 | 92.815 | 93.183 | 93.974 | 44.547 | 54.972 | 58.948 |
| 10 | 7 | 02_1 | que | 139.922 | 91.194 | 91.822 | 93.799 | 48.077 | 56.154 | 59.080 |
| 11 | 8 | 02_1 | van | 141.950 | 91.064 | 91.717 | 92.071 | 49.834 | 58.993 | 61.944 |
| 12 | 9 | 02_1 | a | 40.557 | 91.299 | 92.683 | 94.294 | 51.930 | 54.512 | 55.801 |
| 13 | 10 | 02_1 | rre | 119.644 | 91.582 | 91.839 | 93.049 | 56.972 | 59.840 | 61.496 |
| 14 | 11 | 02_1 | glar | 451.657 | 91.481 | 92.056 | 92.869 | 49.380 | 59.862 | 63.340 |
| 15 | | | | | | | | | | |
| 16 | # min_pitch: 134; median_pitch: 190 max_pitch 300; | | | | | | | | | |
| 17 | 12 | 02_2 | el | 124.219 | 88.572 | 90.616 | 96.558 | 51.055 | 54.128 | 56.503 |
| 18 | 13 | 02_2 | sue | 117.317 | 91.480 | 92.075 | 94.400 | 48.458 | 61.986 | 63.839 |
| 19 | 14 | 02_2 | gro | 97.012 | 91.706 | 92.197 | 92.652 | 60.827 | 64.315 | 66.346 |
| 20 | 15 | 02_2 | lo | 74.451 | 91.957 | 92.116 | 92.197 | 56.906 | 59.513 | 61.688 |
| 21 | 16 | 02_2 | tie | 49.634 | 91.758 | 92.860 | 94.530 | 50.620 | 57.189 | 59.437 |
| 22 | 17 | 02_2 | ne | 133.109 | 90.237 | 91.474 | 92.764 | 51.424 | 54.256 | 57.184 |
| 23 | 18 | 02_2 | ca | 49.634 | 90.598 | 92.124 | 93.702 | 53.660 | 56.123 | 58.710 |
| 24 | 19 | 02_2 | si | 115.061 | 90.278 | 92.116 | 93.673 | 48.243 | 54.311 | 57.973 |
| 25 | 20 | 02_2 | a | 56.402 | 89.661 | 89.954 | 90.278 | 50.698 | 57.104 | 59.873 |
| 26 | 21 | 02_2 | rre | 157.926 | 89.504 | 90.639 | 95.395 | 47.672 | 57.013 | 60.331 |

*Figura 5. Datos importados en Excel.*

Sobre los datos obtenidos se pueden realizar operaciones matemáticas para la obtención de valores necesarios para la identificación de fenómenos prosódicos. A continuación, explicamos brevemente dichas operaciones y los valores que permiten obtener:

a. Recolocación: el script utilizado ofrece valores máximos y mínimos de F0, sin embargo, son los valores inicial y final los que aportan información relevante para la identificación de la inflexión y el reajuste tonal. Por ello, la dirección de la curva se cruza con los valores máximos y mínimos, de manera que en una curva ascendente el valor mínimo es inicial y el máximo, final y en una curva descendente el valor máximo es inicial y el mínimo, final.

b. Resta: una vez identificados los valores iniciales y finales de cada segmento se restan los segundos a los primeros y se obtiene el valor del salto tonal con lo que se identificarán las posibles inflexiones o reajustes tonales.

c. Porcentaje: para obtener el alargamiento de una sílaba cabe obtener la media de duración de las sílabas del GE y calcular mediante una regla de tres si alguna sílaba final de palabra tiene un valor igual o superior a 200 % de la media (véase al respecto lo indicado en §3.1. en este sentido).

Obtendremos de esta manera los datos relativos a la identificación de límites prosódicos (pausa, alargamiento, reajuste tonal e inflexión) y del contorno entonativo (tonema y valores inicial y final).

En cualquier caso, a pesar de las ventajas de objetividad que supone el análisis acústico, sea manual o automatizado, a la hora de transcribir el discurso oral, existen ciertas limitaciones para la segmentación y anotación prosódica:

1. En los casos de pérdida de línea de frecuencia fundamental, solapamiento, habla dudosa, elementos paralingüísticos (risas, toses, bostezos) o en el habla entre risas, solo es posible identificar de manera manual la pausa y no se puede identificar ningún otro fenómeno.

2. Hay ocasiones en los que la curva final del GE no toma una dirección clara por lo que es difícil determinar el tipo de tonema; en este caso, la opción Stylize Pitch de Praat puede ayudar a determinar la dirección tonal precisa.

3. El tonema circunflejo no puede identificarse de manera automatizada, por lo que se requiere una revisión de los finales de GE para identificar la posible presencia de estos tonemas.

4. Los monosílabos tónicos y las palabras agudas suelen presentar una curva tonal propia pero corta, por lo que cabría observar la dirección general del GE y sobre todo de la sílaba previa. En el caso de los monosílabos átonos, estos forman parte de la coda del GE.

### *3.4. Propuesta de transcripción*

Vistas las consideraciones previas, nuestra propuesta de transcripción incluye información sobre límites y sobre el contorno prosódicos. Los fenómenos tenidos en cuenta son la pausa, el alargamiento, la inflexión, el tonema, la F0 inicial y final de grupo entonativo y el reajuste tonal. Algunos fenómenos sirven exclusivamente para la identificación de límites (pausa) o para la descripción del contorno (F0 inicial y final), pero otros sirven para ambos, como es el caso de la inflexión y el tonema, que sirven para la identificación del final de un grupo entonativo (límite) y para la representación del curva melódica (contorno), o el alargamiento final de grupo entonativo que da información sobre el final de esta unidad entonativa y sobre su ritmo interno. Finalmente, como se ha señalado, el sistema Val.Es.Co. actualmente no ofrece ningún símbolo para la representación del reajuste tonal, por lo que los valores iniciales y finales de F0 (en st) transcritos en el superíndice y subíndice, respectivamente, señalarán el salto tonal existente entre grupos entonativos. La Tabla 3 sintetiza la información prosódica que refleja nuestra propuesta y su modo de representación:

| Fenómenos prosódicos | Forma de representación |
|---|---|
| Presencia de silencio no pausa | (< 300 ms) |
| Pausa corta, inferior al medio segundo | / (300-500 ms) |
| Pausa entre medio segundo y un segundo | // (500-1000 ms) |
| Pausa de un segundo o más | /// (> 1000 ms) |
| Alargamiento | aa, ee, ii, oo, uu |
| Inflexión ascendente / Tonema ascendente | ↑ |
| Inflexión descendente / Tonema descendente | ↓ |
| Tonema suspendido | → |
| Inflexión circunfleja / Tonema ascendente | ^ |
| F0 inicial de grupo entonativo | $^{90.12}$digo |
| F0 final de grupo entonativo | arreglaar$_{92.05}$ |

*Tabla 3. Sistema de transcripción de datos prosódicos*

Véase en el siguiente fragmento de una conversación coloquial cuál es el resultado final de la transcripción según nuestra propuesta:

A: $^{90.58}$lo que no me hace graciaa$_{79.35}$↓(045) $^{92.22}$es que con mi hija- mi hija me pone muy$_{87.81}$↓ $^{75.19}$negra$_{87.60}$↑ $^{75.51}$¿eh? $_{80.50}$↑ (012) $^{92.50}$te lo juro$_{88.36}$↑//(679) $^{90.13}$yo de todas se lo dije digo *vamos a ver*$_{89.13}$→/(496) $^{93.42}$*¿qué me dijiste?* digo yo$_{88.69}$→ $^{93.09}$*que no caí en ese momento bien que tiene que dar las nóminas ella*$_{95.99}$→//(595) $^{90.12}$digo [*¿por qué tienes que dar las nóminas?*]§

B: [¿ella tiene que dar las nóminas?]

A: § º($^{90.76}$al suegro porque van a arreglaar$_{92.05}$)º→º($^{90.61}$el suegro lo tiene casi arregla(d)o paraa$_{87.06}$)º→

B: $^{85.52}$la hipoteca$_{95.96}$ (213) [para que la hipo]teca sea para ella solo$_{79.80}$↓

A: [digo](1.098) $^{95.75}$ no$_{89.31}$↓(179) $^{91.89}$para ampliar se ve$_{75.56}$↓(298) $^{86.37}$algo$_{93.67}$↑// (629) $^{89.80}$digo [*mira Jimena*]

C: [yo creo que para] ampliarlo$_{81.65}$↑

A: $^{94.01}$digo *vamos a ver* º(*no es que tenemos que*$_{89.50}$)º↓ $^{90.95}$digo$_{88.17}$↓(032) $^{93.71}$*vamos a ver*$_{91.55}$↓//(922) $^{93.11}$digo *haz el favor*$_{89.15}$↓(057) $^{94.88}$*de explicar un poco tu padre sabe muy bien de bancos*$_{87.89}$↓ (042) $^{89.04}$digo *a ver si vaa a encargarse el Manuel de TODO*$_{97.06}$↑///(1.221) $^{91.50}$*que yo quiero confiaar*$^{96.97}$↑/(313) $^{89.64}$*pero a ver si va arreglar*$_{95.74}$↑(152) $^{94.80}$*cara a que su hiijo salga beneficia(d)o*$_{97.98}$↓/(347) $^{91.28}$*y tú no eres capaz de decirle a tu padre na(da)*↓//(928) $^{99.13}$*no*$_{99.13}$→ $^{90.81}$*porque*

*él ya se arreglaará lo que tenga que arreglar* digo *¡hombre!* $_{91.76}$↑/ (315) $^{89.33}$*haz el favor ¿eh?*$_{98.53}$↑/(355) $^{98.53}$*así que tú coges y le preguntas*$_{87.47}$→(234) $^{93.30}$es que está en un plan$_{95.10}$↑///(1.316) $^{95.83}$es que está en un$_{88.43}$→(044) $^{92.39}$plaan quee-$_{86.22}$→//(598) $^{91.55}$que no$_{86.31}$↓(079)

(Conversación 2016.PT.(20).S6 del Corpus Val.Es.Co. 3.0)

## 4. Conclusiones

Para concluir, el sistema propuesto en este trabajo parece mejorar sistemas anteriores de transcripción donde también se trataban de incorporar los aspectos prosódicos del habla. Ello se debe a varias razones objetivas, entre las que destacamos las siguientes:

— No resulta tan confuso o prolijo como otros sistemas de transcripción prosódica (SAMPROSA, ToBI…).

— Permite una lectura directa de las convenciones empleadas para representar los datos suprasegmentales.

— Cumple los requisitos exigibles a todo sistema de transcripción, particularmente los de exhaustividad y pertinencia.

— Incorpora parámetros no considerados por otros sistemas de transcripción, como es el caso del reajuste tonal.

— Sustituye la cuantificación en hercios de la F0 por su valoración en semitonos (st), lo que hace posible el estudio simultáneo de datos procedentes de individuos de diferente sexo.

— Emplea aplicaciones informáticas útiles que permiten la exportación de los datos y favorecen la realización de un análisis semiautomático de los mismos; así, para la medición del reajuste tonal y la determinación del tipo de tonema se puede llevar a cabo una medición manual, mientras que para la identificación de los alargamientos cabe aplicar un método automatizado.

En todo caso, dada la naturaleza específica del tipo de discurso que se está transcribiendo, el registro coloquial, no siempre el análisis acústico permite obtener datos inequívocos válidos para una transcripción «fiel» de la prosodia discursiva. Los solapamientos de habla, los ruidos, las risas, ciertas alteraciones «extrañas» de la melodía debidas a causas idiosincrásicas o contextuales, etc., son todos ellos factores que no podemos obviar, por cuanto el sistema aquí propuesto, pese a sus ventajas, deberá ser implementado en fases ulteriores con objeto de mejorar en lo posible sus prestaciones. Es necesario, pues, probar todavía

su rentabilidad en sucesivos trabajos de transcripción prosódica que permitan determinar los implementos metodológicos necesarios para su progresivo perfeccionamiento.

## Bibliografía

Alvar Ezquerra, Manuel, y Juan Andrés Villena Ponsoda (coords.) (1994), *Estudios para un corpus del español*, Málaga: Universidad de Málaga.

Atkinson, J. Maxwell, y John Heritage (eds.) (1984), *Structures of social action: studies in conversation analysis*, Cambridge/París, Cambridge University Press/Éditions de la Maison des Sciences de l'Homme.

Ávila Muñoz, Antonio Manuel, y Juan Andrés Villena Ponsoda (2010), *Variación social del léxico disponible en la ciudad de Málaga*, Málaga, Sarriá.

Ávila Muñoz, Antonio Manuel (1996), «Problemas prácticos en la realización de corpus orales: la transliteración del corpus oral del proyecto de investigación de las variedades vernáculas malagueñas (VUM)», en Juan de Dios Luque Durán y Antonio Pamies Bertrán (eds.), *Actas del Primer Simposio de Historiografía Lingüística*, Granada, Método Ediciones: 103-112.

Beckman, Mary, Manuel Díaz-Campos, Julia Tevis McGory y Terrell Morgan (2002), «Intonation across Spanish, in the Tones and Break Indices framework», *Probus*, 14: 9-36.

Blanche-Benveniste, Claire, y Jeanjean Colette (1987), *Le français parlé: transcription et édition*, París, Didier Eridition.

Boersma, Paul, y David Weenink (2022), Praat: doing phonetics by computer [Programa informático]. Version 6.2.14. Disponible en http://www.praat.org/.

Briz, Antonio (1998), *El español coloquial en la conversación: esbozo de pragmagramática*, Barcelona, Ariel.

Briz, Antonio, y Grupo Val.Es.Co. (1995), *La conversación coloquial: materiales para su estudio*, Anejo XVI de *Cuadernos de Filología*, Valencia, Universitat de València.

Briz, Antonio, y Grupo Val.Es.Co. (2002), C*orpus de conversaciones coloquiales*, Anejo I de la revista *Oralia*, Madrid, Arco Libros.

Button, Graham, y John Lee (eds.) (1987), *Talk and social organization*, Clevendon, Multilingual Matters.

Cabedo, Adrián (2011a), «Hacia un modelo predictivo para la segmentación prosódica del discurso oral coloquial: MESTEL (Modelo Estadístico para la Selección de Términos Entonativos Ligados)», *Oralia*, 14: 85-104.

Cabedo, Adrián (2011b), «El reajuste tonal en la delimitación de grupos entonativos», en Antonio Hidalgo Navarro, Yolanda Congosto Martín y Mercedes Quilis Merín (eds.), *El estudio de la prosodia en españa en el siglo xxi, perspectivas y ámbitos*, Valencia, Universitat de València: 209-222

Cantero, Francisco José (2002), *Teoría y análisis de la entonación*, Barcelona, Edicions Universitat de Barcelona.

Cantero, Francisco José (2019), «Análisis prosódico del habla: más allá de la melodía», en María Rosa Álvarez, Silva Álex Muñoz y Alvarado Leonel Ruiz Miyares (eds.) *Comunicación social: lingüística, medios masivos, arte, etnología, folclor y otras ciencias afines*, Volumen II, Santiago de Cuba, Ediciones Centro de Lingüística Aplicada: 485-498.

Cantero, Francisco José, y Dolors Font (2007), «Entonación del español peninsular en habla espontánea: patrones melódicos y márgenes de dispersión», *Moenia*, 13: 69-92.

Cantero, Francisco José, y Dolors Font (2009), «Protocolo para el análisis melódico del habla», *Estudios de Fonética Experimental*, 18: 19-32.

Cantero, Francisco José, y Miguel Mateo (2011), «Análisis melódico del habla: complejidad y entonación en el discurso», *Oralia*, 14: 105-127.

Cestero, Ana María (1994), *Análisis de la conversación: alternancia de turnos en la lengua española*, tesis doctoral, Universidad de Alcalá de Henares.

Chafe, Wallace (1993), «Prosodic and functional units of language», en Jane A. Edwards y Martin D. Lampert (eds.), *Transcription and coding in discourse research*, Nueva Jersey, Lawrence Erlbaum Associates.

DuBois, John W. (1991), «Transcription design principles for spoken discourse research», *Pragmatics*, 1:71-106.

DuBois, John W., Stephan Schuetze-Coburn, Susanna Cumming, y Danae Paolino (1993), «Outline of discourse transcription», en Jane A. Edwards y Martin D. Lampert (eds.), *Talking data: transcription and coding in doscourse research*, Hillsdale, Lawrence Erlbaum Associates: 45-90.

EAGLES (1996), *Preliminary recommendations on spoken texts*, EAGLES Document EAG-TCWG-STP/P.

ELAN (Version 6.8) [Computer software]. (2024), *Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive*. Recuperado de https://archive.mpi.nl/tla/elan.

Elvira-García, Wendy, Paolo Roseano, Ana Fernández Planas y Eugenio Martínez Celdrán (2015), «Una herramienta para la transcripción prosódica automática con etiquetas Sp_ToBI en Praat», en Antonio Hidalgo y Adrián Cabedo (eds.), *Perspectivas actuales en el análisis fónico del habla: tradición y avances en la fonética experimental*, Valencia, Universitat de València: 455-464.

Estebas Vilaplana, Eva, y Pilar Prieto (2008), «La notación prosódica del español: una revisión del Sp-ToBI», *Estudios de Fonética Experimental*, 17: 264-283.

Garrido, Juan María (1991a), *Modelización de patrones melódicos del español para la síntesis y el reconocimiento de habla*, Bellaterra, Universitat Autònoma de Barcelona.

Garrido, Juan María (1991b), «Modelización de patrones melódicos del español para sistemas de conversión texto-habla», *Procesamiento del Lenguaje Natural*, 11: 209-219.

Gibbon, Dafydd (1989), *Survey of Prosodic Labelling for EC Languages. SAM-UBI-1/90, 12 february 1989; Report e.6, en ESPRIT 2589 (SAM) Interim Report, Year 1. Ref. SAM-UCL G002*, University College London.

González, Julio, Teresa Cervera y José Luis Miralles (2002), «Análisis acústico de la voz: fiabilidad de un conjunto de parámetros multidimensionales», *Acta Otorrinolaringológica Española*, 53 (4): 256-268.

Gumperz, John J., y Norine Berenz (1993), «Transcribing conversational exchanges», en Jane A. Edwards y Martin D. Lampert (eds.), *Talking data: transcription and coding in discourse research*, Hillsdale, Lawrence Erlbaum Associates: 91-122.

Hidalgo, Antonio (1997), *La entonación coloquial: función demarcativa y unidades de habla*. Valencia, Anejo XXI de *Cuadernos de Filología*, Valencia, Universitat de València.

Hidalgo, Antonio (2002), *Comentario fónico de textos coloquiales*, Madrid, Arco Libros.

Hidalgo, Antonio (2018), «Unidades discursivas mínimas en la conversación: una aproximación de base prosódico-contextual», en Ester Brenes Peña, Marina González-Sanz y Francisco Grande Alija (eds.), *Enunciado y discurso: estructura y relaciones*, Sevilla, Universidad de Sevilla: 229-250

Hidalgo, Antonio (2019), *Sistema y uso de la entonación en español hablado: aproximación interactivo-funcional*, Santiago de Chile, Universidad Alberto Hurtado.

Hualde, José Ignacio (2003), «El modelo métrico y autosegmental», en Pilar Prieto (coord.), *Teorías de la entonación*, Barcelona, Ariel: 155-184.

Llisterri, Joaquim (1997), *Transcripción, etiquetado y codificación de corpus orales, Seminario de Industrias de la Lengua, Curso Etiquetado y extracción de información de grandes corpus textuales*, Soria, Fundación Duques de Soria.

MacWhinney, Brian (1991), *The CHILDES project: tools for analysis talk*, Hillsdale, Lawrence Erlbaum.

Mateo, Miguel (2010), «Protocolo para la extracción de datos tonales y curva estándar en análisis melódico del habla (AMH)», *Phonica*, 6: 49-90.

Mertens, Piet (2004), «The prosogram: semi-automatic transcription of prosody based on a tonal perception model», en Bernard Bel y Isabelle Marlien (eds.), *Proceedings of Speech Prosody*, Nata: 23-26.

Navarro Tomás, Tomás (1944), *Manual de entonación española*, Nueva York, Hispanic Institute.

Navarro Tomás, Tomás (1982), *Manual de pronunciación española*, Madrid, CSIC.

Ochs, Elinor (1979), «Transcription as a theory», en Elinor Ochs y Bambi B. Schieffelin (eds.), *Developmental pragmatics*, Nueva York, Academic Press: 43-72.

Payrató, Lluís (1995), «Transcripción del discurso coloquial», en Luis Cortés Rodríguez (ed.), *El español coloquial: actas del I*

*Simposio sobre Análisis del Discurso Oral*, Almería, Universidad de Almería: 43-70.

Pierrehumbert, Janet B. (1980), *The phonology and phonetics of English intonation*, tesis doctoral, MIT.

Pino, Marta, y Mercedes Sánchez (1999), «El subcorpus oral del banco de datos CREA-CORDE (Real Academia Española), procedimientos de transcripciones y codificación», *Oralia*, 2:83-138.

Pitrelli, John, Mary E. Beckman, y Julia Hirschberg (1994), «Evaluation of prosodic transcription labelling reliability in the ToBI framework», en *Proceedings of the Third International Conference on Spoken Language Processing*, Yokohama, ICSLP, vol. 2: 123-126.

Pons Bordería, Salvador (2022), *Creación y análisis de corpus orales*, Berna, Peter Lang.

Quilis, Antonio (1975), «Las unidades de la entonación», en *Revista Española de Lingüística*, 5 (2): 261-280.

Quilis, Antonio, Margarita Cantarero, y Manuel Esgueva (1993), «El grupo fónico y el grupo de entonación en español hablado», *Revista de Filología Española*, 73: 55-64.

Sosa, Juan Manuel (2003), «La notación tonal del español en el modelo Sp-ToBI», en Pilar Prieto (ed.) *Teorías de la entonación*, Barcelona, Ariel: 185-208

Stenström, Anna-Brita (1994), *An introduction to spoken interaction*, Londres/Nueva York, Longman.

t'Hart, Johan, René Collier, y Antoine Cohen (1990), *A perceptual study of intonation: an experimental-phonetic approach to intonation*, Cambridge, Cambridge University Press.

Tannen, Deborah (1987), *Conversational style: analyzing talk about friends*, Norwood, Ablex.

Torruella, Joan, y Joaquim Llisterri, (1999), «Diseño de corpus textuales y orales», en José Manuel Blecua *et al*. (eds.), *Filología e informática: nuevas tecnologías en los estudios filológicos*, Barcelona, Editorial Milenio/Universitat Autònoma de Barcelona.

Tusón, Amparo (1995), *Anàlisi de la conversa*, Barcelona, Empúries.

Wells, John (1995), SAMPROSA (SAM Prosodic Transcription), Disponible en: http: //www.icp.grenet.fr/SpeechDat/home.html.

Wells, John, William Barry, Martine Grice, Adrian Fourcin, y Dafydd Gibbon (1992), *Standard Computer-Compatible Transcription. SAM Stage Report Sen. 3 SAM UCL-037, 28 February 1992, en SAM (1992) ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/ Output Assessment, Methodology and Standardisation. Final Report. Year Three: I.III.91-28.II.92*, Londres, University College London.

Salvador Pons Bordería (ORCID: 0000-0001-5788-5506) y Sara Badia-Climent (ORCID: 0000-0002-7993-5997),
«De la transcripción al análisis: desarrollos técnicos del corpus Val.Es.Co. 3.0».
*Biblioteca de Babel: Revista de Filología Hispánica*, vol. extra. 2 (2024), pp. 79-113.
**DOI**: 10.15366/bibliotecababel2024.extra2.003 · **ISSNe**: 2695-6349 · Recibido: 11/09/2024 · Aceptado: 03/11/2024

# De la transcripción al análisis: desarrollos técnicos del corpus Val.Es.Co. 3.0[1]

Salvador Pons Bordería
*Universitat de València*
salvador.pons@uv.es

Sara Badia Climent
*Universitat de València*
sara.badia@uv.es

**Resumen**: Este artículo aborda las decisiones teóricas y técnicas adoptadas en la creación del corpus Val.Es.Co. 3.0. El objetivo principal es detallar todos los pasos que se han seguido para lograr crear un corpus oral accesible en formato digital que no solo permita trabajar a los usuarios con el contenido de las transcripciones y el etiquetado de los fenómenos discursivos, sino también con el análisis de su contenido a partir del modelo de unidades del grupo de investigación: subactos, actos, intervenciones, diálogos y discursos. Para ello, el trabajo se ha dividido en tres secciones. En primer lugar, se hace una breve introducción sobre los dos pilares fundamentales que componen el trabajo de creación del corpus Val.Es.Co. 3.0.: la transcripción y su posterior segmentación. En segundo, se describe el proceso de trabajo técnico interno que cada una de las conversaciones ha recibido, desde la transcripción hasta la segmentación de unidades. Por último, la tercera sección expone cómo se visualizan las transcripciones en la web del corpus y detalla las distintas funcionalidades que este pone a disposición de los usuarios.

**Palabras clave**: corpus oral, lingüística computacional, segmentación discursiva, transcripción, corpus Val.Es.Co., español hablado.

[1] Este trabajo ha sido posible gracias al proyecto CIPROM/2021/038 Hacia la caracterización diacrónica del siglo xx (DIA20), del proyecto PROMETEO de la Generalitat Valenciana, y al proyecto de I+D+I PID2021-125222NB-I00 Aportaciones para una caracterizacion diacronica del siglo xx, financiado por MCIN/AEI/10.13039/501100011033/ y por FEDER Una manera de hacer Europa. Los autores agradecen a los revisores anónimos sus sugerencias y comentarios, que han mejorado notablemente la versión final de este artículo.

## From transcription to analysis: technical developments of the Val.Es.Co. 3.0 corpus

**Abstract**: This article examines the theoretical and technical decisions involved in the elaboration of the Val.Es.Co. 3.0 corpus. Its main goal is to detail the steps taken to develop an accessible digital oral corpus. The Val.Es.Co. 3.0 corpus provides users with spontaneous conversations and a system of discourse-based tags. It also analizes a subset of conversations with the Val.Es.Co. model of discourse units: subacts, acts, interventions, dialogues, and discourses. This article is divided into three sections. Section two outlines the two basic pillars of the creation process for the Val.Es.Co. 3.0 corpus: transcription and its subsequent analysis. Section 3 describes the backend, especially the technical decisions adopted during the processes of transcription and discourse segmentation. Finally, Section 4 explains how the transcriptions are displayed on the website and details the corpus frontend main features.

**Keywords**: oral corpus, computational linguistics, discourse segmentation, transcription, corpus Val.Es.Co., spoken Spanish.

## 1. Introducción

El desarrollo de corpus lingüísticos permite acercar la lingüística al lenguaje empleado por los hablantes en contextos reales (Bolaños 2015; Bernal y Hincapie 2018; García-Miguel 2022), ya sea en su forma oral o escrita. Si bien es cierto que los corpus de discursos escritos cuentan con una serie de dificultades relativas, generalmente, al formato en el que se plasma el texto y a los datos que sitúan el discurso (Torruella y Llisterri 1999), los corpus basados en discursos orales se enfrentan a otro tipo de dificultades, puesto que, por su carácter inmediato (Briz 2010), plantean cuestiones específicas que van desde la fase misma de recolección hasta su procesamiento informático (Briz 1996, Pons Bordería 2022).

Estos problemas aumentan cuando el corpus incluye, además de una transcripción, un análisis del material lingüístico, y no solo por las cuestiones teóricas relativas al tipo de interpretación, sino por las dificultades técnicas que plantea incorporar una carga de información adicional al texto transcrito. Dichas dificultades afectan tanto al diseño del *frontend* (interfaz, disposición de las ventanas de resultados, botones de búsqueda, etc.) como, sobre todo, al *backend* (secuenciación de la información, diseño de la hoja de ELAN, vocabulario de metadatos,

estructura del corpus, etc.). Este trabajo pretende abordar dichas cuestiones a partir del trabajo realizado en la elaboración de la versión 3.0 del corpus Val.Es.Co. (Valencia Español Coloquial), lo que puede ser de interés para la lingüística de corpus en general y para la del español hablado en particular.

Este artículo se organiza en tres partes: primero, se revisarán las características relevantes para el diseño de un corpus en el ámbito del español hablado. En segundo lugar, se mostrarán las decisiones informáticas tomadas en el corpus Val.Es.Co. 3.0. y los pasos seguidos para lograr su adecuada visualización en la página web y su correspondiente exportación. Por último, se mostrará cómo se visualizan las conversaciones en la página web y las distintas funcionalidades accesibles para los usuarios.

## 2. Estado de la cuestión: el trabajo de creación de un corpus

El interés por los corpus orales en la lingüística cobra especial relevancia con el desarrollo del Análisis de la Conversación (Sacks, Schegloff y Jefferson 1974)[2] y, en el ámbito hispánico, con los trabajos pioneros de Criado de Val (1964) y el proyecto PILEI dirigido por Lope Blanch (Lope Blanch 1971, 1976, 1986). Este interés teórico supuso, además, el desarrollo de una metodología para el reflejo del habla , ya que los sistemas ortográficos se revelaron insuficientes para tal fin. Gail Jefferson fue la primera investigadora en abordar estas cuestiones y en desarrollar un sistema de transcripción propio de las conversaciones, denominado hoy en día *jeffersoniano* (Jefferson y Sacks 2000; Jefferson 2004; Margret *et al*. 2009; Bassi 2015). Esta metodología implica aspectos tales como la transcripción de las conversaciones y su etiquetado (§ 2.1) y, en un nivel diferente de estudio, su análisis (§ 2.2.).

### *2.1. La transcripción de la conversación*

Transcribir la oralidad conversacional implica lograr un equilibrio entre fidelidad y utilidad: por un lado, la transcripción debe ser lo suficientemente estrecha como para reflejar los fenómenos que cada corpus pretende reflejar; por otro, debe ser adecuada a los intereses del investigador, es decir, no contar con una sobrecarga de información que no permita o dificulte el estudio de su objeto (Pons Bordería 2022).

[2] La etnometodología del habla (Garfinkel 1967) ya había abordado la recogida y transcripción de material lingüístico oral; no obstante, sus intereses se centraban, esencialmente, en el análisis de las características sociales de los individuos y su reflejo en el habla (Zimmerman 1978; O'Keefe 1979), a diferencia de las disciplinas lingüísticas, que buscan analizar el lenguaje en sí mismo o en relación con rasgos de los individuos, del contexto o la sociedad.

Esta cuestión más general se puede concretar con la pregunta de cuánta información debe contener la transcripción: tratándose un corpus de lenguaje oral, una primera respuesta podría ser «todo aquello que se dice». Pero comunicarse, especialmente cuando se trata de discursos cara a cara, va más allá de las palabras: la entonación, el paralenguaje, los gestos e, incluso, las acciones extralingüísticas cobran un papel tan fundamental que, en ocasiones, dotan de sentido a determinados enunciados que, sin dicha información, serían incomprensibles (Cestero 2014; Cabanes 2023). En este punto el investigador debe buscar un equilibrio entre su objeto de estudio, que puede ser variable, y la cantidad de información que decida incluir en su corpus.

Se sigue de esto que no existe una única forma de transcribir, sino varias, en función del carácter más o menos amplio de la investigación. En este sentido, Pons Bordería (2022: 43) distingue hasta cinco niveles de detalle en la transcripción, que dependen del grado de complejidad y de la cantidad de fenómenos que incluya:

Nivel I: codificación ortográfica.

Nivel II: codificación según los principios del Análisis de la Conversación.

Nivel III: codificación de nivel II con el añadido de la información prosódica.

Nivel IV: codificación de nivel III con el añadido de información kinésica y paralingüística.

Nivel V: codificación de nivel IV con el añadido de vídeo(s) de los participantes.

Asimismo, dentro del nivel de información kinésica y paralingüística (niveles III y IV) pueden codificarse distintos tipos de información (Poyatos 1994, 2018), desde la gesticulación voluntaria o involuntaria del cuerpo hasta las actividades realizadas por los participantes, pasando por el registro de aquellos elementos contextuales o situacionales que puedan influenciar la comunicación (Poyatos 2018: 23-24; Cabanes 2023).

Muchos de los corpus de español hablado actuales se sitúan entre los niveles II y III, como ocurre con el corpus del Proyecto para el estudio sociolingüístico del español de España y de América (PRESEEA), el Macrocorpus de la norma lingüística culta de las principales ciudades de España y América (MC-NC), el Corpus Oral del Lenguaje Adolescente (COLA) o el corpus AMERESCO, entre otros (Rojo 2016; Briz y Carcelén 2019). Si bien aprovechan la ortografía del

español, reflejan, al mismo tiempo, ciertos fenómenos de la oralidad. Por ejemplo, la elisión de la *d* intervocálica en los participios como en *llega(d)o* o los alargamientos vocálicos mediante la repetición de la letra del sonido correspondiente como en *uun*. Suele ser frecuente, en el caso de los corpus de conversaciones, señalar además los solapamientos ([ ]), las intervenciones inmediatas (§) o las pausas en los enunciados (/). Por último, los corpus orales también suelen dan cuenta de ciertos fenómenos paralingüísticos, como las risas (RISAS), o ciertas acciones necesarias para entender la conversación.

Los corpus se consultan en abierto a través de la red mediante un sistema de búsqueda. Este debe permitir la identificación no solo de las palabras, sino también de la información de la que el grupo de trabajo en cuestión haya decidido dar cuenta. Para ello, el estándar que se ha impuesto desde hace años se basa en el lenguaje de etiquetado XML (Santamaría 1999; Brun 2005) y, en el ámbito de la lingüística, frecuentemente en el sistema TEI (Alcaraz y Vázquez 2016; Del Rio y Allés-Torrent 2023). No obstante, este sistema se ha ido modificando y adaptando de acuerdo con los objetivos e intereses de cada grupo de investigación.

Así, en el texto, las etiquetas ofrecen información sobre fenómenos conversacionales (solapamientos, énfasis, habla simultánea), prosodia, elisiones o acortamientos y observaciones necesarias para entender la transcripción. Gracias a este sistema el usuario de los corpus orales puede consultar transcripciones que, además de dar cuenta de las palabras de las grabaciones, identifican también estructuras lingüísticas o fenómenos discursivos, como se puede ver en los siguientes ejemplos:

(1) **PRESEEA – BARR_H22_037**
**I**: <tiempo = "02:33"/> yo de Vill <palabra_cortada/> de Villanueva<alargamiento/> añoro<alargamiento/> // principalmente mis padres // mis viejos // y añoro la paz / que<alargamiento/> // que <vacilación/> se perdió<alargamiento/> // o sea se ha ido perdiendo // eeh // esa tranquilidad<alargamiento/> // esa gente sana // de la cual yo yo <vacilación/> dejé // o sea sí<alargamiento/> había / problemas / normales / como en todos las regiones yo pienso que // del mundo pero // pero<alargamiento/>…

En (1), correspondiente al proyecto PRESEEA, se puede observar el uso de distintas etiquetas, como la localización temporal del fragmento (<tiempo = "02:33"/>), ya que sus transcripciones se transcriben directamente sobre un editor de textos y necesitan mantener una conexión entre el texto y el audio. También se pueden ver ejemplos de

etiquetado relativos a características de la oralidad como las palabras cortadas (<palabra_cortada/>), los alargamientos (<alargamiento/>) o las vacilaciones (<vacilación/>), entre otros fenómenos.

(2) **ESLORA - SCOM_H13_013**
**hab2**: <ininteligible/>
**hab1**: yo qué sé cómo están <ruido tipo="chasquido boca"/> <pausa_larga/>
**hab1**: están ampliando las aceras <pausa/> y están hacien y están pintando muchas fachadas de muchos edificios <pausa/> pues le están dando un toque más <pausa/> unificado a la zona nueva <pausa/> o sea algo más <pausa/> la están haciendo un poco más bonita <pausa/> de lo f <pausa/> o sea partiendo de lo fea que es <risa/> <pausa/>
**hab2**: intentan arreglándola un po
**hab1**: <ininteligible/> sí intentando arreglar un poco y <pausa_larga/>

El corpus ESLORA (Corpus para el estudio del español oral) también emplea el sistema de etiquetado XML. En (2), se observa el uso de etiquetas que marcan elementos de paralenguaje, como el chasquido de la lengua (<ruido tipo="chasquido boca"/>). Además, se indican dos tipos de pausas y se utilizan etiquetas específicas, como (<ininteligible/>), para señalar aquellas partes del audio que no pueden ser transcritas.

Al señalar fenómenos conversacionales, el etiquetado permite que las transcripciones no se limiten a registrar únicamente las palabras de los hablantes, sino que incluyan también la representación de diversos fenómenos lingüísticos y paralingüísticos, lo que facilita el procesamiento de la información, la búsqueda automatizada y la extracción de datos, de manera que los corpus permitan consultas complejas que van más allá del contenido literal del habla. Además, este sistema posibilita el recuento automático de los fenómenos identificados mediante las etiquetas, lo que simplifica la realización de estudios estadísticos.

## *2.2. De la transcripción al análisis de los datos orales*

Un paso más en el proceso de elaboración de corpus consiste en añadir a la transcripción un análisis de los datos. Al incluir este objetivo, se supera la mera representación del material transcrito para adentrarse en el campo de la interpretación[3].

[3] En realidad, toda transcripción es, en cierta medida, una interpretación. Pero si en una transcripción la interpretación busca una representación más o menos fidedigna de la realidad, un análisis busca una explicación del material transcrito. Así, aunque la subjetividad permee ambas

Una de las razones para añadir esta capa extra de información es abordar el problema de la *sintaxis de lo hablado*. La obra de Antonio Narbona (en especial Narbona 1989a, 1989b, 1990) demuestra la inadecuación de la sintaxis oracional para dar cuenta de la organización de la materia hablada en los discursos orales coloquiales, prototípicamente representados por las conversaciones espontáneas. Se inicia así en la lingüística española la pregunta de cómo explicar la estructura de lo hablado desde una base no sintáctica. Los modelos de segmentación discursiva, desarrollados en su mayoría en las lenguas románicas (Pons Bordería 2014.), ofrecen respuestas a esta cuestión desde varios criterios, a menudo superpuestos: prosódico, semántico, sintáctico o pragmático.

El grupo Val.Es.Co ha desarrollado un modelo de este tipo (Briz *et al*. 2003, Briz y Grupo Val.Es.Co. 2014, Pons Bordería 2022) que divide la conversación coloquial en unidades y subunidades sin residuo de un modo similar a como se procede en un análisis sintáctico. Como resultado de este proceso, quedan situados en «ámbitos de estudio diferentes los fenómenos lingüísticos discursivos y, en concreto, del español hablado […] [Además,] se evita así la casuística y la descripción aislada» (Val.Es.Co. 2014: 12). Para comprobar la adecuación teórica del modelo al objeto de estudio, se ha analizado una parte del corpus 3.0 (más de treinta y cinco mil palabras) con dicho modelo y se ha incorporado al corpus en la web. El objetivo de este intento es mostrar la capacidad del modelo para responder a la pregunta de Narbona mediante una contrastación empírica amplia.

La incorporación de este modelo al corpus ha supuesto un reto teórico, pero también aplicado: ha sido necesario incluir en el diseño y elaboración del corpus una serie de procedimientos complejos para incluir este análisis interpretativo al puro y simple proceso de transcripción. La sección § 3 explica cómo se ha llevado a cabo este desarrollo desde un punto de vista técnico.

## 3. La elaboración del corpus Val.Es.Co. 3.0: decisiones técnicas

Aunque el corpus Val.Es.Co. es una obra colectiva que manifiesta su continuidad desde 1995 (Briz *et al*. 1995), su versión 3.0 ha introducido cambios importantes en la forma en la que se tratan las conversaciones del corpus, hasta el punto de que se podría hablar de una refundación del corpus mismo. Asimismo, la introducción del análisis ha obligado a modificaciones de calado en el diseño del corpus. Todo esto ha obligado a abordar las cuestiones que se detallan en esta sección, en la

operaciones, lo hace orientada a dos objetivos completamente diversos.

que se explicará el funcionamiento del *backend* del corpus Val.Es.Co. 3.0., que es la base de su visualización en la versión web. Para ello, se tratarán cuestiones relativas al sistema de transcripción y etiquetado (§ 3.1), a la configuración y organización de las líneas de análisis del programa de transcripción ELAN (§ 3.2.) y a la metodología empleada para segmentar y analizar una conversación (§ 3.3).

## *3.1. Cambios en la transcripción*

El carácter oral, conversacional y coloquial del corpus Val.Es.Co hace necesario reflejar información que va más allá de lo meramente ortográfico:

1) El contenido de las intervenciones.

2) Información prosódica.

3) Información interactiva, que incluye fenómenos dialógicos, como los solapamientos o las interrupciones.

4) Los sucesos extralingüísticos que afectan directamente a la conversación.

A estas cuatro metas se ha añadido, en la versión 3.0, un quinto objetivo:

5) El análisis del contenido de los enunciados (segmentación de unidades de la conversación).

En 1995, el grupo Val.Es.Co. adaptó al español el sistema de transcripción jeffersoniano, que se especializa en los fenómenos conversacionales más relevantes en una conversación: solapamientos, toma de turno, pausas, silencios, tonemas o realizaciones paralingüísticas, entre otros. Este sistema aprovecha los símbolos que ofrece un teclado ASCII para asociar fenómenos conversacionales a signos del teclado, como se muestra en la Tabla 1:

| Fenómeno | Sistema de transcripción Val.Es.Co. |
|---|---|
| Mantenimiento del turno de un participante en un solapamiento | = |
| Lugar donde se inicia un solapamiento o superposición | [ |
| Final del habla simultánea | ] |
| Entonación suspendida | → |
| Entonación ascendente | ↑ |
| Entonación descendente | ↓ |
| Entonación circunfleja | ^ |
| Fragmento ininteligible | (()) |
| Pausa de menos de medio segundo | / |

*Tabla 1. Signos conversacionales del sistema de transcripción Val.Es.Co. (versión 1995).*

(3) **Conversación 2011.PT.S2**
**50A21**: [es la-] es la entidad de la Comunidad Valenciana↑ ¡no! de Europa↑/ [que mejor] paga↓ tía↑ s[í]
**51C27**: [de la Unión EURO]PE que mejor [paga→]/ [a los monitores] [¡hombre!]=
**52B26**: [((¡qué barbaridad!))] [(( )) qué- ¡qué suerte tenéis!]
**53C27**: =[mil nove]cientos euros al mes↑ pues de lunes a viernes por hacer el ((tonto)) [por la mañana]

(4) **Conversación 2011.PT.S5**
**24B13**: ¿el qué? ¿el qué?
**25A12**: que a Vero le he cogido más aprecio↑/ desde que va con vosotras porque me la he encontrao más y aunque ahora vaya con los jarcoretas sigue siendo igual→

El sistema inicial de transcripción aprovechaba los símbolos del teclado para indicar, por ejemplo, los solapamientos con corchetes ([ ]) o el mantenimiento de un turno con el signo igual (=) , pero no disponía de marca asociada para otros elementos, como los fenómenos de fonética sintáctica (destacado en negrita en 4). Este sistema jeffersoniano es un instrumento muy adecuado para leer las conversaciones, algo necesario para los investigadores en pragmática o en análisis conversacional, ya que el comportamiento interactivo de los participantes se desarrolla durante todo el acontecimiento comunicativo y es preciso tener una visión detallada de dicha relación junto a la forma en la que se se está creando material lingüístico. Sin embargo, no es un método adecuado para su procesamiento informático, ya que no cumple las exigencias básicas de toda búsqueda automatizada. Una de las más evidentes es que no todos los símbolos se empleen de forma unívoca.

Un ejemplo claro de esto son los paréntesis, que, en Briz *et al*. (1995), se empleaban para la transcripción dudosa «((palabra dudosa))», la ininteligible «(( ))», los susurros «º(texto susurrado)º» o la delimitación de fenómenos paralingüísticos «(RISAS)». Otro problema consiste en que los símbolos pueden coincidir con la transcripción, como la duplicación de vocales o consonantes, que puede formar parte tanto del texto transcrito (en palabras como *creen* o *innegable*) como del sistema de transcripción (el fenómeno del alargamiento se representaba, precisamente, mediante dos vocales o consonantes idénticas (*Que yo no lo he heecho*).

A partir de 2019[4], y siguiendo los estándares en el campo, el sistema inicial se actualizó para convertirse en un sistema de base XML, en el que las convenciones de transcripción se asociaron a etiquetas. De esta forma, cada suceso podía ser reconocido de manera unívoca por un buscador.

Asimismo, se aprovechó este cambio de sistema para incluir en la transcripción algunos fenómenos que no contaban con ningún tipo de marca explícita previa, como la fonética sintáctica, o que presentaban ambigüedades de lectura, como los alargamientos:

| **Fenómeno** | **Etiquetas TEI** | **Transcripción tradicional** |
|---|---|---|
| Estilo directo | <cita>palabras</cita> | *palabra* |
| Alargamientos | palabra<al/> | palabraa |
| Fragmentos ininteligibles | <in/> | (( )) |
| Fonética sintáctica | <fsr t=”l’almohada”>la almohada</fsr> | l’almohada |
| Tonema ascendente | <ta/> | ↑ |
| Fragmentos entre risas | <e_risas>palabras</e_risas> | Palabras(ENTRE RISAS) |

*Tabla 2. Ejemplo del sistema de transcripción con etiquetas y su traducción al modelo tradicional.*

Esta ventaja, sin embargo, se enfrenta a un inconveniente: la transcripción se sobrecarga con etiquetas que dificultan la lectura frente al sistema tradicional, tal y como ilustran los ejemplos (5a) y (5b):

(5) **Conversación 1994.PT.3**

(5a) **187A88**: [sí<al/>] <e_risas><in/> un pueblerino</e_risas> // [<e_risas>y para decir masovero</e_risas>] no podía decir ni campo ni nada y decía <cita>los que están más <fsr t=”p’allá”>para allá</fsr><ta/></cita> y todo el mundo

[4] Como resultado del proyecto de investigación UDEMADIS (FFI-2016-77841-P).

<cita>ovni estraterrestre</cita> <risas/> <cita>¡no no!</cita> <cita>más <fsr t="p'allá">para allá</cita> pero un poco más <fsr t="p'acá">para acá</fsr> [y mira]

(5b) **187A88**: [síí] (( )) un pueblerino(ENTRE RISAS)// [y para decir masovero(ENTRE RISAS)] no podía decir ni campo ni nada y decía *los que están más p'allá*↑ y todo el mundo *ovni estraterrestre* (RISAS) *¡no no! más p'allá [y mira]*

La versión etiquetada, más completa desde el punto de vista de la transcripción, plantea, sin embargo, un dilema: una transcripción como la jeffersoniana, perfectamente adaptada a su objeto de estudio, no permite un tratamiento informatizado; por su parte, una transcripción de fenómenos conversacionales con el etiquetado XML convierte el texto resultante en algo difícil para la lectura y el análisis del investigador, y todo esto aunque se empleen etiquetas comprensibles como «<ininteligible/>» para fragmentos de audio no comprensibles.

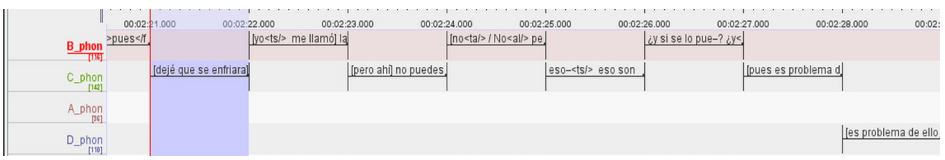
La forma de resolver dicho dilema en el corpus 3.0 ha sido la de mantener ambos tipos de transcripciones, situando cada una en dos partes distintas del corpus: un *backend*, en el que las conversaciones están etiquetadas y sus elementos constitutivos pueden ser recuperados mediante un buscador, y un *frontend* que devuelve al usuario final la conversación transcrita mediante el sistema jeffersoniano original, como se verá en § 4.1. Para lograrlo, a través de la aplicación web del corpus se ha desarrollado un sistema de traducción que vincula cada etiqueta con el símbolo o formato correspondiente en la transcripción tradicional, de manera que en la web del corpus las conversaciones tengan un formato legible, pero sin perder las ventajas informáticas del etiquetado.

No obstante, las novedades de la versión 3.0 no se limitan solo a la combinación de la transcripción tradicional con el sistema TEI/XML; la transcripción ha incorporado herramientas como el programa ELAN, que permite una organización más detallada y flexible de los datos. A continuación, se describe cómo se ha estructurado la hoja de trabajo de ELAN en el corpus 3.0, que, como se verá, presenta una complejidad considerable.

## *3.2. Una hoja de ELAN compleja*

El programa ELAN es, hoy en día, uno de los estándares más utilizados para la transcripción de audio y video. Esta herramienta permite integrar información de diferentes formatos (audio, video y texto) y separar los datos de la transcripción por niveles, lo que permite exportar la información para su cuantificación y posterior procesamiento. Para

ello, este *software* permite la creación de líneas (*tiers*) en las que pueden clasificarse los fenómenos que se quiere estudiar. Al estar asociadas dichas líneas a una marca temporal, se hace posible asociar fenómenos lingüísticos al lugar de la transcripción en que aparezcan y al hablante




que los produce:
*Imagen 1. Interfaz de ELAN con anotaciones.*

Una transcripción en ELAN utiliza un sistema *en pentagrama* (Vázquez *et al*. 2021), que tiene la forma de un papiro (Pons Bordería 2022) que se despliega de forma horizontal sobre el eje temporal (en la parte superior) y que consta de una línea por cada hablante. Dentro de cada una de las líneas se crean cajas alineadas temporalmente con el audio para cada fragmento de transcripción.

Estas cajas pueden diseñarse según diferentes criterios, pero el estándar comúnmente empleado, tanto en corpus alineados como en tradicionales, ha sido identificar la intervención o turno como la unidad base para la transcripción. Esta elección se fundamenta en la necesidad de segmentar el discurso en unidades que permitan analizar tanto las dinámicas de toma de turno como la estructura interna de la interacción (Briz 2000; Pons Bordería 2022). Además, la intervención permite delimitar cada una de las contribuciones del hablante, aspecto esencial para comprender fenómenos como la organización de los turnos de habla.

Sin embargo, aunque las transcripciones realizadas por el grupo Val.Es.Co. se presentan tradicionalmente a través de la unidad intervención, siempre se ha tenido en cuenta el grupo entonativo, dado su interés en la segmentación del discurso (Pons Bordería 2016). Este constituye la unidad física reconocible, definida por una pausa o un tonema marcado (Cabedo 2009 y 2011) que, a su vez, conforma las intervenciones. En la versión 3.0, gracias al uso del programa ELAN y al sistema de análisis descrito en § 3.3., ha sido posible crear *cajas*[5] que separen los grupos entonativos, manteniendo al mismo tiempo su identificación dentro de las intervenciones que los contienen. En el ejemplo siguiente se muestra la línea de transcripción de C (C_PHON) con cuatro grupos entonativos que están recogidos dentro de una única intervención (ver § 3.3.2.).

[5] En el programa ELAN, una caja es un espacio de trabajo digital vinculado a un fragmento temporal de un audio o video. En ellas, no solo es posible escribir la transcripción de lo que se dice en la grabación, sino también analizar la información en diferentes niveles, como se verá en § 3.3.

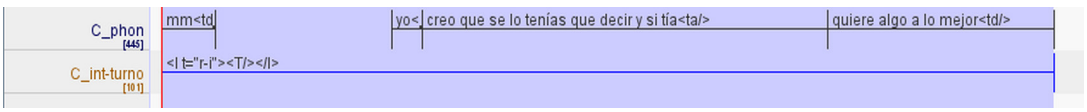



*Imagen 2. Grupos entonativos recogidos dentro de una intervención.*

La línea de transcripción con cajas que separan cada grupo entonativo se tokeniza en una segunda línea (HABLANTE_palabras). Este proceso utiliza el espacio gráfico como frontera y genera un espacio para cada una de las palabras que conforman un grupo entonativo (imagen 3).

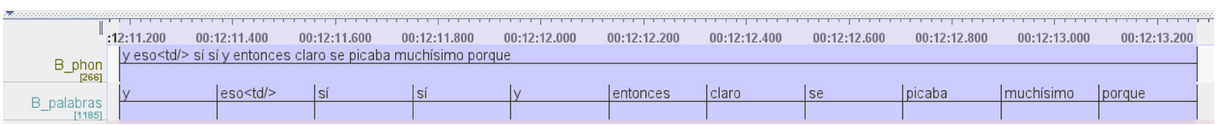



*Imagen 3. Fragmento de transcripción en ELAN con tokenización de palarbas - Conversación 011.PT.S4.*

De este modo, se crea una caja para cada una de las palabras que componen el grupo entonativo, con sus correspondientes identificadores temporales de inicio y fin[6]. Estos identificadores son básicos, puesto que son los que se emplean para asociar las cajas de las distintas líneas de análisis (§ 3.3.2).

La utilidad de este proceso se aprecia en la imagen 4, en la que los límites temporales del subacto sustantivo directivo (SSD) no se establecen desde la línea _PHON, que incluye todo el grupo entonativo, ya que el subacto solo afecta a una parte de este. Sin embargo, los límites del subacto coinciden exactamente con el inicio de la caja de la primera palabra que compone el subacto (*aquí*) y con el fin de la que lo termina («((aquí))»). Por esta razón, la tokenización, al dividir por palabras, permite una división exacta por unidades discursivas.

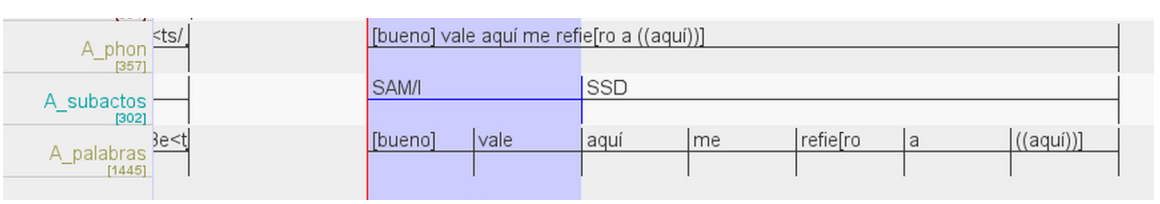



*Imagen 4. Ejemplo de segmentación de subactos a través de ELAN – Conversación 2018. PT.S11*

[6] Cabe reseñar que las cajas creadas automáticamente para las palabras no se corresponden con el fragmento exacto en el que esta ha sido pronunciada, sino que sus marcas temporales son virtuales dentro del grupo entonativo en el que se han pronunciado. En otras palabras, aunque los identificadores temporales de los tokens creados no sean reales, sí que se mantienen dentro del tiempo del grupo entonativo en el que fueron pronunciadas.

Con estos elementos ya se cuenta con las herramientas necesarias para crear un corpus que no solo permita la búsqueda de concordancias o de determinados fenómenos, sino que también pueda mostrar los fragmentos exactos de audio que se corresponden con la transcripción, tal y como implementan los grupos ESLORA o AMERESCO. El corpus Val.Es.Co. 3.0 introduce además la novedad de consultar en línea todas las conversaciones (§ 4.2.) y, algunas de ellas, completamente segmentadas. En § 3.3 se detallará cómo se han adaptado las funcionalidades del programa para alcanzar este objetivo.

## *3.3. Introducción del análisis jerárquico de unidades discursivas*

La segmentación del discurso propuesta por el grupo Val.Es.Co. se basa en un modelo jerárquico y recursivo (§ 2). *Jerárquico* implica que las unidades menores están incluidas dentro de las superiores. Al mismo tiempo, el análisis es recursivo, lo que significa que un nivel puede estar compuesto por unidades del mismo nivel o de nivel superior. Por ejemplo, los subactos directores, que suelen ser el núcleo de los actos, pueden contener en su interior subactos adyacentes, como se observa en (6), donde un subacto adyacente modal integra un subacto sustantivo director :

(6) **(2011.PT.S2)**
**126B76**: #{$_{SSD}$ qué bueno {$_{SAM}$ ¿eh? $_{SAM}$} el español coloquial(ENTRE RISAS)$_{SSD}$}# #(RISAS)#

Este sistema teórico se tiene que adaptar al programa ELAN y esto supone adoptar una serie de decisiones que, vistas en conjunto, han supuesto un desarrollo técnico considerable. En ELAN cada línea está asociada a un hablante, lo que crea un esquema del tipo {A, B, C…}, donde {A, B, C…} son los hablantes. En la versión 3.0 se han asociado los distintos niveles de análisis (subacto, acto, intervención, turno, diálogo y discurso) a cada uno de los hablantes, lo que produce un esquema del tipo {A($l_1$, $l_2$, $l_3$, …$l_n$), B($l_1$, $l_2$, $l_3$, …$l_n$)…}, donde {A, B, C…} son los hablantes y {($l_1$, $l_2$, $l_3$, …$l_n$} son las diferentes líneas que se asocian a este, lo que implica, tanto una línea por nivel de análisis, como las líneas PHON_ y PALABRAS_, descritas en (§ 3.2).

Además de por el fenómeno que contienen, las líneas de la transcripción se pueden dividir en función del tipo de contenido que reproduzcan; así, se distinguen líneas de contenido libre y líneas de contenido cerrado. Las primeras se asocian directamente con la transcripción y son las que aparecen en primer lugar en la hoja de ELAN: X_ PHON, X_ PALABRAS y _OBSERVACIONES. Las segundas no permiten la escritura manual, sino la selección del vocabulario programado

dentro de ellas, ya que ELAN permite crear vocabularios controlados que optimizan el proceso de etiquetado.

Con estos vocabularios se reduce en buena medida el riesgo de errores humanos debidos a la escritura incorrecta de las etiquetas. Las líneas de contenido cerrado se han utilizado para el análisis en unidades discursivas: HABLANTE_subactos, HABLANTE_actos, HABLANTE_ intervención-turno, HABLANTE_diálogos y HABLANTE_discurso.

En resumen, el formato de la hoja de ELAN desarrollada para la transcripción de las conversaciones del corpus 3.0 está formado por los siguientes elementos (Pons Bordería 2022: 138-140 [adaptación]):

| **Tipo de contenido** | **Contenido** | **Línea** | **Tipo** | **Vocabulario** |
|---|---|---|---|---|
| Libre | Transcripción por grupos entonativos del hablante | A/ B/ C... _phon | phon | - |
| | Palabras del hablante | A/ B/ C... palabras | word-tokenización | - |
| | Acciones, gestos o información que influencia o transcurre durante la conversación | Observaciones | obs | - |
| Cerrado | Subacto del hablante | A/ B/ C…_subactos/ subactos_II | subactos | SSD<br>SSS<br>SAM<br>SAI<br>SAT<br>SS/SA |
| | Actos del hablante | A/ B/ C…_actos / actos_II | actos | Acto |
| | Intervenciones del hablante | A/ B/ C…_int-turno | Int_turno | <I t="i"><T/></I><br><I t="i-r"><T/></I><br><I t="r"><T/></I><br><I t="r"><br><I t="ind"> |
| | Diálogos de la conversación | Diálogos_I Y _II | dialogo | <Di t="n"/> |
| | Discursos de la conversación | Discurso | discurso | <DSC/> |

*Tabla 3. Formato de las conversaciones en ELAN para el corpus Val.Es.Co.*

Así, cada línea posee unos rasgos distintivos que permiten su posterior procesamiento informático. A continuación, se describirá el funcionamiento de estos dos tipos de líneas.

### *3.3.1. Líneas de contenido abierto*

Las líneas de contenido libre son tres: HABLANTE_phon, HABLANTE_palabras y Observaciones. Se caracterizan por rellenarse de forma libre.

La línea HABLANTE_phon incluye el texto de la transcripción, como se muestra en la imagen 5. Se trata de la más importante del sistema de transcripción y análisis, dado que es la primera en alinearse con el audio y sirve como base para las demás, al contener los referentes temporales reales asociados a cada fragmento de habla. El resto de las líneas, en cambio, se alinean sobre las cajas de las palabras que corresponden a cada grupo entonativo (§ 3.2.), de manera que cada elemento de las líneas de análisis se vincula al contenido de la transcripción y no al audio de la conversación.

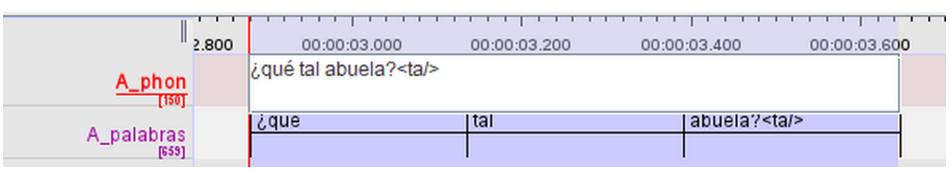


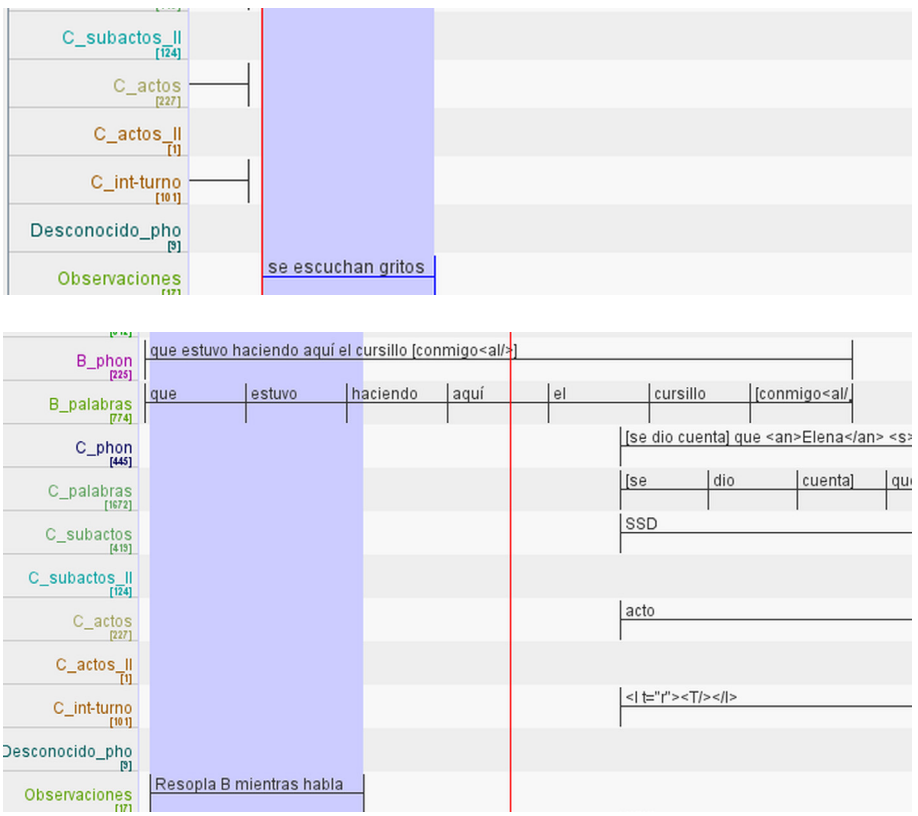

*Imagen 5. Forma de transcripción del audio a través de ELAN*

A pesar de que los fenómenos prosódicos o discursivos recurrentes en la conversación cuentan con un sistema de etiquetas estable, estas se introducen manualmente en este campo, ya que se trata de fenómenos que afectan directamente al contenido de la grabación y, por lo tanto, se codifican en el texto. En otras palabras, no se ha definido un vocabulario para estas etiquetas en ELAN porque se emplean dentro del contenido de la caja de transcripción asociadas a las palabras que afectan.

A partir de HABLANTE_phon se lleva a cabo el proceso de tokenización en la línea HABLANTE_palabras. Gracias a este proceso, el análisis realizado en las líneas de contenido cerrado puede hacerse palabra por palabra (§ 3.3.2). Por ejemplo, para vincular la información de la línea «subactos» al segmento correspondiente, se asocia el punto de inicio de la primera palabra con el punto final, como se ha visto en la imagen 4 del apartado anterior.

Por último, dentro de las líneas de contenido abierto, se encuentra (_Observaciones), destinada a incluir comentarios sobre el contexto o

las acciones de los participantes[7]. Esta línea puede alinearse de forma libre o en relación con _phon, ya que su contenido puede estar relacionado o no con la intervención de un hablante (imágenes 6 y 7, respectivamente). Cabe señalar que, aunque dicha información se utiliza para facilitar una mejor comprensión de la situación, no se emplea directamente en el análisis.




*Imágenes 6 y 7. Transcripción de observaciones a través de ELAN*

### *3.3.2. Líneas de contenido cerrado*

Las líneas de contenido cerrado están vinculadas a las unidades de segmentación definidas por el modelo Val.Es.Co. (2014, Pons Bordería 2022). Cada línea cuenta con un vocabulario predefinido (ver tabla 3), que permite seleccionar directamente la unidad correspondiente sin necesidad de escribirla manualmente. La jerarquía estricta del sistema de unidades se reproduce en la página de ELAN, donde cada línea representa un nivel de información que abarca, desde la unidad estructural mínima (los subactos) hasta la unidad dialógica máxima (los discursos). Este diseño asegura la coherencia del modelo, de modo que

[7] Es la traducción, en la versión 3.0, de las notas a pie de página de las versiones en papel del corpus.

un subacto no puede abarcar dos actos, ni un acto puede distribuirse entre dos intervenciones, como se muestra en la imagen 8.

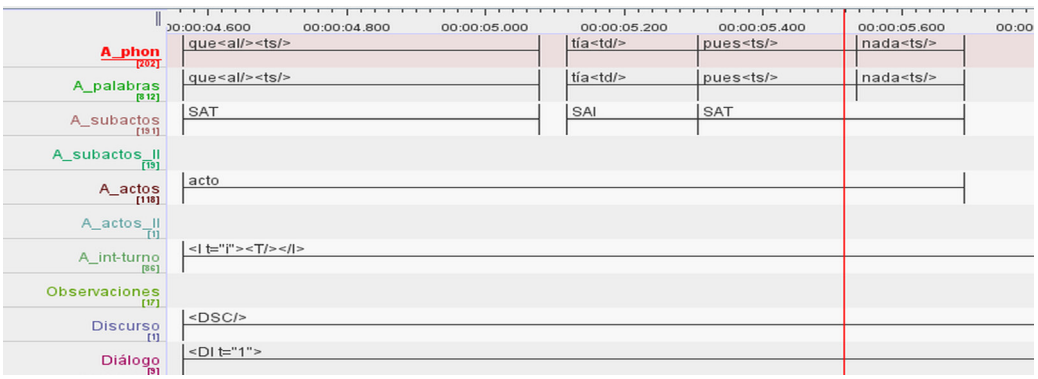



*Imagen 8. Muestra de segmentación de unidades – Conversación 1994.PT.S1.*

Esta estructura, por así decirlo, vertical de las líneas de análisis, se completa con una doble distinción basada en sus características: según el número de participantes a los que afectan las líneas de análisis y según el tipo de vocabulario implementado en ellas.

La primera clasificación se basa en la distinción entre unidades monologales y dialogales del modelo Val.Es.Co. Las unidades monologales (subactos, actos e intervenciones) están asociadas a un único participante, ya que se derivan exclusivamente de las emisiones de un solo hablante. Por lo tanto, cada una de estas unidades se repetirá tantas veces como participantes haya en la conversación, al igual que ocurre con las líneas de transcripción y tokenización, tal como se ilustra en la imagen 9 con los hablantes A y B:

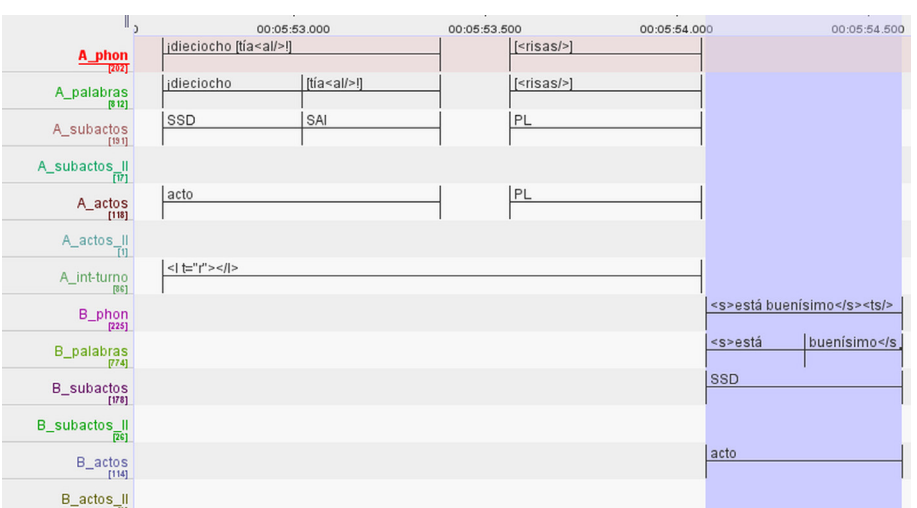



*Imagen 9. Interfaz de ELAN con líneas monologales para A y B. Extraído de 1992.PT.S1.*

Asimismo, los subactos y los actos presentan recursividad, de modo que un acto puede estar dentro de otro acto, o un subacto dentro de otro subacto. Así, estas líneas se pueden duplicar añadiendo el sufijo _II, lo que permite incorporar una línea más de análisis dentro de la categorización general de una estructura, como se observa en la imagen 10:

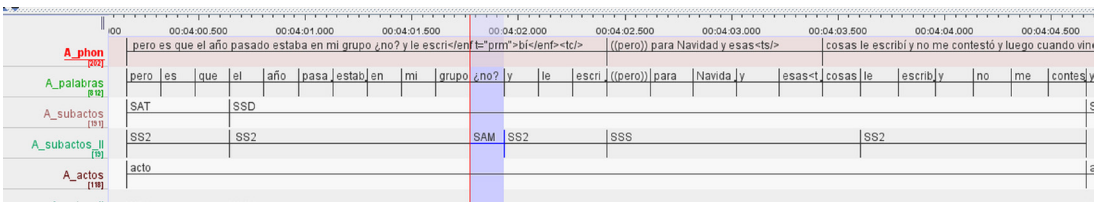
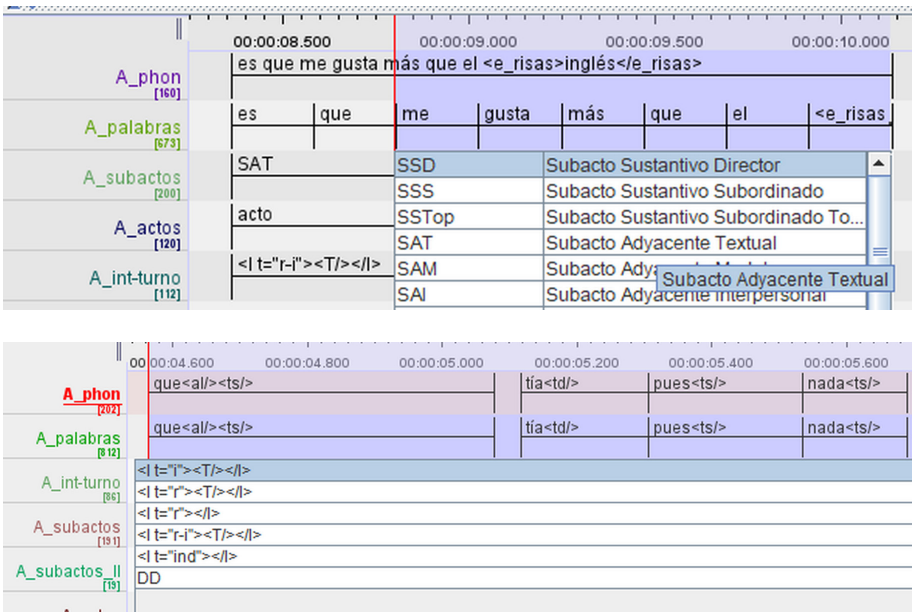



*Imagen 10. Interfaz de ELAN con análisis de segundo nivel para los subactos de A.*

En la línea HABLANTE_SUBACTOS_II de la Figura 10 se puede observar una etiqueta denominada SS2. Esta etiqueta se emplea para satisfacer una exigencia inherente a este sistema de análisis: garantizar que todas las líneas de análisis monologales estén completamente rellenas. En otras palabras, son etiquetas de carácter exclusivamente técnico, sin relación directa con el análisis lingüístico.

La segunda clasificación necesaria para describir las líneas se basa en el tipo de vocabulario que conforman las líneas de contenido cerrado. Según este criterio, se pueden distinguir tres tipos de líneas. En primer lugar, aquellas que incluyen una tipología, como sucede con los subactos y las intervenciones, ya que estas unidades se subdividen en varios tipos. A cada una de ellas le corresponde una etiqueta que permite su identificación en ELAN y se muestra en un desplegable (imágenes 11, 12 y 13, siguiente página).

El segundo grupo lo conforman las unidades sin tipología, como el acto y el discurso. Estas cuentan solo con un tipo de etiqueta que se corresponde con la misma unidad, como se ve en la imagen 13 (siguiente página):




*Imágenes 11 y 12. Interfaz de ELAN con el desplegable del vocabulario de la línea de subactos y de intervenciones.*

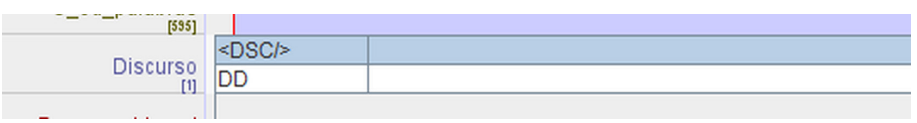



*Imagen 13. Interfaz de ELAN con el desplegable del vocabulario de la línea de discurso*

Por último, aunque las líneas de diálogos no tienen una tipología específica, se numeran desde el análisis debido a un fenómeno que sucede en conversaciones con más de dos participantes: es posible que un diálogo no haya finalizado y que, antes de su conclusión o de manera simultánea, comience otro. Para dar cuenta de esta situación, se introduce una segunda línea de análisis para los diálogos (_Diálogo II). De este modo, puede indicarse el inicio de un nuevo diálogo antes de que haya finalizado el anterior, como se observa en la imagen 14:

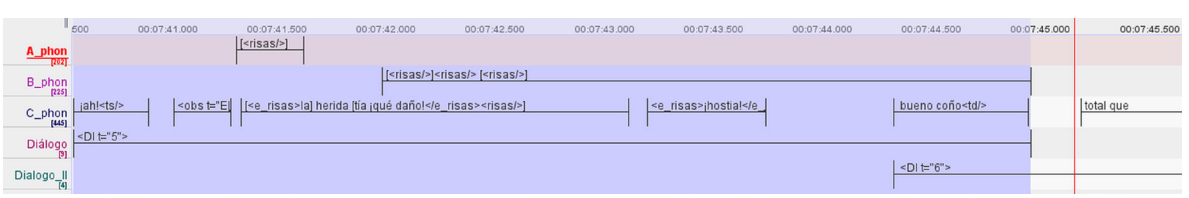



*Imagen 14. Interfaz de ELAN con dos diálogos solapados.*

Esta configuración en cajas y en líneas de vocabulario cerrado resulta muy útil para el investigador, ya que solo debe preocuparse por

crear y ajustar las cajas alineadas de acuerdo con las palabras a las que afecte cada unidad. Una vez creada la caja, el programa no permite la escritura directa como en las líneas de contenido libre, sino que ofrece un desplegable con las opciones disponibles, de entre las que el segmentador solo debe elegir la etiqueta correcta. Esta cuestión es de vital importancia para la construcción del corpus: un error de escritura puede provocar que la base de datos del corpus genere un fallo de lectura, lo que derivaría en un procesamiento de la información incorrecto y, consecuentemente, en una visualización errónea en el corpus.

Por último, toda la estructura de unidades descrita hasta el momento se replica para el estilo directo[8], que se analiza como un nuevo discurso. Para ello, se traslada el contenido de los fragmentos de estilo directo a una nueva línea llamada «HABLANTE_ed» y se duplica el resto de líneas con la marca «ed»:

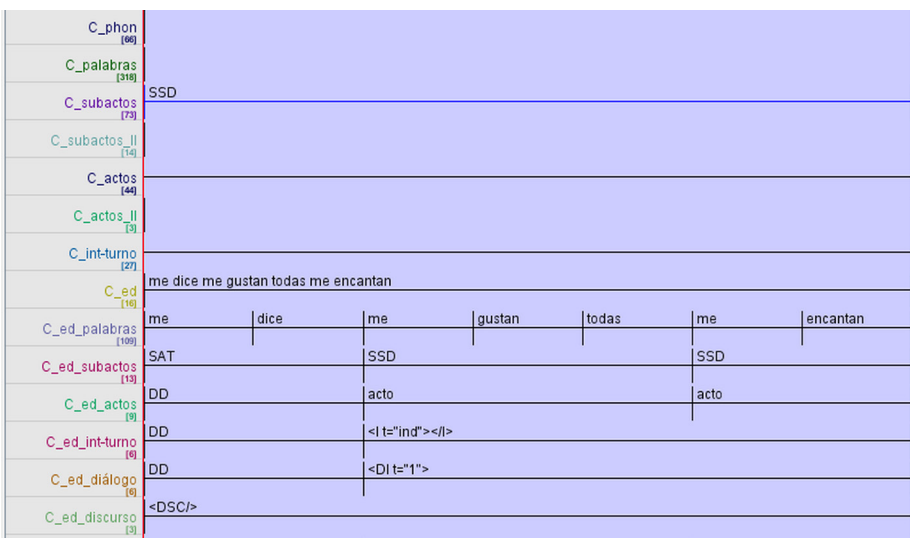



*Imagen 15. Fragmento de transcripción de estilo directo – conversación 2016.PT.(20).S6.*

Como resultado, el análisis del contenido de la transcripción en fragmentos de estilo directo se aborda en un doble nivel. Con respecto al marco que los contiene en el estilo principal, estos fragmentos suelen categorizarse como un SSD, ya que narran algo dicho por otro hablante en un momento previo, es decir, se trata de contenido puramente conceptual. Por otro, debido a que implican un desplazamiento de los ejes deícticos de la conversación (*yo*, *aquí*, *ahora*), se analizan de manera específica como un nuevo discurso, compuesto por diálogos, intervenciones, actos y subactos. Este enfoque permite reflejar la dualidad y complejidad inherente al discurso reportado.

[8] El análisis del estilo directo siguiendo a Benavent (2024) es más complejo y desborda los límites del presente trabajo. Al igual que con el estilo principal, solo se ha explicado su parte técnica.

En conclusión, si se consideran todas las líneas que se han presentado en este apartado, una conversación completamente segmentada incluye 7 líneas por hablante para el estilo principal (2 de contenido libre y 5 de contenido cerrado), 7 para el estilo directo y 4 líneas grupales. Esto da como resultado que la plantilla de cada grabación cuente con un mínimo de 32 líneas, asumiendo que cualquier conversación involucra, al menos, dos hablantes[9].

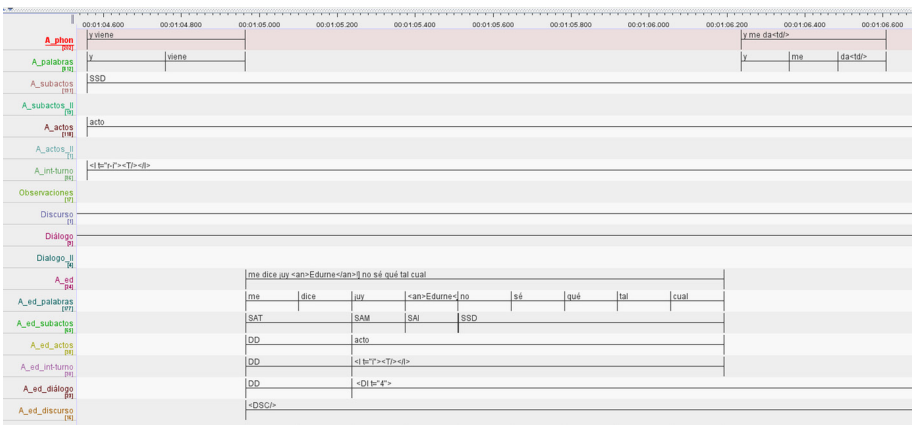



*Imagen 16. Fragmento de transcripción con todas las líneas para un hablante – conversación 1994.PT.S1.*

Esta sección ha explicado los aspectos fundamentales desarrollados en el corpus Val.Es.Co. 3.0 para crear una transcripción segmentada desde el punto de vista técnico. Estos aspectos incluyen, desde la estructura básica del sistema de transcripción, hasta la organización inicial en ELAN. Se trata de una metodología ciertamente laboriosa que ha permitido reflejar en el corpus un análisis tan complejo como la segmentación del discurso oral. En el siguiente apartado, se expondrá cómo todo este trabajo de *backend* se muestra a los usuarios a través de la web del corpus.

## 4. El corpus Val.Es.Co. 3.0 en la web: el *frontend*

La sección 3 ha detallado cómo el *backend* del corpus Val.Es.Co. 3.0 se ha adaptado al objeto de estudio, lo que ha implicado importantes adaptaciones en todos los aspectos de la transcripción. Sin embargo, las etiquetas y marcas de este sistema no pueden mostrarse de manera literal a los usuarios, ya que harían imposible el estudio de las transcripciones y exigirían un conocimiento técnico específico del formato

[9] De acuerdo con el carácter flexible de toda conversación, se podría ampliar el número de líneas si se quisiera incluir, por ejemplo, los gestos o el paralenguaje. El análisis de los gestos en las conversaciones coloquiales (Cabanes 2023) añade siete líneas adicionales:

de análisis en ELAN, lo que reduciría enormemente su utilidad. En este apartado, se explicará cómo este trabajo técnico se ha adaptado para ofrecer una visualización funcional y accesible en la web que permita a los usuarios consultar las transcripciones y explorar los datos segmentados sin enfrentarse a la complejidad del *backend*, optimizando así la experiencia de consulta y análisis del corpus Val.Es.Co. 3.0.

### *4.1. La visualización de la transcripción*

La transcripción constituye el punto de partida del corpus Val.Es.Co. 3.0, ya que sobre ella se construye el resto de las líneas de análisis y segmentación detalladas en el apartado anterior. Sin embargo, el formato técnico empleado para la transcripción no resulta adecuado para su consulta directa: por un lado, la visión horizontal de la conversación no resulta amigable para la lectura y, por otro, el etiquetado de fenómeno vuelve la lectura de las transcripciones casi incomprensible.

Por esta razón, cuando una conversación se sube a la aplicación web, esta sufre tres transformaciones: primero, toda la transcripción, agrupada en grupos entonativos, se reestructura en sus correspondientes intervenciones, gracias a su análisis en el *backend*[10]. Luego, estas intervenciones se ordenan verticalmente de acuerdo con sus índices temporales, un formato mucho más familiar para la lectura. Por último, la aplicación traduce las etiquetas XML al sistema de transcripción jeffersoniano que, como se ha mencionado anteriormente, resulta más legible. Como consecuencia de estas transformaciones, el usuario accede a la conversación de la imagen 17 en el formato de la imagen 18 (siguiente página).

La visualización de la transcripción en el corpus Val.Es.Co. 3.0 responde a la necesidad de adaptar el formato técnico original a un entorno más accesible y familiar para los usuarios. Sin embargo, este corpus, además de permitir el acceso a las conversaciones mediante la búsqueda de concordancias, también hace posible la consulta en línea de las transcripciones completas.

[10] Esta característica obliga a que, como mínimo, todas las conversaciones del corpus estén segmentadas en intervenciones.


*Imagen 17. Interfaz de ELAN para la conversación 1989.PT.56(1).*

| | | |
|---|---|---|
| 213C87 | 1 - Ir | al principio (RISAS) |
| 214A103 | 2 - Ii | en fin mañana a las ocho te levantarás tu madre ↑ te va a levantar a las ocho- a las siete |
| 215C88 | 3 - Iri | mi madre se aburre y se levanta a las siete [(RISAS)] todo el mundo lo sabe(ENTRE RISAS) |
| B58 | 4 - Ir | [(RISAS)] |

*Imagen 18. Muestra de la conversación 1994.PT.3 mediante la visualización de la búsqueda.*

## *4.2. Los modos del corpus: de la búsqueda en conversaciones a su consulta*

En la actualidad, la mayoría de los corpus lingüísticos orales incorporan un buscador que permite localizar palabras o fenómenos en los documentos que contienen, ya que el usuario consulta las transcripciones para encontrar elementos concretos, ya sean palabras, estructuras lingüísticas o fenómenos discursivos (Rojo 2024).

Al igual que otros corpus, Val.Es.Co. 3.0 cuenta con un sistema de búsqueda por concordancias que incorpora cuatro tipos de filtros, organizados según distintos criterios: la consideración lingüística del término de búsqueda, los metadatos de las conversaciones, las características prosódicas y las unidades discursivas. El primer tipo de filtro permite definir cómo debe interpretarse el término introducido, especificando aspectos como forma o lema, la distinción entre mayúsculas y minúsculas, la inclusión o no de acentos, y si el fragmento está precedido o seguido de signos como interrogaciones, exclamaciones o pausas. Además, permite delimitar la búsqueda por la categoría gramatical a la que pertenece el término[11] (imagen 19).

[11] Esta clasificación es posible gracias a la implementación del etiquetador morfológico XIADA (Centro Ramón Piñeiro para la investigación en humanidades), desarrollado por Mario Barcala y el corpus ESLORA.

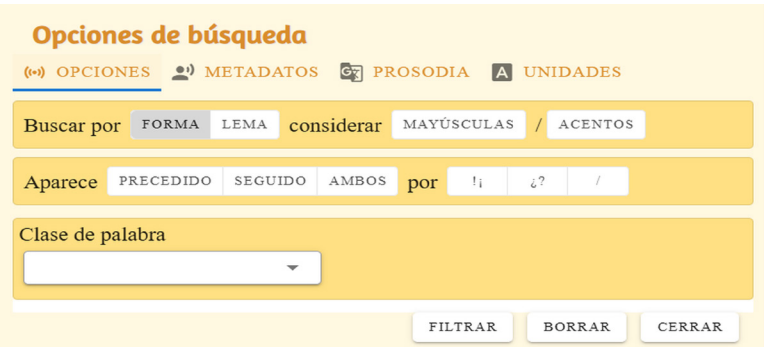
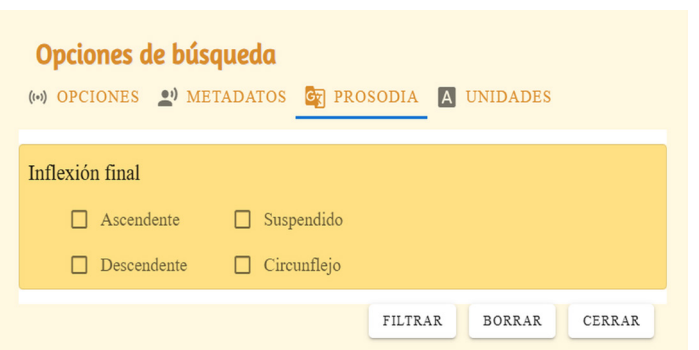



*Imagen 19. Interfaz de filtros de búsqueda - opciones generales.*

El segundo tipo de filtro se refiere a los datos asociados a las conversaciones y permite seleccionar dos categorías de información: los aspectos relacionados con la identificación general de la conversación, como el año de registro o si esta es considerada prototípica, y los datos sociolingüísticos de los hablantes, como su género, nivel educativo o grupo etario (imagen 20)[12]. Por su parte, el filtro Prosodia permite buscar fragmentos de transcripción que finalicen con un determinado tonema (imagen 21).


*Imágenes 20 y 21. Interfaz de búsqueda – filtro de metadatos y de prosodia.*

El último filtro está relacionado con la segmentación de unidades. Gracias al análisis del *backend*, realizado durante el proceso de segmentación en ELAN (§ 3.3), el sistema de búsqueda permite al usuario seleccionar y filtrar por el tipo específico de subacto o intervención que desee consultar. Por ejemplo, es posible buscar únicamente subactos sustantivos directores (SSD) o intervenciones de tipo independiente, como se muestra en la imagen 22 (siguiente página):

[12] Esta es la traducción de la tradicional «ficha técnica», tal y como se presentaba en las versiones previas en papel.

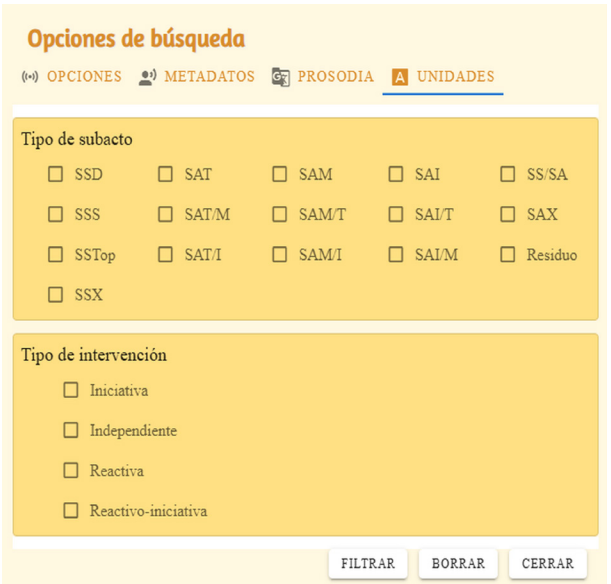



*Imagen 22. Interfaz de búsqueda - filtro de unidades.*

Los resultados obtenidos en el buscador se presentan en una tabla de concordancias que muestra la intervención completa a la que pertenece cada resultado. Cada ejemplo incluye la opción de reproducir el fragmento de audio correspondiente y de ampliar el contexto hasta en cinco intervenciones, previas o posteriores. Esta información es exportable en formato .docx o .xml para facilitar su análisis o uso externo. Además, las conversaciones segmentadas, identificadas con el marcador «S», permiten la exportación adicional con el nivel de análisis que el usuario seleccione.

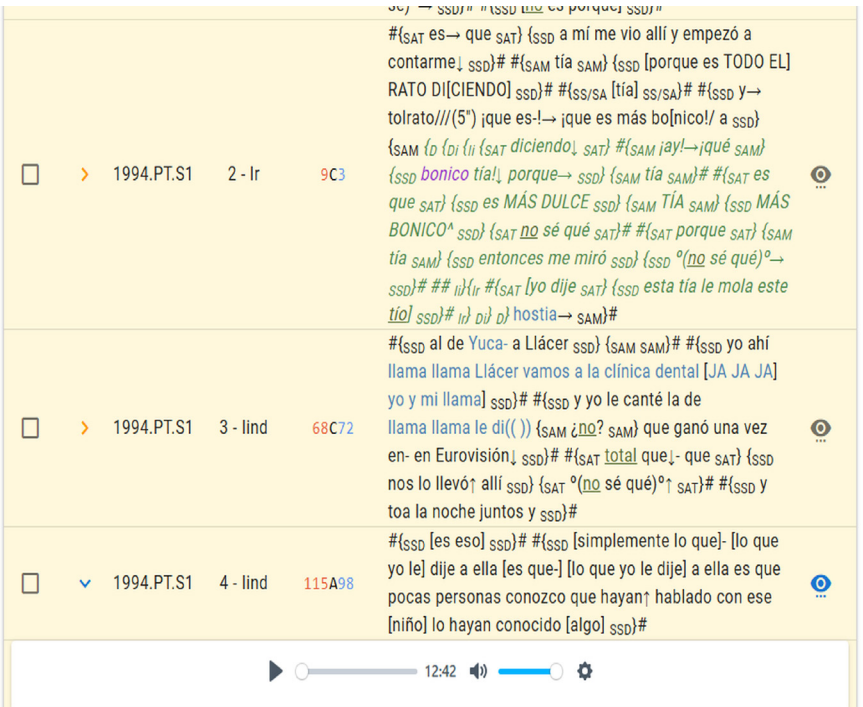




*Imagen 23. Interfaz de búsqueda – resultado de concordancias y ampliación de contexto con análisis de subactos.*

Desde mediados de 2024, se ha incluido en el corpus una modalidad de búsqueda cronológica que, utilizando las funciones del buscador previamente descrito, posibilita realizar búsquedas dobles dentro de un rango específico de años. Esta búsqueda responde al carácter microdiacrónico del corpus Val.Es.Co., que, con casi treinta años de existencia, contiene datos correspondientes a dos generaciones de hablantes. Siguiendo con los objetivos de los proyectos de investigación DIAXX y DIA20, citados al inicio de este artículo, el corpus Val.Es.Co. 3.0 ha transformado su estructura para convertirse en el primer corpus diacrónico oral del dominio hispánico.

Gracias a esta doble búsqueda, los resultados se pueden presentar en paralelo, facilitando de este modo su comparación, tal y como se muestra en la imagen 24:

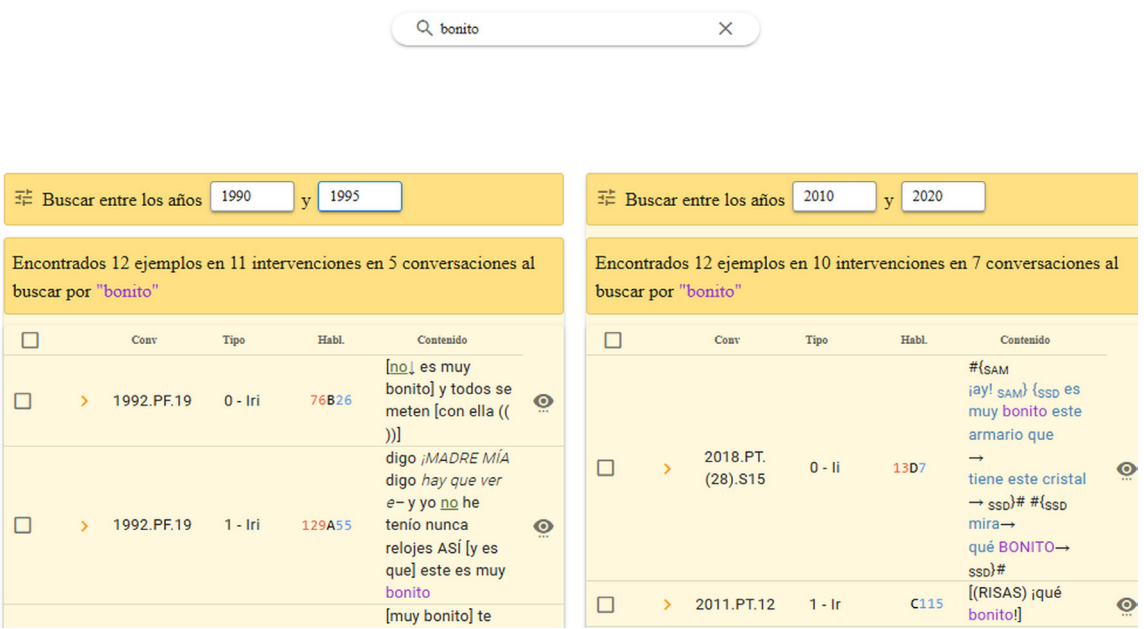



*Imagen 24. Resultados de la búsqueda cronológica*

Además de la búsqueda, el corpus también ofrece la posibilidad de consultar las conversaciones completas en línea mediante el modo REPRODUCCIÓN AVANZADA. Este sistema no solo presenta el contenido de cada transcripción en formato vertical (§ 4.1), sino que permite seguir el desarrollo de la grabación en tiempo real en modo *karaoke*: a medida que el audio avanza, la intervención correspondiente se resalta, como se muestra en la imagen 25 con la intervención 5A3. Asimismo, al igual que en el modo búsqueda, la reproducción avanzada permite visualizar distintos niveles de análisis de segmentación, siempre que dicha información esté disponible en la conversación, como se ilustra en 26.

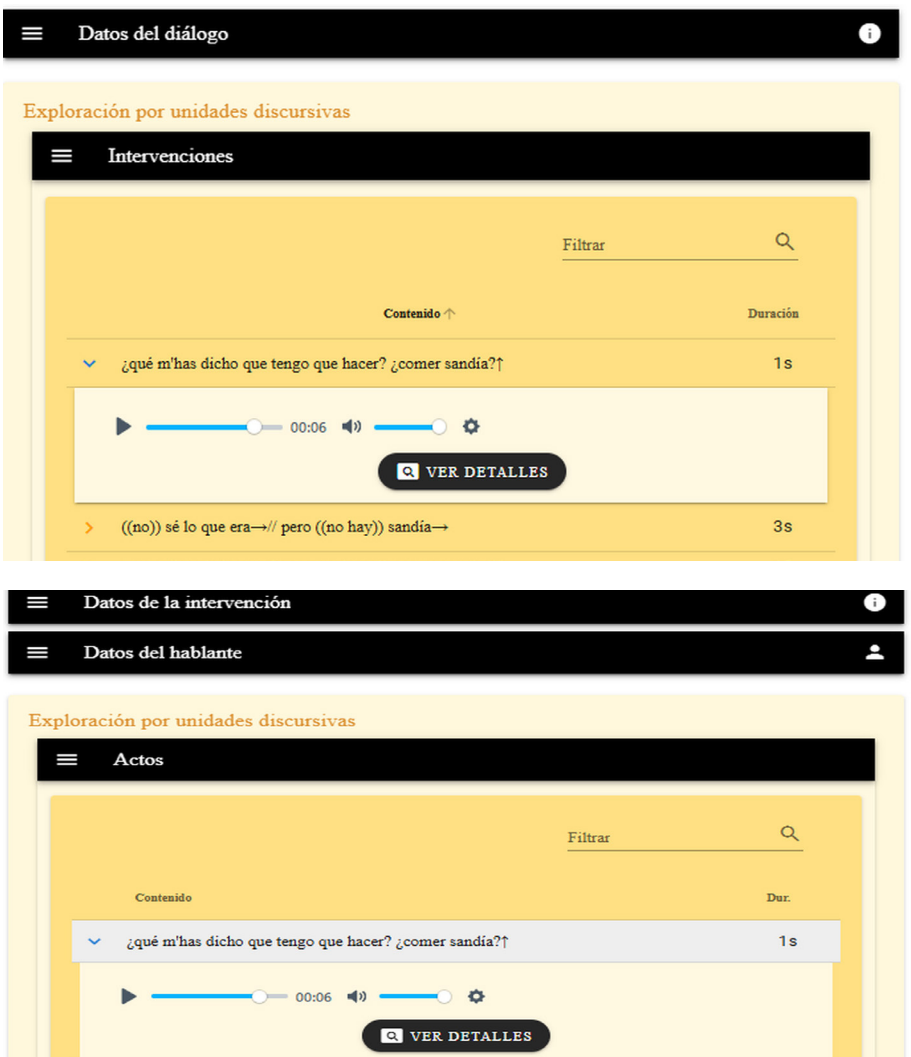

*Imágenes 25 y 26. Interfaz de reproducción avanzada – modo karaoke y transcripción segmentación de subactos*

Esta funcionalidad, junto con la posibilidad de visualizar distintos niveles de análisis, facilita una exploración detallada y dinámica del corpus, de modo que el usuario pueda acceder casi de manera completa a la totalidad del trabajo del grupo de investigación Val.Es.Co.

Por último, cada conversación ofrece una opción de VER DETALLES que no solo proporciona los datos identificativos de la conversación y de los hablantes, sino que también permite acceder a la transcripción a través de las distintas unidades discursivas que se han segmentado (ver § 3.3.2.). Gracias a este trabajo y al proceso de lectura que hace la aplicación web, es posible consultarla desde los discursos, diálogos, intervenciones, actos y subactos que la conforman, manteniendo su correspondencia con el fragmento de audio asociado. Además, cada unidad incluye información sobre su duración, el tiempo de inicio y fin

dentro del audio, y, en el caso de las unidades monologales, los datos del hablante que las ha pronunciado.

Cabe señalar que esta consulta respeta la estructura jerárquica de las unidades descrita en el apartado §3.3.2. Así, al explorar un diálogo, solo es posible acceder a sus intervenciones; al consultar una intervención, únicamente se pueden visualizar sus actos, y así sucesivamente, tal como se ilustra en las imágenes 27 y 28.




*Imágenes 27 y 28. Consulta de la conversación a partir de las unidades diálogos e intervención.*

## 5. Conclusión

En conclusión, el propósito de este artículo ha sido mostrar los objetivos del corpus Val.Es.Co. 3.0, las decisiones técnicas necesarias para su creación y el resultado final, con el fin de ofrecer una guía para el desarrollo de nuevos corpus en línea que quieran ir más allá de la transcripción. Con ello, se ha tratado de destacar la importancia de integrar técnicas de transcripción con análisis cualitativos, como la segmentación del discurso. El resumen del trabajo expuesto en el artículo puede sintetizarse en la siguiente tabla:

<table>
<tr><th>Fenómeno</th><th>Tratamiento de datos</th><th>Codificación informática</th><th>Visualización web del corpus</th></tr>
<tr><td>Contenido</td><td rowspan="4">Transcripción (líneas de contenido libre en ELAN)</td><td colspan="2" rowspan="2">Ortotipográfica</td></tr>
<tr><td>Paralenguaje y sucesos</td></tr>
<tr><td>Prosodia</td><td rowspan="3">Etiquetas XML</td><td rowspan="2">Símbolos (sistema de base jeffersoniana)</td></tr>
<tr><td>Fenómenos discursivos</td></tr>
<tr><td>Análisis de contenido</td><td rowspan="2">Segmentación (líneas con vocabulario cerrado en ELAN)</td><td rowspan="2">Símbolos ortotipográficos</td></tr>
<tr><td>Análisis de interacción</td></tr>
</table>

*Tabla 4. Resumen metodología de trabajo y visualización en la web.*

Creemos que estas decisiones abordan (y, en algunos casos, resuelven) problemas a los que se pueden enfrentar quienes deseen adentrarse en el complejo, duro y laborioso mundo de la creación de corpus, que es un campo de creación colectiva en el que la lingüística española brilla por méritos propios y que la sitúa muy por delante de cualquier otra lengua románica. Este artículo se propone como una invitación para utilizar estos hallazgos en futuras investigaciones que sigan desarrollando nuevos métodos de transcripción, análisis y trabajo para capturar y, en última instancia, comprender ese inasible apoyado en el tiempo que es el lenguaje oral coloquial.

## Bibliografía

Albelda, Marta, y Maria Estellés (dirs.), *Corpus Ameresco*. Disponible en: https://corpusameresco.com. [Fecha de consulta: 8 de septiembre de 2024].

Alcaraz Martínez, Rubén, y Elisabet Vázquez Puig (2016), «TEI: un estándar para codificar textos en el ámbito de las humanidades digitales», *BiD: Textos Universitaris de Biblioteconomia i Documentació*, 37: s.p. DOI: 10.1344/BiD2016.37.24.

Bolaños Cuéllar, Sergio (2015), «La lingüística de corpus: perspectivas para la investigación lingüística contemporánea», *Forma y Función*, 28 (1): 31-54. DOI: 10.15446/fyf.v28n1.51970.

Briz, Antonio (1996), *El español coloquial: situación y uso*, Barcelona, Ariel.

Briz Antonio. (2010), «Lo coloquial y lo formal, el eje de la variedad lingüística», en Castañer, R. M. y Lagüéns, V. (eds.): *«De moneda nunca usada»: Estudios dedicados a José Ma Enguita Utrilla*, Zaragoza, Instituto Fernando El Católico: 125-133.

Briz, Antonio *et al*. (1995), *La conversación coloquial: materiales para su estudio*, Valencia, Universitat de València.

Briz, Antonio y Carcelén, A. (2019): «El futuro iberoamericano del español: la investigación del español oral y en español», en Richard Bueno Hudson (dir.), El español en el mundo: anuario del Instituto Cervantes 2019, Madrid, Bala Perdida/Instituto Cervantes: 189-217.

Brun, Rircardo Eíto (2005). «XML y la gestión de contenidos», *Hipertext. net: Revista Académica sobre Documentación Digital y Comunicación Interactiva*, 3: s.p.

Cabedo Nebot, Adrián (2011). «El reajuste tonal en la delimitación de grupos entonativos», en Antonio Hidalgo Navarro, Yolanda Congosto Martín y Mercedes Quilis Merín (eds.), *El estudio de la prosodia en España en el siglo XXI: Perspectivas y ámbitos*, Valencia, Universitat de València, 209-222.

Cabanes Pérez, Sandra (2023), *Análisis multimodal en la distinción entre intervención y turno: efectos en la segmentación de la conversación desde el modelo Val.Es.Co.*, tesis doctoral, Universitat de València.

Cestero Mancera, Ana M.ª (2014), «Comunicación no verbal y comunicación eficaz», *ELUA*, 28: 125-150.

CORPES = Real Academia Española, *Corpus del Español del Siglo XXI*. Disponible en: https://www.rae.es/corpes. [Fecha de consulta: 8 de septiembre de 2024].

Criado de Val, Manuel (1964), *Fisonomía del español y de las lenguas modernas*, Madrid, Aguilar.

Del Rio Riande, Gimena, y Susanna Allés-Torrent (2023). «Treinta años de TEI en español: usos y comunidad». *Journal of the Text Encoding Initiative*, 16: 1-8.

ESLORA = *Corpus para el estudio del español oral*, versión 2.3. Disponible en: <http://eslora.usc.es>. [Fecha de consulta: octubre de 2024].

García-Miguel, José M. (2022), «Lingüística de corpus», *Estudios de Lingüística del Español*, 45: 11-42.

Garfinkel, Harold (1967), *Studies in ethnomethodology*, Englewood Cliffs, Prentice-Hall.

Jefferson, Gail (2004), «Glossary of transcript symbols with an introduction», en Gene Lerner (ed.), *Conversation analysis: studies from the first generation,* Amsterdam (Phil.), John Benjamin: 13-31. DOI: 10.1075/pbns.125.02jef.

Llamazares, Milka Villayandre (2008), «Lingüística con corpus (I)», *Estudios Humanísticos. Filología*, 30: 329-349. DOI: 10.18002/ehf.v0i30.2847.

Lope Blanch., Juan M. (1971), «El léxico de la zona maya en el marco de la dialectología mexicana», *Nueva Revista de Filología Hispánica*, 20 (1): 1-63. DOI: 10.24201/nrfh.v20i1.1557.

Lope Blanch, Juan M. (1976), «Algunos casos de polimorfismo fonético en México», *Revista de Dialectología y Tradiciones Populares*, 32 (1): 247-262.

Lope Blanch, Juan M. (1986), E*l estudio del español hablado culto: historia de un proyecto*, Ciudad de México, Universidad Nacional Autónoma de México.

Marcos Marín, Francisco (dir.), *Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC)*. Disponible en: https://cvc.cervantes.es/lengua/corlec.htm. [Fecha de consulta: 8 de septiembre de 2024].

Narbona, Antonio (1989), *Sintaxis española: nuevos y viejos enfoques*, Barcelona, Ariel.

O'Keefe, Daniel J. (1979), «Ethnomethodology», *Journal for the Theory of Social Behaviour*, 9 (2): 187-219.

Pons Bordería, Salvador (dir.), *Corpus Val.Es.Co*. Disponible en: http://www.valesco.es. [Fecha de consulta: 8 de septiembre de 2024].

Pons Bordería, Salvador (ed.) (2014): *Discourse segmentation in Romance languages*. Amsterdam (Phil.), John Benjamins.

Pons Bordería, Salvador (2016). «Cómo dividir una conversación en actos y subactos», en Antonio Miguel Bañón *et al.* (eds.), *Oralidad*

*y análisis del discurso: homenaje a Luis Cortés Rodríguez*, Almería, Universidad de Almería, 545-566.

Pons Bordería, Salvador (2022), *Creación y análisis de corpus orales: saberes prácticos y reflexiones teóricas*, Berna, Peter Lang.

Poyatos, Fernando (1994), *La comunicación no verbal*, Madrid, Istmo.

Poyatos, Fernando (2018), *Advances in non-berbal communication*, Amsterdam (Phil.), John Benjamins.

PRESEEA = *Proyecto para el estudio sociolingüístico del español de España y América*. Disponible en: https://preseea.linguas.net. [Fecha de consulta: 8 de septiembre de 2024].

Rojo, Guillermo (2016), «Los corpus textuales del español», *Enciclopedia lingüística hispánica*, 2: 285-296. DOI: 10.4324/9781315792942.

Rojo, Guillermo (2024). «El futuro de los corpus de referencia», *Studia Linguistica Romanica*, 12: 18-33.

Roulet, Eddy, Laurent Fillietaz, y Anne Grobet (2002), «Un modèle et un instrument d'analyse de l'organisation du discours», en Patrick Charaudeau y Dominique Maingueneau (eds.), D*ictionnaire d'analyse du discours*, París, Seuil.

Roulet, Eddy, *et al*. (1981), *L'articulation du discours en français contemporain*, Berna, Peter Lang.

Sacks, Harvey, Emanuel A. Schegloff, y Gail Jefferson (1974), «A simplest systematics for the organization of turn-taking for conversation», *Language*, 50 (4): 696-735.

Sacks, Harvey, y Gail Jefferson (2000), «Convenciones de transcripción», en Teun A. Van Dijk (comp.), *El discurso como estructura y proceso. Estudios del discurso: introducción multidisciplinaria*, Barcelona, Gedisa: 442-444.

Torruella, Joan, y Joaquim Llisterri (1999), «Diseño de corpus textuales y orales», en José Manuel Blecua, Gloria Clavería, Carlos Sánchez y Joan Torruella (eds.): *Filología e informática: nuevas tecnologías en los estudios filológicos*, Barcelona, Milenio/Universidad Autónoma de Barcelona: 45-77.

Val.Es.Co. (2014), «Las unidades del discurso oral: la propuesta Val. Es.Co. de segmentación de la conversación (coloquial)», *Estudios de Lingüística del Español*, 35: 11-71.

Vázquez Rozas, Victoria, *et al*. (2020), «Codificación y anotación del habla en un contexto bilingüe: el corpus ESLORA de español de

Galicia» en Ángel Gallego y Francesc Roca (eds.), *Dialectología Digital del Español*, Santiago de Compostela, Universidade de Santiago de Compostela, 189-224.

Venegas, Rene, Iris Viviana Bosio, y Constanza Ceda-Canales (2022), «Los corpus sincrónicos del español: descripción y potencialidades para la investigación teórica y aplicada de la lengua», *Revista de Lexicografía y Lingüística Aplicada*, 22 (3): 45-67.

Zimmerman, Don H. (1978), «Ethnomethodology», *The American Sociologist*, 13 (1), 6-15.

Marta Garrote Salazar (ORCID: 0000-0002-5566-9073), «El etiquetado gramatical automático en el procesamiento del habla coloquial». *Biblioteca de Babel: Revista de Filología Hispánica*, vol. extra. 2 (2024), pp. 115-128.
**DOI**: 10.15366/bibliotecababel2024.extra2.004 · **ISSNe**: 2695-6349 · Recibido: 26/07/2024 · Aceptado: 28/10/2024

# El etiquetado gramatical automático en el procesamiento del habla coloquial

Marta Garrote Salazar
*Universidad Autónoma de Madrid*
marta.garrote@uam.es

**Resumen**: El etiquetado gramatical (*part-of-speech tagging*) es una técnica esencial del Procesamiento del Lenguaje Natural (PLN) que consiste en asignar etiquetas gramaticales a cada palabra en un determinado texto. Mientras que el etiquetado gramatical se ha estudiado de manera extensa en datos lingüísticos formales y bien estructurados, el etiquetado preciso de textos de habla coloquial plantea dificultades específicas. Este artículo tiene como finalidad explorar las dificultades a las que se enfrentan las técnicas de etiquetado gramatical a la hora de analizar e interpretar el habla coloquial. Se discuten fenómenos como el impacto en la precisión del etiquetado gramatical de los dialectos, la jerga, el contexto cultural y social y las disfluencias discursivas, entre otros. Además, tras una revisión del estado de la cuestión, se identifican soluciones potenciales y se establece una prospectiva de estudio que mejore el desempeño del etiquetado gramatical en contextos de habla coloquial.

**Palabras clave**: etiquetado gramatical automático, procesamiento del lenguaje natural (PLN), habla coloquial, dificultades de etiquetado, prospectiva de investigación.

## Automatic POS tagging in colloquial speech processing

**Abstract**: Part-of-speech (POS) tagging is a fundamental Natural Language Processing (NLP) task that involves assigning grammatical labels to each word in a given text. While POS tagging has been extensively studied in formal, well-structured language data, the accurate tagging of colloquial speech corpora presents unique challenges. This paper aims to explore the difficulties faced when employing POS tagging techniques on colloquial speech texts. We discuss the impact of dialects, slang, cultural and social context, and speech disfluencies on the accuracy of POS tagging. Furthermore, after a review of the state of

the art, we identify potential solutions and future research directions to improve the performance of POS tagging in colloquial speech contexts.

**Keywords**: automatic POS tagging, natural language processing (NLP), colloquial speech, tagging difficulties, research prospective.

## 1. Introducción

El procesamiento del lenguaje natural (PLN) es un área de estudio interdisciplinar que se sitúa entre la inteligencia artificial y la lingüística computacional, y que se centra en el análisis de la interacción entre la máquina y el ser humano. El objetivo principal del PLN es permitir que las máquinas comprendan, interpreten y respondan al lenguaje de manera similar a como lo hace el ser humano (Bolaños 2015). El PLN es esencial para muchas aplicaciones, como los asistentes virtuales (como Siri o Alexa), los traductores automáticos (como Google Translate), los sistemas de recomendación (como los de Amazon o Youtube), los chatbots (como ChatGPT), la extracción de información y muchos otros sistemas que interactúan a través del lenguaje humano. Para lograr esa interacción, el PLN abarca una gran variedad de técnicas, entre las cuales se encuentra el etiquetado gramatical o *part-of-speech* (POS) *tagging*.

El etiquetado gramatical es una técnica del PLN que consiste en asignar categorías gramaticales (como nombre, verbo, adjetivo, preposición, etc.) a las palabras de un texto con la finalidad de analizar y comprender la estructura gramatical de las oraciones mediante la clasificación de sus componentes en categorías sintácticas y, así, interpretar su significado correcto, resolviendo la ambigüedad en tareas de análisis sintáctico estructural (Chiche y Yitagesu 2022).

El etiquetado gramatical juega un papel esencial en el PLN, como en la traducción automática, en la minería de datos o en la extracción de información. En la mayoría de los estudios y desarrollos, los modelos de etiquetado gramatical han sido entrenados y evaluados a partir de corpus escritos de lengua formal (Taulé et al. 2015). No obstante, el creciente interés por analizar y procesar lenguaje coloquial genera nuevos retos debido a su naturaleza informal. Si el lenguaje informal supone mayor dificultad que el formal en la precisión del análisis del etiquetador, esta dificultad incrementa cuando se trata de analizar el habla coloquial, es decir, oral, ya que, además de las características propias del lenguaje informal, la lengua oral incluye elementos comunicativos, como la prosodia o la discontinuidad de la cadena hablada, que

hacen que la precisión de los etiquetadores automáticos disminuya. Profundizaremos en estos fenómenos más adelante.

Este estudio plantea las dificultades que, a pesar de los enormes avances en el PLN, aún existen en el etiquetado gramatical del habla coloquial y explora estrategias potenciales para solventar estas dificultades de manera eficiente. Para ello, se explicará en profundidad qué es el etiquetado gramatical, cuáles son los principales métodos existentes, qué carencias presentan las distintas técnicas a la hora de analizar e interpretar el habla coloquial y qué soluciones se deben adoptar en futuros trabajos.

## 2. ¿Qué es el etiquetado gramatical?

Como se ha mencionado, el etiquetado gramatical es una técnica o tarea muy habitual dentro del PLN que consiste en asignar etiquetas gramaticales a cada palabra en un texto (Chiche y Yitagesu 2022). Un ejemplo es el etiquetario utilizado en el corpus CHIEDE (Garrote 2010: 86) que se reproduce en la Tabla 1:

| **Categoría** | **Etiqueta** | **Ejemplo** |
|---|---|---|
| Sustantivo | N | mesa |
| Nombre propio | NPR | Natalia |
| Adjetivo | ADJ | rojo |
| Artículo | ART | el, la, lo |
| Posesivo | POSS | mi, tu, su |
| Demostrativo | DEM | este, ese |
| Cuantificador | Q | uno, primero, muchos |
| Pronombre | P | yo, tú, lo_que |
| Relativo | REL | que |
| Verbo | V | comer |
| Auxiliar | AUX | **he** comido |
| Preposición | PREP | ante, bajo, con |
| Adverbio | ADV | aquí, así, allí |
| Conjunción | C | y, pero, ni |
| Marcador discursivo | MD | oye, o_sea, es_decir |
| Interjección | INTJ | madre_mía, uy |

*Tabla 1. Etiquetario del corpus CHIEDE (Garrote 2010: 86).*

Mediante el etiquetado gramatical, la máquina no solo puede interpretar la oración de manera adecuada, atendiendo a las relaciones sintácticas de los elementos que la conforman, sino que también puede evitar la ambigüedad semántica, entre otras funciones. Por ejemplo, si el etiquetado es preciso –siguiendo los ejemplos de la Tabla 1–, la máquina interpretará que *madre_mía* es una interjección, no una referencia a la madre del hablante.

## *2.1. Usos del etiquetado gramatical en el PLN*

La información que proporciona el etiquetado gramatical (Ghosh y Mishra 2020) es crucial para tareas del PLN como:

— **Análisis sintáctico** (*parser*): el etiquetado gramatical es un paso preliminar a tareas más complejas como el análisis sintáctico (Castillo Velásquez *et al.* 2020), mediante el cual se establecen las relaciones entre palabras y su papel sintáctico. El análisis sintáctico automático es una herramienta fundamental para analizar y procesar la estructura de un texto, de acuerdo con las reglas definidas, para asegurar que se ajusta a la gramática esperada y permite una interpretación posterior adecuada.

— **Reconocimiento de entidades nombradas** (*Named-entity recognition*): la identificación de la categoría gramatical de cada palabra ayuda en el reconocimiento de entidades, como nombres propios de personas, organizaciones, localizaciones, etc. (Landolsi *et al.* 2024). El reconocimiento de entidades nombradas es una herramienta poderosa en el PLN que permite la extracción automática y estructurada de información clave a partir de textos. No obstante, uno de los desafíos a los que se enfrentan los sistemas de reconocimiento de entidades nombradas es la ambigüedad, ya que una misma palabra (o grupo de palabras) puede referirse a diferentes entidades según el contexto. Para ello, un etiquetador gramatical puede ser de ayuda para identificar la categoría gramatical y facilitar la correcta interpretación. Por ejemplo, la palabra *chile* puede etiquetarse como N o como NPR (ver Tabla 1) y, dependiendo de la categoría gramatical que se le asigne, el significado del enunciado oracional variará.

— **Traducción automática**: entender la estructura gramatical de una oración es esencial para acometer una traducción rigurosa (Sánchez-Cartagena 2024). Este proceso de traducción automática implica varios niveles de análisis lingüístico, desde el nivel léxico hasta el sintáctico y semántico, pasando por el etiquetado gramatical.

— **Recuperación de información**: el etiquetado gramatical puede ser de utilidad en los sistemas de recuperación de información para mejorar los motores de búsqueda al considerar las categorías gramaticales y los papeles sintácticos de las palabras (Cherradi y Haddadi 2024).

— **Reconocimiento y síntesis de voz** (*Speech-to-Text and Text-to-Speech systems*): conocer la categoría gramatical de cada palabra ayuda a interpretar y generar un discurso más natural y adecuado al contexto (Ying *et al*. 2024).

Respecto a los métodos, existen distintos enfoques para abordar el etiquetado gramatical de un texto. En la siguiente sección haremos un breve repaso de la información más relevante.

## *2.2. Enfoques y métodos*

El etiquetado gramatical nace de la necesidad de desambiguar palabras cuya categoría gramatical varía de un contexto a otro. El etiquetado manual es una tarea ardua; de ahí la búsqueda de la automatización. Para ello, se han desarrollado distintos enfoques. Los más generalizados son los enfoques basados en reglas (*rule-based approach*), los enfoques probabilísticos (*probabilistic approach*) y los basados en transformaciones (*transformational-based approach*) (Martínez 2012).

Las técnicas basadas en reglas asignan categorías gramaticales a las unidades léxicas a partir de reglas lingüísticas creadas manualmente; por ejemplo, se establece que ante una palabra ambigua como *bajo*, esta será etiquetada como nombre, y no como preposición, si va precedida de un determinante. Uno de los desarrollos más conocido dentro de los modelos basados en reglas es el *Brill tagger* (Brill 1993).

Los enfoques probabilísticos parten de la frecuencia de las etiquetas presentes en un corpus ya etiquetado manualmente y asignan la etiqueta basándose en una mayor probabilidad (modelos estocásticos) de la categoría de una determinada palabra en un determinado contexto. El modelo oculto de Markov (*Hidden Markov Model*) es uno de los más comunes entre los estocásticos.

El tercer enfoque, el basado en transformaciones, aplica un sistema híbrido, combinando los dos anteriores y generando, así, reglas a partir de un corpus[1].

Recientemente, se están utilizando técnicas de la inteligencia artificial (IA) para realizar el etiquetado gramatical de manera automática. Dos ejemplos son el uso del aprendizaje automático (*machine learning*) y del aprendizaje profundo (*deep learning*). Chiche y Yitagesu (2022) realizan un análisis exhaustivo del estado de la cuestión, comparando distintos métodos basados en la IA para etiquetar automáticamente las partes del discurso de un texto y concluyen que, hasta el momento, los más eficientes son los basados en el aprendizaje profundo, seguidos de aquellos basados en el aprendizaje automático y, por último, los métodos híbridos.

De acuerdo con la mayoría de los estudios (Bonilla 2024), el porcentaje de precisión de los etiquetadores gramaticales automáticos se sitúa

[1] Véase Chiche y Yitagesu (2022), Kumawat y Jain (2015) y Martínez (2012) para una mayor información sobre los distintos métodos de etiquetado gramatical.

alrededor del 97 %. No obstante, esta cifra hace referencia al lenguaje formal. Este porcentaje disminuye considerablemente (dependiendo de la lengua y del dominio lingüístico) cuando nos referimos al lenguaje oral. Es decir, que incluso en textos formales y editados el porcentaje de acierto no alcanza el 100 %. Si bien dichos porcentajes de precisión son un éxito para el PLN, el reto consiste en conseguir una precisión del 100 % para cualquier tipo de registro lingüístico. Las peculiaridades idiosincrásicas de la lengua informal, especialmente del habla coloquial, hacen que su procesamiento automático aún esté lejos de ser una realidad. En la siguiente sección se abordará la definición de habla coloquial, sus características y, derivadas de estas, la dificultad de su procesamiento automático.

## 3. Dificultades para etiquetar gramaticalmente el habla coloquial

### *3.1. Definición de habla coloquial*

El evento de habla coloquial se caracteriza por su espontaneidad y por ciertos elementos intrínsecos a ella, derivados de la presencialidad y la contextualidad (Garrote 2010: 66):

— Interacciones cara a cara multimodales: importancia del contexto.

— Referencia intersubjetiva a un espacio deíctico.

— Programación simultánea de la producción.

— Comportamiento lingüístico impredecible sujeto al contexto.

— Discontinuidad de la cadena hablada.

— Lenguaje corporal.

— Aspectos suprasegmentales.

— Aspectos pragmáticos.

— Características sociales y contextuales que influyen en el evento del habla.

El contexto, el espacio, la espontaneidad, la impredecibilidad, la ruptura del flujo discursivo (disfluencias), los elementos paralingüísticos o el lenguaje no verbal son elementos comunicativos que el ser humano, en circunstancias normales, procesa con naturalidad. No obstante, todos estos fenómenos comunicativos son difíciles de formalizar en un lenguaje que pueda procesar la máquina (Rojo 2021).

Según Briz (2016: 463), el habla coloquial es «[u]n registro o uso lingüístico […] empleado en situaciones de inmediatez comunicativa, a la vez que favorecido por estas». Añade el autor que dichas situaciones se definen por las relaciones entre los hablantes, de igualdad social, proximidad, cotidianidad temática y espacial, el fin interpersonal, la falta de planificación y el tono informal, lo que hace que el discurso se caracterice por rasgos dialectales y sociolectales (propios de las características sociolingüísticas de los hablantes). Además, Briz (2016: 464) matiza que «a veces se escribe como si se hablara». Esto se aprecia especialmente en interacciones a través de redes sociales. Cuando hablamos con un amigo por Whatsapp, escribimos como si estuviéramos hablando, intentando reproducir, de hecho, fenómenos característicos de la lengua oral, como rasgos prosódicos (utilización de mayúsculas como si estuviéramos gritando), pérdida de sonidos (*pa mí*), léxico específico (sociolecto), marcadores discursivos prototípicamente orales (*bueno*, *mira*...) o una sintaxis que sigue un orden menos neutro que el estándar sujeto-verbo-objeto, etc. Si tomamos en consideración alguno de los ejemplos citados, el lexicón (diccionario) de un sistema de etiquetado gramatical no recogerá la palabra *pa* como apócope de la preposición *para*; y, probablemente, tampoco incluya neologismos y palabras propias de un sociolecto; además, el sistema de etiquetado podría categorizar de manera errónea la palabra *bueno* como adjetivo cuando esté funcionando como marcador discursivo.

Todas estas características del habla coloquial hacen que la descodificación e interpretación del discurso por parte de la máquina sean más complejas y costosas. El procesamiento del habla coloquial añade retos al PLN que detallaremos a continuación.

### *3.2. ¿Por qué es difícil etiquetar gramaticalmente el habla coloquial?*

El etiquetado gramatical automático de textos de lenguaje coloquial (concretamente, del habla coloquial, aunque, como se ha mencionado, los textos escritos en ciertos contextos presentan características similares a las de la oralidad) plantea varios desafíos debido a la particularidad y a la variabilidad de este registro lingüístico. A pesar de que, en primer lugar, detallaremos aquellos propios de la oralidad, también se mencionarán rasgos propios de la escritura coloquial que afectan a la eficacia y al rendimiento del proceso de etiquetado gramatical automático. Respecto a la oralidad, los principales retos son la variedad léxica, la ambigüedad, las disfluencias, el contexto social y cultural o la innovación y creatividad lingüísticas (entre otros).

La variedad léxica, debido a cuestiones como el uso de dialectos, de jergas o de préstamos lingüísticos, es difícil de procesar mediante

un programa de etiquetado gramatical automático porque determinadas unidades léxicas no aparecen en los lexicones, ni son habituales en los corpus etiquetados utilizados como modelo de entrenamiento. La imposibilidad de que un corpus contenga todos los elementos de una lengua impide que se pueda recopilar un conjunto de datos completo con el que entrenar a la máquina (Rojo 2021). Así, se alcanza una mayor precisión cuando se trabaja con corpus de documentos de dominios lingüísticos específicos (Neunerdt *et al.* 2013) para los que se ha entrenado a la máquina.

La ambigüedad (que puede ser léxica o contextual), junto con la polisemia, plantea un gran reto a los etiquetadores gramaticales. De hecho, la desambiguación es la principal función de un etiquetador. En la lengua coloquial es más habitual que las palabras puedan tener diferentes significados según el contexto; es más, la falta de contexto claro en oraciones cortas y fragmentadas puede hacer difícil la desambiguación de términos. La ambigüedad y la presencia de palabras desconocidas reducen el índice de acierto o precisión de los etiquetadores gramaticales automáticos (Bonilla 2024).

Las disfluencias, entendidas como fenómenos orales que interrumpen el flujo del habla (Crible *et al*. 2019), como las pausas, las autointerrupciones, las repeticiones, los falsos inicios, los alargamientos vocálicos o las pausas sonoras (o *fillers*, como *eh*), complican el trabajo de etiquetado automático. Las disfluencias son un fenómeno característico del habla, especialmente de la coloquial o informal (aunque también de la formal). Tradicionalmente, se han considerado un obstáculo para la comunicación; sin embargo, en la actualidad, los expertos consideran que contribuyen al proceso de comunicación, facilitando la comprensión del discurso (Barr y Seyfeddinipur 2009). Las disfluencias ayudan a desambiguar el significado, evitando interpretaciones erróneas, permiten al interlocutor predecir el tema y apreciar las intenciones del hablante (si existe duda, si se pretende matizar un significado, etc.). Es decir, tienen una función pragmática que el cerebro humano sabe descodificar e interpretar. No obstante, estos fenómenos son un inconveniente para un etiquetador gramatical automático y para el PLN en general.

Los contextos social y cultural influyen enormemente en la generación del habla coloquial. Las referencias a eventos, personajes o chistes específicos de la cultura o del grupo social pueden resultar difíciles de interpretar sin el contexto adecuado. Si a esto añadimos las diferencias regionales (dialectos) y/o sociales (sociolectos) en el uso del lenguaje, estos usos pueden impedir que los modelos generalicen adecuadamente todas las formas de habla (Farasyn *et al.* 2022; Jamatia *et al.* 2015; Jørgensen *et al.* 2015; Rozovskaya *et al.* 2006).

La innovación y creatividad lingüísticas, tan propias de una lengua viva, implican la creación constante de nuevas palabras y expresiones en el lenguaje coloquial que pueden dejar obsoletos rápidamente a los modelos entrenados (Taulé *et al.* 2015). Los modelos necesitan actualizarse constantemente para mantenerse al día con los cambios en el uso del lenguaje.

Por último, a pesar de no ser fenómenos propios de la oralidad, la escritura coloquial también presenta retos para el etiquetado gramatical automático que merecen ser mencionados. Así, en la escritura también se producen alteraciones gráficas (dejando de lado los errores involuntarios) que obedecen a la innovación y la creatividad, como el uso intencionado de errores ortográficos y tipográficos con efectos estilísticos o humorísticos, el uso de formas abreviadas y siglas o el uso de símbolos y emoticonos, que pueden alterar el significado del texto y complicar el análisis gramatical (Neunerdt *et al.* 2013).

## 4. Propuestas y prospectiva de investigación

Abordar estos desafíos implica desarrollar modelos de PLN que sean robustos y adaptables. El uso de técnicas avanzadas de aprendizaje profundo, entrenadas con grandes y variados conjuntos de datos que incluyan ejemplos de habla coloquial, así como la implementación de algoritmos que puedan manejar la corrección ortográfica y la contextualización, son enfoques necesarios para mejorar la precisión del etiquetado gramatical en textos coloquiales en cualquier lengua. Además, la combinación de enfoques basados en reglas con modelos estadísticos y de aprendizaje automático puede ayudar a mejorar la precisión del etiquetado gramatical en textos coloquiales.

La investigación sobre el etiquetado gramatical ha identificado varios retos, incluyendo la necesidad de reducir la ambigüedad espuria (percepción incorrecta de que una oración tiene múltiples interpretaciones cuando, en realidad, solo hay una interpretación lógica o contextual correcta) en la gramática de dependencias categoriales (Alfared 2012), el impacto de las palabras desconocidas y la ambigüedad en la precisión (Anbananthen 2017) y el problema de etiquetar palabras desconocidas de manera eficiente (Gupta 2014). Para abordar estos desafíos, se han propuesto varios enfoques. Alfared (2012) sugiere usar un etiquetador gramatical automático destinado a mejorar la tasa de ambigüedad espuria, mientras que Anbananthen (2017) compara metodologías estocásticas y basadas en reglas para incrementar la precisión del etiquetado automático. Gupta (2014) propone un modelo para etiquetar palabras desconocidas y Tsai (2003) introduce un modelo de reglas de contexto para lograr una alta precisión y reducir la

revisión manual. Estos estudios, en conjunto, resaltan la importancia de abordar la ambigüedad, las palabras desconocidas y la escasez de datos para entrenar a los etiquetadores gramaticales.

A grandes rasgos, se podría hablar de cuatro prospectivas de investigación principales: modelos de etiquetado específicos para el habla coloquial, incrustaciones contextuales y redes neuronales, aprendizaje activo y colaboración abierta y métricas de evaluación y puntos de referencia.

En primer lugar, es necesario contar con modelos de etiquetado específicos para el habla coloquial. Desarrollar modelos de etiquetado gramatical automático especializados que se entrenen con corpus de habla coloquial puede ayudar a abordar muchos de los desafíos discutidos. La incorporación de características lingüísticas específicas del habla coloquial y el aprovechamiento de técnicas de adaptación del dominio lingüístico pueden mejorar la precisión de dichos modelos. Evidentemente, este es un trabajo arduo, ya que no solo supone recopilar y anotar corpus de habla coloquial, sino que, además, estos modelos deben tener en cuenta los distintos y variados dominios lingüísticos y deben ser específicos para cada lengua (Rozovskaya *et al.* 2006).

En segundo lugar, el uso de incrustaciones contextuales (*contextual embeddings*) y de redes neuronales, fruto del reciente avance en métodos de aprendizaje profundo, se ha mostrado prometedor en la solución de varios desafíos del PLN. Las incrustaciones contextuales consisten en otorgar significados dinámicos a las palabras según el contexto en el que aparecen, en contraposición a la tradicional asignación de significados estáticos que no tienen en cuenta el contexto (Kanade *et al.* 2020). En cuanto a las redes neuronales, son útiles para tareas de etiquetado de secuencias, predicción de la siguiente palabra, reconocimiento de patrones, etc. Aplicar estas técnicas al etiquetado gramatical automático de conjuntos de datos de habla coloquial puede capturar matices de significado contextuales y mejorar el rendimiento general de los etiquetadores.

En tercer lugar, aprovechar los métodos de aprendizaje activo y la colaboración abierta (*crowdsourcing*) puede ayudar a construir corpus anotados de habla coloquial a gran escala. Utilizar anotadores humanos especializados en lenguaje coloquial puede proporcionar datos valiosos para el entrenamiento y la evaluación de modelos de etiquetado gramatical automático en este registro lingüístico.

Por último, desarrollar métricas de evaluación y crear puntos de referencia (*benchmarks*) estandarizados específicamente adaptados al etiquetado gramatical del habla coloquial es vital para medir su rendimiento (Tintinago *et al.* 2018). Las métricas de evaluación comparan

valores como la precisión y la exhaustividad en el resultado de los etiquetadores. Los puntos de referencia son conjuntos de datos y estándares utilizados para evaluar y comparar diferentes modelos de etiquetado gramatical. Estos dos procedimientos pueden usarse para valorar objetivamente diferentes modelos de etiquetado e identificar las áreas de mejora.

## 5. Conclusiones

El etiquetado gramatical automático de corpus de habla coloquial presenta numerosos desafíos debido, entre otros fenómenos propios del registro, a la variedad léxica, a la ambigüedad, a las disfluencias, a los contextos social y cultural o a la innovación y la creatividad lingüísticas. Abordar estos desafíos requiere más investigación y desarrollo en la creación de modelos especializados en los distintos dominios lingüísticos, el aprovechamiento de la información contextual, la adopción de técnicas de aprendizaje activo y la comparación y evaluación de etiquetadores. A estos retos, se añade la dificultad de crear etiquetadores estándar para una lengua, ya que es imposible contar con un corpus que recoja por completo todos los dominios lingüísticos para entrenar a la máquina. A medida que aumente el interés en procesar datos de habla informal, los avances en el etiquetado gramatical para el lenguaje coloquial permitirán un PLN más preciso y aplicaciones en escenarios del mundo real.

## Bibliografía

Alfared, Ramadan, y Denis Béchet (2012). «POS taggers and dependency parsing», *International Journal of Computational Linguistics and Applications*, 3 (2): 107-122.

Anbananthen, Kalaiarasi S. M., Jaya K. Krishnan, M. Shohel Sayeed y Praviny Muniapan (2017), «Comparison of stochastic and rule-based POS tagging on Malay online text», *American Journal of Applied Sciences*: 14 (9), 843-851.

Barr, Dale J., y Mandana Seyfeddinipur (2009), «The role of fillers in listener attributions for speaker disfluency», *Language and Cognitive Processes*, 25 (4): 441–455. DOI: 10.1080/01690960903047122.

Bolaños, Sergio (2015), «La lingüística de corpus: perspectivas para la investigación lingüística contemporánea», *Forma y Función*, 28 (1): 31-54. DOI: 10.15446/fyf.v28n1.51970.

Bonilla, Johnatan. E. (2024), «Spoken Spanish POS tagging: gold standard dataset», *Language Resources and Evaluation*: 1-30. DOI: 10.1007/s10579-024-09751-x.

Brill, Eric D. (1993), *A corpus-based approach to language learning*, Philadelphia, University of Pennsylvania.

Briz, Antonio (2016), «Español coloquial», en Javier Gutiérrez-Rexach (ed.) *Enciclopedia de lingüística hispánica*, vol. 2, Londres/Nueva York, Routledge: 463-476.

Castillo Velásquez, Francisco A., José Luis Martínez Godoy, María del Consuelo P. Torres Falcón, Jonny P. Zavala De Paz, Adela Becerra Chávez, y José A. Rizzo Sierra (2020), «Atribución de autoría de mensajes de Twitter a través del análisis sintáctico automático», *Research in Computer Science*, 149 (11): 91-101.

Cherradi, Mohamed, y Anass Haddadi (2024), «Exploration of scientific documents through unsupervised learning-based segmentation techniques», *Seminars in Medical Writing and Education*, 3: 1-9. DOI: 10.56294/mw202468.

Chiche, Alebachew, y Betselot Yitagesu (2022), «Part of speech tagging: a systematic review of deep learning and machine learning approaches», *Journal of Big Data*, 9: 1-25. DOI: 10.1186/s40537-022-00561-y.

Crible, Ludivine, Amandine Dumont, Lulia Grosman y Ingrid Notarrigo (2019), «(Dis)fluency across spoken and signed languages: spplication of an interoperable annotation scheme», en Liesbeth Degand, Gaëtanelle Gilquin, Laurence Meurant y Catherine Simon (eds.) *Fluency and disfluency across languages and language varieties*, Lovaina, Presses universitaires de Louvain: 17-40.

Farasyn, Melisa, Anne-Sophie Ghyselen, Jacques Van Keymeulen y Anne Breitbarth (2022), «Challenges in tagging and parsing spoken dialects of Dutch», *Journal of Historical Syntax*, 6 (4-11): 1-36.

Garrote, Marta (2010), *Los corpus de habla infantil: metodología y análisis*, Madrid, UAM Ediciones.

Ghosh, Soumitra, y Brojo Kishore Mishra (2020), «Parts-of-speech tagging in NPL: utility, types, and some popular POS taggers», en Brojo Kishore Mishra y Raghvendra Kumar (eds.), *Natural language processing in artificial intelligence*, Palm Bay, Apple Academic Press: 131-165.

Gupta, Aastha, Rachna Rajput, Richa Gupta, y Monika Arora (2014), «Improved POS tagging for unknown words», *International Journal of Soft Computing and Engineering*, 4: 47-50.

Jamatia, Anupam, Jjörn Gambäck y Amitava Das (2015), «Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages», *Proceedings of Recent Advances in Natural Language Processing*: 239–248.

Jørgensen, Anna K., Dirk Hovy y Anders Søgaard (2015), «Challenges of studying and processing dialects in social media», *Proceedings of the Workshop on Noisy User-Generated Text*: 9-18.

Kanade, Aditya, Petros Maniatis, Gogul Balakrishnan y Kensen Shi (2020), «Learning and evaluating contextual embedding of source code», *International Conference on Machine Learning*: 5110-5121.

Kumawat, Deepika, y Vinesh Jain (2015), «POS tagging approaches: a comparison», *International Journal of Computer Applications*, 118 (6): 32-38.

Landolsi, Mohamed Y., Lotfi Ben Romdhane, y Lobna Hlaoua (2024), «Hybrid medical named entity recognition using document structure and surrounding context», *The Journal of Supercomputing*, 80 (4): 5011-5041.

Martínez, Ángel R. (2012), «Part-of-speech tagging», *Wiley Interdiciplinary Reviews: Computational Statistics*, 4 (1): 107-113. DOI: 10.1002/wics.195.

Neunerdt, Melanie, Bianca Trevisan, Michael Reyer y Rudolf Mathar (2013), «Part-of-Speech tagging for social media texts», en Iryna Gurevych, Chris Biemann y Torsten Zesch (eds.) *Language processing and knowledge in the web: lecture notes in computer science*, Berlín/Heidelberg, Springer: 139-150. DOI: 10.1007/978-3-642-40722-2_15.

Rojo, Guillermo (2021), *Introducción a la lingüística de corpus en español*, Londres/Nueva York, Routledge.

Rozovskaya, Alla, Richard Sproat, y Elabbas Benmamoun (2006) «Challenges in processing colloquial Arabic», *Proceedings of the International Conference on the Challenge of Arabic for NLP/MT*: 4-14.

Sánchez-Cartagena, Víctor M., Juan Antonio Pérez-Ortiz, y Felipe Sánchez-Martínez (2024), «Understanding the effects of word-level linguistic annotations in under-resourced neural machine translation», *arXiv*: 2401.16078. DOI: 10.48550/arXiv.2401.16078.

Taulé, Mariona, M. Antonia Martí, Ann Bies, Montserrat Nofre, Aina Garí, Zhiyi Song, Stephanie Strassel, y Joe Ellis, J. (2015), «Spanish treebank annotation of informal non-standard web text», en *Current Trends in Web Engineering: 15th International Conference, ICWE 2015 Workshops*, Rotterdam, Springer International Publishing: 15-27.

Tintinago, Alfonso, Yordan Muñoz, Gustavo A. Uribe, y Pedro H. Álvarez (2018), «Etiquetado asistido de documentos de investigación mediante procesamiento de lenguaje natural y tecnologías de la web semántica», *Scientia et Technica*, 23 (4), 528-537.

Tsai, Yu-Fang, y Keh-Jiann Chen (2003), «Context-rule model for POS tagging», *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, ACL Anthology: 146-151.

Ying, Zelin, Chen Li, Yu Dong, Qiuqiang Kong, Qiao Tian, Yuanyuan Huo, y Yuxuan Wang (2024), «A unified front-end framework for English text-to-speech synthesis», *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: 10181-10185.

Beatriz Méndez Guerrero (ORCID: 0000-0001-9510-9989), «Los corpus orales en la investigación pragmática: el caso de la locución *un poco*». *Biblioteca de Babel: Revista de Filología Hispánica*, vol. extra. 2 (2024), pp. 129-147.
**DOI**: 10.15366/bibliotecababel2024.extra2.005 · **ISSNe**: 2695-6349 · Recibido: 02/07/2024 · Aceptado: 07/10/2024

# Los corpus orales en la investigación pragmática: el caso de la locución *un poco*

Beatriz Méndez Guerrero
*Universidad Autónoma de Madrid*
beatriz.mendez@uam.es

**Resumen**: La investigación lingüística basada en corpus ofrece datos empíricos esenciales para identificar patrones recurrentes y llegar a interpretaciones certeras en los estudios pragmáticos. En este artículo, se evidencia la utilidad de tres corpus orales (PRESEEA-Palma, Val. Es.Co. y COJEM) en el estudio de la locución *un poco* en español. Para realizar la investigación, se ha partido de la metodología cualitativa y cuantitativa, la cual ha permitido examinar los usos, valores y frecuencias de la expresión, así como explorar sus variantes morfológicas. Los resultados del estudio sugieren que se puede establecer una distinción entre usos semánticos con carácter cuantificador de la locución y usos pragmáticos con valores como atenuador, intensificador, aproximativo y reformulador (Vigara Tauste 1992; Fuentes y Alcaide 2002; Matte Bon 1995; Briz 2005; *NGLE* 2009; Mariottini 2012; Repede 2023, 2024). Asimismo, indican que el género discursivo y las variables sociales de los hablantes inciden parcialmente en el uso que se hace de la locución. Las conclusiones del trabajo llevan a pensar que los corpus aportan información crucial para la investigación comunicativa y que constituyen una herramienta fundamental para los investigadores.

**Palabras clave**: pragmática, corpus orales, *un poco*, *un poquito*, usos semánticos, usos pragmáticos.

## Oral corpora in pragmatic research: the case of the expression *un poco*

**Abstract**: Corpus-based linguistic research provides essential empirical data for identifying recurring patterns and arriving at accurate interpretations in pragmatic studies. This article demonstrates the utility of three oral corpora (PRESEEA-Palma, Val.Es.Co. and COJEM) in studying the expression *un poco* in Spanish. The research employs both qualitative and quantitative methodologies, allowing for an examination of the uses, values, and frequencies of the expression, as well as an

exploration of its morphological variants. The study's results suggest that a distinction can be made between semantic uses with a quantifying character and pragmatic uses with attenuating, intensifying, approximative, and reformulating values (Vigara Tauste 1992; Fuentes y Alcaide 2002; Matte Bon 1995; Briz 2005; *NGLE* 2009; Mariottini 2012; Repede 2023, 2024). They also suggest that the discursive genre and social variables of the speakers partially influence the use made of the expression. The study's conclusions indicate that corpora provide crucial information for communicative research and constitute a fundamental tool for researchers.

**Keywords**: pragmatics, oral corpora, *un poco*, *un poquito*, semantic uses, pragmatic uses.

## 1. Introducción: los corpus orales en la investigación pragmática[1]

Cada vez se ve más clara la necesidad de estudiar los fenómenos lingüísticos a partir de datos empíricos que, en cantidad suficiente, proporcionen evidencias de las características del objeto de estudio y de la realidad que se quiere observar. La lingüística basada en corpus ha permitido establecer una relación entre la teoría y los datos para comprobar ideas que anteriormente se fundamentaban muchas veces en las impresiones del investigador. También ha hecho más sencillo el análisis contrastivo entre lenguas, variedades lingüísticas y variables de cualquier tipo, así como la exploración de los aspectos cuantitativos y probabilísticos (Torruella y Llisterri 1999; Jucker, Schneider y Bublitz 2018).

Los corpus lingüísticos incluyen desde pequeñas compilaciones de datos específicos, seleccionados manualmente, hasta enormes colecciones (semi)automáticas de datos disponibles electrónicamente; desde datos formateados y anotados para un objetivo de investigación concreto hasta corpus más generales; y desde corpus exclusivamente basados en texto hasta corpus con archivos de audio y vídeo (Landert *et al.* 2023). La lingüística de corpus ofrece algunas ventajas que, de acuerdo con Landert *et al.* (2023: 7-8), pueden resumirse en cinco puntos: (1) la búsqueda de patrones, puesto que son conjuntos de datos bien definidos que permiten la identificación de patrones recurrentes; (2) la

[1] Esta investigación ha sido posible gracias a los proyectos Corpus de composicionalidad e informatividad léxica: anotación, análisis y aplicaciones (INFOLEXIS) (referencia PID2022-138135NB-I00) y Estudio de los condicionantes sociales del español actual en el centro y norte de España: nuevas identidades, nuevos retos, nuevas soluciones (referencia PID2023-148371NB-C42), financiados por MCIN/AEI/10.13039/501100011033 y por el FSE.

sistematicidad, ya que los corpus incluyen muestras que representan el uso de la lengua en una variedad o dominio determinado de forma sistemática; (3) la generalización, pues las observaciones de patrones recurrentes en diferentes conjuntos de datos permiten generalizar los hallazgos para el uso de la lengua, siempre que el corpus se compile a partir de un muestreo representativo; (4) la reproducibilidad, en la medida en que los datos del corpus sean accesibles para otros investigadores y las explicaciones metodológicas se presenten con suficiente detalle; y (5) la transparencia, cuando se use una metodología lingüística de corpus bien establecida y suficientemente detallada para que pueda reproducirse.

En la investigación pragmática, son especialmente útiles los corpus orales conformados por muestras de lengua hablada. Estos corpus constituyen, actualmente, la base de la descripción, explicación y teorización lingüística en cualquiera de los niveles de la lengua. Son numerosos los corpus orales que existen ahora para el estudio del español. Destacan, por ejemplo, los corpus PRESEEA, Val.Es.Co., CORPES XXI, COSER, COLA, C-Oral-Rom, COJEM y, en menor proporción, el CREA, entre otros. De los anteriores, resultan muy interesantes para la pragmática los corpus orales que recogen muestras de habla coloquial y espontánea. Estos representan el género más natural y cotidiano de los hablantes y son, asimismo, los que dan cuenta en mayor medida de cómo emergen las nuevas estructuras lingüísticas y de cómo se producen los procesos de gramaticalización y lexicalización que originan el cambio lingüístico (Recalde y Vázquez 2009).

Ahora bien, la compilación de estos corpus no está exenta de problemas y la mayoría de los investigadores que recogen muestras de este tipo se enfrentan a obstáculos éticos, a «dificultades técnicas y metodológicas que supone recabar muestras de conversación natural» (Recalde y Vázquez 2009: 53) y a largos periodos de recolección. Asimismo, los corpus todavía tienden a caracterizarse «por la falta de acceso al contexto, privilegiando los resultados cuantitativos sobre la interpretación cualitativa y centrándose en las formas lingüísticas más que en sus funciones, todo lo cual puede dificultar los estudios pragmáticos» (Landert *et al.* 2023: la traducción es nuestra).

Este trabajo tiene como objetivo principal realizar un estudio pragmático de la locución *un poco* en español a partir del análisis de tres corpus orales: PRESEEA-Palma, Val.Es.Co. y COJEM. Con ello se pretende mostrar la utilidad de los corpus orales como herramientas para la recogida de datos y el estudio lingüístico.

## 2. La locución *un poco*

La expresión *un poco* se describe en la *Nueva gramática de la lengua española* (*NGLE* 2009) como una locución cuantificadora que, desde una perspectiva nocional, presenta usos semánticos y usos pragmáticos. En el caso de los primeros, el verbo restringe semánticamente su complemento directo como en el ejemplo *Había un poco de comida en el frigorífico* (§ 20.8d), donde *un poco* se refiere a la comida y expresa únicamente cantidad pequeña. En los segundos, la información pragmática se obtiene en el entorno discursivo y actúa tanto en el nivel del contenido proposicional como en el nivel ilocutivo o de fuerza del acto, ya que puede dar lugar a procesos de modificación performativa. Es lo que ocurre en *Dame un poco de tu bocadillo*, donde *un poco* pretende aumentar la vaguedad expresiva y disminuir la fuerza de la petición (Vigara Tauste 1992; Fuentes y Alcaide 2002; Matte Bon 1995; Briz 2005; *NGLE* 2009; Mariottini 2012; Repede 2023, 2024).

Los valores de la locución se han establecido a lo largo de las últimas décadas. A los valores de frecuencia (*Le gusta pasar un poco de tiempo en casa*), intensidad (*Se esfuerza un poco*) y tiempo (*Viene un poco tarde*), propuestos por Matte Bom (1995) y la ***NGLE*** (§ 20.8c), se añaden los de intensificador escalar (*Es un poco maleducado*), atenuador (*Llegaré un poco tarde*), aproximador (*La sintaxis es un poco como las matemáticas*), operador modal-enunciativo (*Aproximémonos un poco al concepto de interlengua*) y reformulador (*Vivir en el campo me parece aburrido, puedes entretenerte solo hasta cierto punto, un poco pasear, tomar el fresco, pero ya*), estudiados por Vigara Tauste (1992), Haverkate (1994), Sánchez (1999), Fuentes y Alcaide (2002), Briz (2005), Sedano y Guirado (2009), Mariottini (2012) y Repede (2023, 2024), entre otros.

Haverkate (1994) explica que *un poco* tiene valor mitigador cuando se intenta disminuir el valor peyorativo de lo expresado o atenuar la calificación negativa que se hace de una persona u objeto. Un ejemplo de este valor sería el observado en *¿Me dejarías un poco de dinero?* Este uso se encuentra igualmente en otras lenguas, entre ellas el francés (*un peu*), como explica Ducrot (1970). Por otra parte, se puede recurrir a la locución *un poco* cuando lo que se pretende es intensificar el referente (Fuentes y Alcaide 2002; Sedano y Guirado 2009; Mariottini 2012; Repede 2023, 2024). En estos casos, *un poco* transforma lo dicho en una propiedad negativa. Este *un poco* atenúa la propiedad que modifica, como explica Mariottini (2012), pero al mismo tiempo provoca un efecto pragmático de realce o intensificación, tal y como se aprecia en *Estás siendo un poco exagerado, ¿no te parece?* Dicha propiedad se puede intensificar aún más con el uso del diminutivo *Estás siendo un poquito exagerado, ¿no te parece?* (Sedano y Guirado 2009).

El valor aproximativo, descrito por Fuentes y Alcaide (2002), alcanza para Mariottini (2012) una doble dimensión: una oracional, muy cerca de otros aproximativos como *casi*, *una especie de*, *en cierta medida*, que, como se aprecia en *La vasija parece un poco romana*, indican cercanía a un punto tomado como referencia (Schwenter y Pons Bordería 2005); y otra conversacional, en el interior de una secuencia de apertura y tras un verbo imperativo exhortativo con la que se pretende seleccionar un tópico discursivo y un participante, como en el caso de *Vamos a hablar ahora un poquito de las diversas técnicas de formulación*.

También se ha descrito para *un poco* un uso modal-enunciativo que aparece normalmente en situaciones formales o no estrictamente coloquiales. Se considera un modalizador que pretende hacer un uso afectivo del discurso para amenizarlo o mostrar proximidad con el destinatario (Fuentes y Alcaide 2002; Mariottini 2012). Un ejemplo de este tipo sería el propuesto por Mariottini (2012), *Lo que tiene un poquito de misterio es que…*, expresado durante una comunicación académica. Por último, Repede (2023, 2024) asigna un valor reformulador a la expresión *un poco* en los casos en los que sirve para definir un concepto o una idea cuando el hablante considera que no están quedando claros. Este valor es el que se observa, por ejemplo, en *Estudiar matemáticas es relativamente sencillo, es un poco que tienes que recordar lo que ya sabes*.

La locución *un poco* presenta, además, algunas restricciones y preferencias de uso según la situación comunicativa, el registro o la variedad dialectal del hablante, tal y como sugiere la NGLE (2009: § 20.8i). En el español coloquial de Colombia y Venezuela, por ejemplo, se utiliza a veces con el sentido de *mucho* o *bastante* en expresiones eufemísticas del tipo *Es un poco tonto* o *Está un poco borracho*. Con la incorporación del diminutivo -ito (*Es un poquito burra*), se acentúa, además, la cualidad negativa que se predica ('es muy burra' o 'es burrísima'). También se admite en estas variedades el uso de la locución para expresar cantidad superior a la normal como en el ejemplo *En el concierto había un poco de gente*, en el sentido de 'mucha gente'. En el español de Chile, las Antillas y algunos países centroamericanos, como Costa Rica, se utiliza la expresión *un buen poco* como equivalente a *mucho* (*Todos sabemos que falta un buen poco para eso*).

Hasta el momento, se han registrado un buen número de variantes morfológicas de la expresión. Destacan en este sentido *un poquico*, *un poquín*, *un poquino*, *un poquitico*, *un poquillo*, *un poquitillo*, *un poquitín*, *un poquitito*, *un poquito*, *un poconón* y *un poconotóncon* con sus variantes de género y número (*NGLE* 2009: § 9.2e). En los corpus orales manejados en esta investigación, solo se han registrado casos de *un poco* y *un poquito*. En los siguientes apartados, se presenta el análisis de los corpus PRESEEA-Palma, Val.Es.Co. y COJEM para determinar el uso que se realiza en español de la expresión *un poco* y su variante morfológica *un poquito*.

## 3. Metodología

Para realizar el estudio se ha partido del análisis de tres corpus orales que reflejan la producción oral de hablantes españoles de la variedad castellana. En primer lugar, se ha recurrido al estudio de los casos de la locución *un poco* y su variante morfológica *un poquito* en el corpus PRESEEA-Palma, compuesto por 54 entrevistas semidirigidas de unos 45 minutos de duración. Los informantes de la muestra están estratificados por edad, sexo, nivel de instrucción y grupo etnolingüístico. El segundo corpus manejado ha sido el Val.Es.Co., que cuenta con un total de 72 conversaciones coloquiales también estratificadas por edad, sexo y clase social. Por último, se ha utilizado el corpus COJEM organizado en torno a 7 conversaciones espontáneas de casi 3 horas de duración (en total 20 horas de grabación) entre mujeres y hombres jóvenes con estudios universitarios. Con la selección de estos tres corpus, se ha pretendido, por una parte, controlar el origen de los informantes (hablantes de castellano de la zona oriental de España) y, por otra, permitir el contraste entre dos géneros textuales orales —la conversación y la entrevista semidirigida— para la identificación de fenómenos discursivos similares y diversos relacionados con estos géneros. Véase la información detallada sobre cada uno de los corpus en la Tabla 1 (siguiente página).

El análisis se ha centrado en determinar los usos semánticos y pragmáticos de la locución *un poco* y su variante morfológica *un poquito* que utilizan los hablantes de las muestras y en medir los valores que le otorgan a cada caso. El estudio sigue una metodología mixta de tipo cualitativo-cuantitativo, si bien, por limitaciones de espacio, no se ha tenido en cuenta las variables sociales que, probablemente, incidirán en algunos de los usos observados. No obstante, se mencionarán algunas de estas incidencias en los resultados para dar cuenta de su posible influencia. En el estudio, se han analizado todos los casos de *un poco* y *un poquito* registrados en los tres corpus, a excepción de los usos cortados o los casos registrados en pasajes inciertos de las grabaciones.

| | PRESEEA-Palma | Val.Es.Co. | COJEM |
|---|---|---|---|
| Tamaño de la muestra | 54 entrevistas de 45 minutos (2 430 minutos aproximadamente) | 72 conversaciones en su mayoría coloquiales (1707 minutos aproximadamente) | 7 conversaciones coloquiales de 3 horas (3 330 minutos aproximadamente) |
| Estratificación | 54 informantes:<br>Sexo: 27 mujeres y 27 hombres<br>Edad: 18 jóvenes (18-34 años); 18 adultos (35-55 años); y 18 mayores (más de 55 años)<br>Nivel de instrucción: 18 con estudios primarios, 18 con estudios secundarios; y 18 con estudios universita-rios.<br>Grupo etnolingüístico: 18 castellanohablantes; 18 catalanohablantes; y 18 bilingües equilibrados | 226 informantes:<br>Edad: 86 jóvenes (18-34 años); 28 adultos (35-55 años); y 24 mayores (más de 55 años)<br>Clase social: 38 baja, 132 media y 56 alta<br>La estratificación por sexo y grupo etnolingüístico puede consultarse en cada una de las fichas técnicas de las conversaciones | 10 informantes:<br>Sexo: 5 mujeres y 5 hombres<br>Edad: menores de 25 años<br>Nivel de instrucción: estudios universitarios<br>Grupo etnolingüístico: castellanohablantes |
| Fecha de recogida | Entre 2010 y 2014 | Entre 1994 y 2022 | Entre 2010 y 2012 |
| Acceso | https://www.corpusmallorca.es/preseea/ | https://www.valesco.es/#/pa-ges/cod_hj3y7hwvuuajtlkq0ik/cod_fa393ih5l4jx9zssv7 | https://ebuah.uah.es/dspace/handle/10017/25298 |

*Tabla 1. Información sobre los corpus orales utilizados en el estudio*

## 4. Resultados y discusión de los usos de *un poco* y *un poquito* en los corpus PRESEEA-Palma, Val.Es.Co. y COJEM

El apartado de resultados y discusión está dividido en dos partes. En primer lugar, se presenta el análisis discursivo de *un poco* y *un poquito*, atendiendo los usos globales y estableciendo una distinción entre los usos semánticos y los pragmáticos. Y, posteriormente, se examinan los valores de la locución y su variante morfológica de los que los informantes han hecho uso: atenuador, intensificador, aproximativo y reformulador (Fuentes y Alcaide 2002; Sedano y Guirado 2009; Mariottini 2012; Repede 2023, 2024, entre otros).

## *4.1. Análisis discursivo de* un poco *y* un poquito *en los corpus analizados*

El primer dato reseñable de la investigación es que se han registrado un total de 916 casos de *un poco* y *un poquito* en los tres corpus estudiados. Predomina claramente el uso de *un poco* y, en menor medida, de *un poquito* en el corpus de entrevistas PRESEEA-Palma, frente a los usos de estas expresiones en los corpus conversacionales Val.Es.Co. y COJEM. La distribución de los casos puede consultarse con detalle en la Tabla 2:

| | **PRESEEA-Palma** | **Val.Es.Co.** | **COJEM** |
|---|---|---|---|
| *un poco* | 564 (61,6 %) | 59 (6,4 %) | 110 (12 %) |
| *un poquito* | 157 (17,1 %) | 18 (2 %) | 8 (0,9 %) |
| **Total** | **721 (78,7 %)** | **77 (8,4 %)** | **118 (12,9 %)** |

*Tabla 2. Frecuencias de uso de* un poco *y* un poquito.

Atendiendo a la duración de las muestras estudiadas, 2 430 minutos aprox. del corpus PRESEEA-Palma, 1 707 minutos aprox. del corpus Val.Es.Co. y 3 330 minutos aprox. del corpus COJEM, se observa que la proporción en el uso de *un poco* en los dos corpus orales espontáneos (Val.Es.Co. y COJEM) apenas difiere. En Val.Es.Co., se localiza un caso de la locución *un poco* cada 28,9 minutos de grabación y, en el COJEM, cada 30,3 minutos. Esta proximidad en los resultados no se halla, sin embargo, en el caso de *un poquito*, donde la diferencia entre los corpus es mucho mayor: un caso cada 94,8 minutos en Val.Es.Co. y un caso cada 416,2 minutos en el COJEM. Como se apunta más abajo, es posible que el uso menor de *un poquito* que hacen los informantes del COJEM tenga que ver con las características sociológicas de los individuos de la muestra: jóvenes menores de 25 años con estudios universitarios. No obstante, es necesario mostrarse cautos en el análisis actual y, por ello, será necesario realizar estudios más exhaustivos en el futuro que confirmen o desmientan esta afirmación y permitan determinar si nos encontramos o no ante un cambio lingüístico en marcha impulsado por los jóvenes.

En cuanto a la frecuencia de uso de la expresión *un poco* y su variante *un poquito* en el corpus PRESEEA-Palma, como se ha indicado ya, la proporción por minuto es mucho mayor que en los otros dos corpus analizados (un caso cada 3,4 minutos en la expresión *un poco* y un caso cada 15,5 minutos en *un poquito*). Posiblemente haya incidido en esta diferencia el hecho de que las muestras del primer corpus y de los últimos pertenezcan a géneros textuales diferentes (la entrevista semidirigida con grabación consciente en el primer caso y la conversación espontánea con grabación secreta en los últimos). Esta línea argumental se relaciona con la de otras investigaciones previas que

asocian diferencias discursivas, por ejemplo en el uso de la atenuación, con el género discursivo (entrevista semidirigida o conversación) en el que se enmarca la enunciación (Estellés y Cabedo, 2017).

Si nos centramos ahora en los usos semánticos y pragmáticos registrados en la muestra, se aprecia que predominan los pragmáticos, con 628 casos totales (68,6 %), sobre los semánticos, con 288 casos (31,4 %), lo cual sugiere una consolidación de los usos pragmáticos de la locución. Estos resultados coinciden con la tendencia observada en *un poco* por Repede (2023, 2024) en el corpus PRESEEA-Sevilla y con los usos registrados, en PRESEEA-Palma, en otras expresiones como *lo que pasa es que* (Méndez, 2022). Como decimos y como se aprecia también en las Tablas 3 y 4, la frecuencia de los usos pragmáticos duplica a la de los usos semánticos, si bien la proporción de la locución *un poco* y su variante morfológica *un poquito* se mantiene constante y equilibrada respecto a la presentada en la Tabla 2, lo cual demuestra la consistencia de los datos:

| | **PRESEEA-Palma** | **Val.Es.Co.** | **COJEM** |
|---|---|---|---|
| *un poco* | 189 (65,6 %) | 14 (4,9 %) | 21 (73 %) |
| *un poquito* | 54 (18,7 %) | 9 (3,1 %) | 1 (0,3 %) |
| **Total** | **243 (84,4 %)** | **23 (8 %)** | **22 (7,6 %)** |

*Tabla 3. Frecuencias de usos semánticos de* un poco *y* un poquito.

| | **PRESEEA-Palma** | **Val.Es.Co.** | **COJEM** |
|---|---|---|---|
| *un poco* | 375 (59,7 %) | 45 (7,2 %) | 89 (14,2 %) |
| *un poquito* | 103 (16,4 %) | 9 (1,4 %) | 7 (1,1 %) |
| **Total** | **478 (76,1 %)** | **54 (8,6 %)** | **96 (15,3 %)** |

*Tabla 4. Frecuencias de usos pragmáticos de* un poco *y* un poquito.

Las mayores frecuencias de usos semánticos de *un poco* y *un poquito* se producen en el corpus PRESEEA-Palma, coincidiendo con lo que ya se vio en los datos globales (Tabla 2). Llama la atención el uso de la expresión *un poquito*, tanto en los usos semánticos como en los pragmáticos, en mayor medida, en el corpus Val.Es.Co., que cuenta con una muestra estratificada por edad y nivel formativo, frente a la del corpus COJEM, constituido únicamente por mujeres y hombres jóvenes con estudios universitarios. Es posible, como se indicaba anteriormente, que este resultado se vea afectado por las variables edad y nivel de instrucción de los informantes, coincidiendo con lo que ya se ha visto en otros estudios sobre el diminutivo en español en los corpus PRESEEA (Paredes, 2011; León-Castro, 2020; León-Castro y Jiménez, 2022, Repede, 2023, entre otros).

Para profundizar en esta idea, contrastamos a continuación los casos registrados en el corpus PRESEEA-Palma con los resultados de los estudios previamente citados, realizados sobre los corpus

PRESEEA-Madrid y PRESEEA-Sevilla. En primer lugar, resulta reseñable que, en el PRESEEA-Palma, sean el segundo y tercer grupo de edad los que utilizan *un poquito* con mayor frecuencia, dato que se ve reforzado en el corpus COJEM, exclusivamente de jóvenes mallorquines, que presenta frecuencias muy bajas (Tabla 4). Este resultado contraviene, sin embargo, lo observado por Paredes (2011) en el corpus PRESEEA-Madrid o por León-Castro (2020), León-Castro y Jiménez (2022) y Repede (2023) en el corpus PRESEEA-Sevilla, donde eran los jóvenes los propulsores del uso del diminutivo -ito. No obstante, conviene mostrarse cautos con esta diferencia al tratarse de estudios con planteamientos distintos. Respecto al nivel formativo, sí se observan más similitudes entre los datos registrados en el presente estudio y los consultados en estudios previos, pues son los informantes con estudios superiores/universitarios los que más utilizan *un poquito*, al igual que ocurría en Paredes (2011) con la variedad del castellano madrileño y Repede (2023) en la variedad andaluza. En León-Castro (2020) y León-Castro y Jiménez (2022), en cambio, fue el grupo de hablantes sevillanos con menor formación educativa quien utilizó con una frecuencia mayor el diminutivo *-ito*. Queda pendiente, pues, la revisión de este resultado, atendiendo las diferencias diastráticas y diatópicas.

En otro orden de cosas, en lo que se refiere a los usos semánticos registrados en la muestra, estos funcionan, tal y como se explicó anteriormente, como cuantificadores estrictos en combinación con sustantivos no contables, adjetivos o adverbios o complementando a un verbo, en consonancia con lo observado en la NGLE:

(1) **I**: pues sí / hasta que me aclare **un poco** / y nada / que no he tenido muy buena experiencia la verdad (PRESEEA-Palma).

(2) **I**: **un poco** / bueno **un poco** de mantenimiento / **un poco** de pesas / luego miro **un poco** de natación / también a veces voy a correr / **un poco** de todo / hasta hace 2 años siempre he jugado a fútbol (PRESEEA-Palma).

La locución expresa, en estos casos, cantidades pequeñas de alguna propiedad. En las situaciones en las que los usos de la locución tienen un valor pragmático, la noción va más allá del significado cuantificador proposicional y añade al enunciado una fuerza ilocutiva potencialmente perlocutiva:

(3) **I**: estaba muy mal distribuida porque es un edificio muy antiguo y **un poco** mal construido y estaba / estaba muy mal distribuida y mi padre (chasquido de boca) se encaprichó mi padre es constructor / y se le encaprichó // y porque él él es

muy artista y él / como artista que es / es muy cuco y lo quiso ///(1) todo hacer a su gusto // y la verdad es que le quedó muy bonito (PRESEEA-Palma).

(4) I: me parece **un poco** locura / no sé / ya te digo yo creo que para tener un hijo / yo mi idea es esperar a que / bueno no sé cuántos años tendré / pero la cuestión es que / quiero tener una situación económica / buena (PRESEEA-Palma).

En los ejemplos (3) y (4), la locución se usa para tratar de mitigar o, como explica Vigara Tauste (1992: 393), de «matizar la expresión del significado para atenuar los efectos del sentido para conseguir una mayor aceptación de lo que decimos por parte de nuestro/s interlocutor/es».

## *4.2. Valores de* un poco *y* un poquito *en los corpus analizados*

Si nos centramos ahora en los valores de la expresión *un poco* y su variante morfológica *un poquito*, también observamos resultados reseñables. Como se aprecia en la Tabla 5, el valor más frecuente en los corpus analizados es el atenuador con un total de 372 casos (59,2 %), como también se observó en Repede (2023, 2024). Le siguen el valor intensificador con 113 casos (18 %), el aproximativo con 94 casos (15 %) y el reformulador con 49 casos (7,8 %).

| | | **PRESEEA-Palma** | **Val.Es.Co.** | **COJEM** |
|---|---|---|---|---|
| Intensificador | *un poco* | 82 (13,1 %) | 6 (1,1 %) | 10 (1,7 %) |
| | *un poquito* | 13 (2,1 %) | 0 (0 %) | 2 (0,4 %) |
| Atenuador | *un poco* | 182 (28,9 %) | 36 (5,5 %) | 56 (9 %) |
| | *un poquito* | 84 (13,4 %) | 9 (1,4 %) | 5 (0,9 %) |
| Aproximativo | *un poco* | 73 (11,5 %) | 2 (0,4 %) | 14 (2,5 %) |
| | *un poquito* | 5 (0,8 %) | 0 (0 %) | 0 (0 %) |
| Reformulador | *un poco* | 38 (6,1 %) | 1 (0,2 %) | 9 (1,4 %) |
| | *un poquito* | 1 (0,2 %) | 0 (0 %) | 0 (0 %) |
| **Total** | | **478 (76,1 %)** | **54 (8,6 %)** | **96 (15,9 %)** |

*Tabla 5. Frecuencias de los valores de* un poco *y* un poquito.

Respecto a las proporciones de uso de *un poquito* en relación con los valores de la locución, se aprecia una clara preferencia por la variante morfológica con el valor atenuador frente al resto de los valores. Este resultado se observa principalmente en el corpus PRESEEA-Palma, aunque la tendencia se repite en Val.Es.Co. y en el COJEM. No se han localizado casos de *un poquito* para el resto de los valores en los corpus conversaciones, salvo dos casos en el COJEM con valor intensificador, que permiten aumentar el grado de intensificación (Sedano y Guirado

2009). Esta tendencia, tan similar en los tres corpus, sugiere que la variante morfológica *un poquito* está muy focalizada en los usos atenuadores de la locución.

El valor atenuador es el más versátil de todos, pues, como ya observaron Haverkate (1994) y Matte Bon (1995), la mitigación puede aplicarse tanto a la disminución del significado peyorativo, como a las calificaciones negativas de la persona o del objeto referido o al impacto en la imagen social de quien emite el enunciado:

(5) I: no / es una urbanización / o sea solo hay una zo mmm una carretera para llegar / es **un poco** estrechita y tal / pero claro hombre / hay como 4 o 5 urbanizaciones y todas todas salen y todas entran / hay veces que se montan atascos ahí que te cagas pero (PRESEEA-Palma).

(6) I: me da **un poco** de envidia sana // por pero me alegro mucho por ellos porque a lo mejor es gente que lo ha pasado muy mal porque ahora con los tiempos que corren pues ///(1) mmm es una alegría / me da envidia pero es sana // (PRESEEA-Palma).

(7) I: bueno hay **un poco** de rifi rafe ahora porque desde que la jefa está medio estudiando algo que a nosotras no nos gusta pero bueno lo que es el ambiente hasta ahora sí lo que pasa es que cuando hay un punto negro en algo todo se empieza como a remover (PRESEEA-Palma).

(8) I: sí sí sí de hecho pues nos gusta siempre cocinar lo que sea o o esmerarnos **un poquito** ya que es Navidad pero siempre nos sale bien ///(1) (PRESEEA-Palma).

En los ejemplos (5) y (7) la locución atenúa la calidad negativa que se le atribuye al referente, la urbanización (*un poco estrechita*) en el primer caso y a la relación con la jefa (*un poco de rifi rafe*) en la segunda. En (5), además, dicha atenuación viene reforzada por el uso del diminutivo en *estrechita*. En cambio, en los ejemplos (6) y (8), se pretende que la mitigación aplique sobre la imagen de quien habla, pues podría parecer que es una persona envidiosa en (6) (*me da un poco de envidia sana*) o presuntuosa en (8) (*esmerarnos un poquito ya que es Navidad*). En este último caso, vemos cómo el uso del diminutivo o variante morfológica *un poquito* contribuye a este fin y aumenta el valor atenuador de la locución.

El valor intensificador, por su parte, se utiliza en las muestras para maximizar la fuerza ilocutiva del acto, reforzar el punto de vista propio y realzar la afectación de los implicados en la actividad:

(9) **I**: me parece **un poco** locura / no sé / ya te digo yo creo que para tener un hijo / yo mi idea es esperar a que / bueno no sé cuántos años tendré / pero la cuestión es que / quiero tener una situación económica / buena (PRESEEA-Palma).

(10) **I**: encuentro que tiene **un poco** de culpa la televisión y las tiendas de ropa / influyen en eso // influyen en eso porque / porque tú vas a comprarte un pantalón y y ves nada más tallas pequeñas y tallas pequeñas y tallas pequeñas / entonces ves que que no te entran / ves que los los maniquís son / son nada / o sea son unas mujeres muy muy muy delgadas /// (1) igual que en la televisión las modelos están en los huesos entonces pues tú intentas ser como ellas / ves que ellas son guapas / que ellas son famosas / que tal por el cuerpo // y supongo que estas chicas pues intentan parecerse a ellas // yo desde luego yo personalmente lo veo una tontería / cada uno es como es / está claro que si te ves con quilos de más pues vas a intentar adelgazar pero ya no / ya no como una obsesión por el peso / o sea eso ya lo veo / exagerado (PRESEEA-Palma).

(11) **I**: bueno no sé / creo que de momento soy **un poco** joven para tener hijos pero sí que siempre me ha gustado porque sí me gustan mucho los niños pequeños (PRESEEA-Palma).

(12) **I**: es una zona que / bueno / (chasquido de boca) me gusta por el / por la cercanía a las cosas pero sí que es verdad que / que cada vez más va / se va como mmm habiendo un ambiente **un poquito** más / más conflictivo / es una zona en donde está muy / muy próxima a según que zonas que son algo con conflictivas y que / y que bueno por el hecho de / de la inmigración y todo esto pues se van / se van formando muchos / muchos guetos (PRESEEA-Palma).

En (9), (11) y (12) *un poco* precede a adjetivos y adverbios y se usa para expresar una valoración negativa de lo que se cuenta o su inaceptabilidad, manifestando, como explica Matte Bon (1995: 77), cierta atención hacia las expectativas del interlocutor: «al no saber lo que éste espera, el hablante prefiere suponer que se trata de una valoración positiva o una aceptación de aquello de lo que se está hablando, y suaviza los argumentos negativos», como se ve en *un poco locura*, *soy un poco joven para tener hijos* o *un poquito más/ más conflictivo*. Por su parte, en (10), lo

que ocurre es que la locución aumenta o intensifica el valor negativo en lo dicho, sin necesidad de recurrir a un adjetivo o adverbio (*encuentro que tiene un poco de culpa la televisión y las tiendas de ropa*). Beinhauer (1978 [1929]) y Spitzer (2007) ven en el valor descrito en los ejemplos anteriores un doble uso mitigador e intensificador de la locución al que denominan eufemístico y que, con posterioridad, ha sido recogido también por Mariottini (2012) en el corpus Val.Es.Co.

Para acabar, se han localizado en las muestras usos aproximativos y reformuladores de *un poco*. De acuerdo con Fuentes y Alcaide (2002), el uso aproximativo de la locución se encuentra a medio camino entre la cuantificación, presente de forma más o menos implícita en la mayor parte de los usos registrados, y la atenuación. Este valor se caracteriza por flexibilizar los significados semánticos de los elementos de la proposición y por dotar al enunciado de una mayor imprecisión. Por su parte, el valor reformulador, descrito por Repede (2023, 2024), implica operaciones de reorganización discursiva tales como la explicitación, precisión, ampliación de información, rectificación o (auto)corrección:

(13) **I**: no / es la típica carpa de pueblo que viene un grupo y (risa = "I") lo que pasa que claro cada pueblo también tiene luego **un poco** su pub así para gente más joven y tal **un poco** así es (PRESEEA-Palma).

(14) **I**: no no // en ese momento creo que no eres consciente // hasta que un poco evalúas // lo que ha pasado / los daños / lo que hubiera podido ser // y ves a la gente / que viene a socorrerte / que en este caso // mis padres estaban de viaje y bueno / llamé a mi novia / (risa = "I") y vino mi novia // vino mi suegro y aparte bueno vino una patrulla de la guardia civil y demás / además de un conductor de autocar / que había presenciado el el accidente // y fue el único / quiero hacerlo constar (risa = "I") fue el único la única persona que se paró // a ver qué había pasado // porque pasaron varios coches ///(1,5) (PRESEEA-Palma).

(15) **I**: una especialización o máster o bueno / algo **un poco** relacionado así (PRESEEA-Palma).

(16) **I**: es **un poco** / bueno vale que te puedan hacer eeh mmm en la (ininteligible) te pueden dar una hoja de robo y tal pero necesito mi DNI / no puedo estar sin él (PRESEEA-Palma).

En los ejemplos (13) y (14), observamos el valor aproximativo de la locución, puesto que tanto en *tiene luego un poco su pub* como en *hasta que*

*un poco evalúas* se pretende trasmitir la idea de que tanto el pub como la evaluación del accidente se describen o realizan de forma imprecisa o aproximada. En cuanto a los ejemplos (15) y (16), el valor es otro y mucho más próximo al de un operador conversacional que pretende, en unos casos, introducir una ampliación de la información (*algo un poco relacionado así*) y, en otros, reformular y precisar la información que le precede (*es un poco / bueno vale que te puedan hacer […] una hoja de robo*).

## 5. Conclusiones

La utilidad de los corpus en la investigación lingüística es indudable, especialmente en lo que se refiere a la investigación de fenómenos pragmáticos. La lingüística basada en corpus ha demostrado ser necesaria en la investigación al ser una fuente fidedigna de la que obtener datos empíricos. Son este tipo de datos los que permiten identificar patrones recurrentes, generalizar hallazgos y reproducir estudios con precisión. Los corpus orales, que incluyen muestras de lengua hablada en contextos naturales, resultan especialmente valiosos, pues reflejan de manera fiel el uso cotidiano de la lengua y los procesos de cambio lingüístico.

En este trabajo se ha presentado un estudio basado en corpus orales para determinar el uso que se hace en español de la locución *un poco* y de su variante morfológica *un poquito*. El estudio ha seguido una metodología mixta (cualitativa-cuantitativa) que ha permitido determinar los usos semánticos y pragmáticos de la expresión en tres macrocorpus del español: el PRESEEA-Palma, el Val.Es.Co. y el COJEM. También se ha pretendido en el estudio diferenciar los valores de la locución en cada una de las muestras analizadas y localizar patrones de uso que refuercen o desmientan las afirmaciones existentes en la bibliografía.

Los resultados de la investigación indican un predominio evidente del uso de *un poco* y de su variante morfológica *un poquito* en el corpus PRESEEA-Palma con ocurrencias de un caso cada 3,4 segundos en *un poco* y un caso cada 15,5 segundos en *un poquito*. Como decimos, la frecuencia de la locución es significativamente superior a la observada en los corpus Val.Es.Co. y COJEM, que tienen un carácter más conversacional. Este resultado podría asociarse, como ya se ha hecho en otros estudios, con el género discursivo en el que se enmarcan los intercambios. Será interesante determinar en futuros estudios si la menor aparición de la locución en los corpus conversacionales guarda relación también con los diversos temas o tópicos tratados.

La variante *un poquito*, por su parte, ha demostrado tener una menor presencia en el corpus COJEM. A partir del contraste entre el corpus PRESEEA-Palma, estudiado aquí, y los corpus PRESEEA-Madrid y

PRESEEA-Sevilla, analizados en estudios previos sobre el sufijo *-ito*, se ha determinado que la edad y el nivel de instrucción son variables a tener en cuenta en el estudio de la locución y que será necesario profundizar en el análisis de dichas variables para determinar hasta qué punto inciden en la producción de la expresión.

Respecto a los usos globales de la expresión, se observa que predominan los pragmáticos, que suponen más de dos tercios de los 916 casos localizados en las muestras. Los usos semánticos de *un poco* son estrictamente cuantificadores, como ya explica la *NGLE*, y modifican a verbos y a adjetivos; mientras que los usos pragmáticos, si bien conservan ese valor cuantificador en mayor o menor medida, añaden nuevos valores a la locución, más allá de la proposición, que inciden sobre el acto comunicativo y sobre la perlocución. Los valores registrados en los corpus orales analizados son el atenuador, intensificador, aproximador y reformulador, ya descritos en la bibliografía previa (Vigara Tauste 1992; Fuentes y Alcaide 2002; Matte Bon 1995; Briz 2005; *NGLE* 2009; Mariottini 2012; Repede 2023, 2024).

Así pues, los resultados de este estudio refuerzan, en parte, las conclusiones de investigaciones anteriores que han identificado valores similares para *un poco* y *un poquito*. Siguiendo la línea de los estudios de Mariottini (2012) y Repede (2023, 2024), se han podido establecer en esta investigación los valores más frecuentes para la locución, con el valor atenuador y el intensificador como los predominantes. Todo ello sugiere una estabilidad en los usos pragmáticos de esta locución en diferentes zonas hispanohablantes. No obstante, al mismo tiempo, nuestra investigación apunta hacia distintos condicionantes, hasta ahora no estudiados, en el uso de la expresión (como, por ejemplo, el género discursivo) y señala la necesidad de realizar más análisis contrastivos que permitan explicar si los usos de la locución *un poco* y su variante *un poquito* se ven influenciados por las variables sociológicas asociadas a los hablantes (como parece ser en el caso de la edad y el nivel de instrucción), por los tópicos conversacionales o por la variedad diatópica de los hablantes en cuestión.

En el futuro, será imprescindible superar algunas de las limitaciones de la presente investigación y, como se ha dicho, determinar hasta qué punto existen diferencias relacionadas con las variables sociolingüísticas de los hablantes o con el tópico conversacional. También será interesante comparar distintas comunidades de prácticas o variedades geolectales para trazar una explicación más completa de los usos que hacen los hispanohablantes de la locución *un poco* y de su variante morfológica *un poquito*.

## Bibliografía

Beinhauer, Werner (1978 [1929]), *El español coloquial*, Madrid, Gredos.

Briz Gómez, Antonio (2005), «Eficacia, imagen social e imagen de cortesía», en Diana Bravo (ed.), *Estudios de la (des)cortesía en español*, Buenos Aires, Dunken: 53-91.

COJEM = Méndez Guerrero, Beatriz (2015), «Corpus Oral Juvenil del Español de Mallorca (COJEM)», *LinRed*, 13. Disponible en: http://hdl.handle.net/10017/25298.

Corpus PRESEEA-Mallorca <http:// https://www.corpusmallorca.es/preseea/> [12 de marzo de 2024].

Ducrot, Oswald, (1970), «*Peu* et *un peu*», *Cahiers de lexicologie*, 16 (1): 21-52.

Estellés Arguedas, Maria y Adrián Cabedo Nebot (2017), «La atenuación fónica en entrevistas (proyecto PRESEEA) y en conversaciones (corpus Val.Es.Co): un estudio de campo», *LinRed*, 15.

Fuentes Rodríguez, Catalina y Esperanza Alcaide Lara (2002), *Mecanismos lingüísticos de la persuasión*, Madrid, Arco/Libros.

Haverkate, Henk (1994), *La cortesía verbal*, Madrid, Gredos.

Landert, Daniela, Daria Dayter, Thomas C. Messerli, y Miriam A. Locher (2023), *Corpus Pragmatics*, Cambridge, Cambridge University Press. Doi:10.1017/9781009091107.

Jucker, Andreas H., Klaus P. Schneider, y Wolfram Bublitz (2018) (eds.), *Methods in Pragmatics*, Berlín/Boston, De Gruyter Mouton. DOI: 10.1515/9783110424928-022.

León-Castro Gómez, Marta (2020), «El empleo del diminutivo en la ciudad de Sevilla: perspectivas sociolingüística y pragmática», *Lengua y Habla*, 24: 112-131.

León-Castro Gómez, Marta, y Rafael Jiménez Fernández (2022), «La alternancia *–ito/-illo* en hablantes sevillanos de nivel educacional bajo: un estudio en tiempo real», *Literatura y Lingüística*, 45: 543-569.

Mariottini, Laura (2012), «Modalidad y atenuación: análisis de *un poco* y de sus alternaciones morfológicas en las conversaciones coloquiales», *Oralia*, 15: 177-204.

Matte Bon, Francisco (1995), *Gramática comunicativa del español: de la idea a la lengua*, Madrid, Edelsa.

Méndez Guerrero, Beatriz (2022), «La expresión gramaticalizada *lo que pasa es que* en español: estudio contrastivo de PRESEEA-Palma y PRESEEA-Alcalá». *Revista Signos: Estudios De Lingüística*, 55 (110): 844-872. DOI: 10.4067/S0718-09342022000300844.

*NGLE* = Real Academia Española y Asociación de Academias de la Lengua Española (2009), *Nueva gramática de la lengua española*, Madrid, Espasa.

Paredes García, Florentino (2011), «Variación en el uso del diminutivo en el habla de Madrid: avance de un estudio sociolingüístico», en Ana María Cestero Mancera, Isabel Molina Martos y Florentino Paredes García (eds.), *La lengua, lugar de encuentro: actas del XVI Congreso Internacional de la ALFAL*, Alcalá de Henares, Universidad de Alcalá: 3709-3719.

Recalde Fernández, Montserrat y Vázquez Rozas, Victoria (2009), «Problemas metodológicos en la formación de corpus orales», en Pascual Cantos Gómez y Aquilino Sánchez Pérez (eds.), *A survey of corpus-based research*, Murcia, Asociación Española de Lingüística del Corpus: 51-64.

Repede, Doina (2023), «La locución *un poco* en el corpus oral PRESEEA-Sevilla: funciones discursivas y distribución social», *Forma y Función*, 36 (1): 1-24. DOI: 10.15446/fyf.v36n1.97379.

Repede, Doina (2024), «Análisis sociopragmático de *un poco* en las entrevistas semidirigidas», *Onomázein*, 63: 1-19. DOI: 10.7764/onomazein.63.01.

Sánchez López, Cristina (1999), «Los cuantificadores: clases de cuantificadores y estructuras cuantitativas», en Ignacio Bosque y Violeta Demonte (eds.): *Gramática descriptiva de la lengua española*, Madrid, Espasa-Calpe: 1025-1128.

Schwenter, Scott A., y Salvador Pons Bordería (2005), «*Por poco (no)*: explicación sincrónica y diacrónica de sus componentes de significado», *Lingüística Española Actual*, 27 (1): 131-158.

Sedano, Mercedes, y Krístel Guirado (2009), «*Compré un poco de libros*: ¿un uso característico del español de Venezuela?», *Verba*, 36: 67-87.

Spitzer, Leo (2007), *Lingua italiana del dialogo*, Milano, il Saggiatore.

Torruella, Joan y Llisterri, Joaquim (1999), «Diseño de corpus textuales y orales», en José Manuel Blecua, Gloria Clavería, Carlos Sánchez y Joan Torruella (eds.), *Filología e informática: nuevas tecnologías en los estudios filológicos*, Barcelona, Milenio: 45-77.

Val.Es.Co. = Pons Bordería, Salvador (dir.): Corpus Val.Es.Co 3.0. <http://www.valesco.es> [Fecha de consulta: 12 de marzo de 2024].

Vigara Tauste, Ana María (1992), *Morfosintaxis del español coloquial: esbozo estilístico*, Madrid, Gredos.

Marta Blanco Domínguez (ORCID: 0000-0002-8635-6950) y María Eugenia Conde Noguerol (ORCID: 0000-0002-5568-7423), «*Un poco* como recurso pragmático-discursivo en corpus orales de nativos y de aprendientes de español: un estudio comparado». *Biblioteca de Babel: Revista de Filología Hispánica*, vol. extra. 2 (2024), pp. 149-169.
**DOI**: 10.15366/bibliotecababel2024.extra2.006 · **ISSNe**: 2695-6349 · Recibido: 16/09/2024 · Aceptado: 28/10/2024

# *Un poco* como recurso pragmático-discursivo en corpus orales de nativos y de aprendientes de español: un estudio comparado

Marta Blanco Domínguez
*Universidade de Santiago de Compostela*
marta.blanco@usc.es

María Eugenia Conde Noguerol
*Universidade da Coruña*
eugenia.noguerol@udc.es

**Resumen**: Esta contribución se enmarca en la explotación de corpus aplicada a la enseñanza de lenguas. Concretamente, su objetivo reside en investigar el empleo que en la lengua oral hacen los aprendientes de español L2/LE respecto al que realizan los hablantes nativos del cuantificador *un poco* como recurso pragmático-discursivo. El corpus ESLORA de español oral (https://eslora.usc.es/) y el corpus Spanish Learner Language Oral Corpora (SPLLOC) (http://www.splloc.soton.ac.uk/index.html) constituyen las fuentes para el análisis empírico. El análisis comparativo de los datos arrojados por los corpus manejados permitirá comprobar las similitudes y diferencias en el uso que de esta partícula hacen los hablantes nativos y no nativos, de las que se podrán extraer algunas consideraciones sobre el tratamiento didáctico del componente pragmalingüístico en el aula de español como segunda lengua o lengua extranjera.

**Palabras clave**: *un poco*, atenuación, lingüística de corpus, español L2/LE, componente pragmalingüístico.

# *Un poco* as a pragmatic-discursive resource in oral corpora of native speakers and learners of Spanish: a comparative study

**Abstract**: This contribution is framed within the framework of corpus exploitation applied to language teaching. Specifically, its aim is to investigate the use of the discourse operator *un poco* as a pragmatic-discursive resource in the spoken language by learners of Spanish as L2/LE compared to native speakers. The ESLORA corpus of spoken

Spanish (https://eslora.usc.es/) and the Spanish Learner Language Oral Corpora (SPLLOC) corpus (http://www.splloc.soton.ac.uk/index.html), constitute the sources for the empirical analysis. The comparative analysis of the data from the corpora will make it possible to verify the similarities and differences in the use of this particle by native and non-native speakers, from which it will be possible to extract some considerations on the didactic treatment of the pragmalinguistic component in the Spanish as a second/foreign language classroom.

**Keywords**: *un poco*, attenuation, corpus linguistics, Spanish as L2/LE, pragmalinguistic component.

## 1. Introducción

La bibliografía atribuye dos valores principales a la partícula *un poco*: i) uno semántico, como cuantificador (*NGLE* 2009) (1) y ii) otro pragmático, como operador discursivo con función atenuadora, que los hablantes pueden emplear fundamentalmente para rebajar la fuerza de lo dicho, expresar su opinión de forma humilde o indicar un valor aproximativo (Fuentes Rodríguez y Alcaide Lara 2002; Fuentes Rodríguez 2009; Mariotttini 2012; Albelda y Briz 2013; Repede 2023, 2024) (2):

(1) Había **un poco** de comida en el frigorífico; La tela es **un poco** áspera; Ha trabajado **un poco** (*NGLE* 2009: § 20.8d).

(2) Es **un poco** latoso; Estoy **un poco** desesperada; Intentamos **un poco** en esta serie promover **un poco** una magnitud de imágenes (Mariottini 2012: 184-185).

Siguiendo a la Academia (*NGLE* 2009: § 19.2a-f, 20.5a, 20.8, 30.4), *un poco* es un *cuantificador débil (o indefinido) evaluativo* que cuantifica grados, esto es, «expresa los diversos grados en los que se predica una propiedad o tiene lugar un proceso», y se caracteriza por «evaluar una cantidad interpretándola como inferior o superior a alguna norma o a alguna expectativa: *poca agua*, *mucho público*. Otras veces, la magnitud se evalúa como adecuada o inadecuada en relación con cierta finalidad que puede expresarse o no». *Un poco* modifica a adjetivos y adverbios que suelen expresar cualidades negativas o tenidas por tales (*Resulta un poco raro*; *Parecía un poco torpe*; *Un poco atolondradamente*) y suele orientar en esa dirección a los que no expresan léxicamente

una valoración negativa (*un poco azul*, *un poco lejos*)[1]; nocionalmente, adquiere muy diversos valores en función de la naturaleza semántica de lo que se cuantifica: frecuencia, intensidad, duración, tiempo o alguna magnitud similar (*Aquí huele un poco a humedad*; *Voy a dormir un poco*, etc.)[2].

Los estudios más específicos sobre *un poco* se centran fundamentalmente en su carácter de operador discursivo y lo describen desde un enfoque pragmalingüístico. Fuentes Rodríguez y Alcaide Lara (2002: 401-404) apuntan que la expresión adverbial *un poco* funciona en español como cuantificador-intensificador escalar (*Ya eres un poco más libre*) y como atenuativo desrealizante; en este último caso se utiliza, por ejemplo, para rebajar la fuerza de lo dicho (*Es un poco patético ver a un ministro echando culpas…*) o para expresar la opinión de forma humilde y de ese modo no dañar la imagen pública del oyente y no imponerse a los interlocutores (*Yo creo que se ha magnificado un poco la importancia de Draskovic…*). Además, estas autoras también hablan de un uso aproximativo, a medio camino entre el cuantificador y el atenuativo (*Ahora Javier Rihoyo que nos introduce un poco en el libro que hoy es noticia*). Por su parte, Fuentes Rodríguez (2009: 336) indica que *un poco* funciona como «operador argumentativo de suficiencia en posición baja de la escala, en dirección ascendente, que puede actuar como "desrealizante" o "atenuador de fuerza argumentativa" para expresar cortesía o delicadeza» (*Aléjate un poco del foco de tensión, que ya tú hiciste tu parte*). Mariottini (2012: 184-185, 195-196) analiza las funciones de *un poco* desde un punto de vista pragmático y señala para este elemento cuatro valores: a) *un poco* con valor eufemístico que, en función de los niveles en los que opere, tiene un valor intensificador (en el nivel lingüístico) o un valor atenuante (en el nivel social) (*Hombre en eso ya empezamos a disentir un poco*); b) *un poco* atenuador de la fuerza ilocutiva de actos de habla directivos, comisivos, asertivos y expresivos (*Dame un poco de dinero*; *¿Se lo envuelvo un poquito?*; *Es un poco latoso*; *Estoy un poco desesperada*); c) *un poco* modal-enunciativo, caracterizado por modificar el registro de una situación interactiva e introduciendo así un grado de confianza que no corresponde a la situación real (En el acto de clausura de un congreso internacional, la presidenta de mesa afirma: «Hay un poco de morriña»), y d) *un poco* aproximativo oracional, equivalente a 'casi, una especie de, en cierta medida' (*Yo no quise hacer una historia…*, *hacer un poco una contrahistoria*), o *un poco* aproximativo conversacional que suele aparecer tras un verbo imperativo exhortativo y en el interior de una secuencia de abertura (*Hoy debuta Nuria. Nuria, cuéntanos*

[1] Para un estudio comparado de las diferencias entre *poco* y *un poco* + adjetivo, véase Sedano (2009).

[2] Conviene señalar que una de las gramáticas pedagógicas de referencia en el ámbito de la enseñanza de español/LE apunta que para matizar en intensidad el uso de un adjetivo o de un adverbio se suele usar, entre otros, el operador *un poco*, como en *Es un poco caro, pero realmente bueno*; *Es precioso, aunque me parece un poco pequeño para lo que necesitamos* (cf. Matte Bon 2001: 71-72).

*un poquito*…). En su modelo para el análisis sociolingüístico y pragmático de la atenuación, Albelda y Briz (2013: 292-293, 306) incluyen entre los procedimientos lingüísticos (tácticas) para la atenuación el uso de los cuantificadores aproximativos como, por ejemplo, *un poco*. Estos autores señalan que la atenuación es una categoría pragmática «en tanto mecanismo estratégico y táctico (por tanto, intencional), que tiene que ver con la efectividad y la eficacia del discurso»; también es una estrategia «puesto que se atenúa, argumentativamente hablando, para lograr el acuerdo o aceptación del otro (incluso, cuando esta sea solo una aceptación social)» y, por último, es un mecanismo retórico «para convencer, lograr un beneficio, persuadir y, a la vez, para cuidar las relaciones interpersonales y sociales o evitar que estas sufran algún tipo de menoscabo». Repede (2023 y 2024) analiza la locución *un poco* a partir de una muestra tomada del corpus PRESEEA-Sevilla para la que documenta dos valores diferentes. Un valor semántico cuando funciona como cuantificador (… *pero hoy me he levantado un poco más tarde*), y un valor pragmático cuando se emplea como operador discursivo para el que identifica cuatro funciones distintas: atenuadora (la más común) (*Soy una persona un poco negativa en verdad ¿eh?*), intensificadora (*¿Te sentiste un poco desprotegido/a a lo mejor? Un poco sí, un poco sí*), aproximadora (… *entonces en eso se resumiría un poco eh lo que te estoy diciendo ¿no?*) y reformuladora (*Una de ellas se hace con vino vi o sea el pan eh un poco asentado un poco duro se remoja*).

## 2. Metodología

Teniendo, pues, en consideración las propuestas descritas en el apartado anterior acerca de los usos y valores del cuantificador *un poco*, en tanto que recurso pragmático-discursivo, este trabajo tiene como objetivo comprobar el uso en contextos orales que de esta partícula hacen los hablantes nativos y compararlo con el de los aprendientes de español. Pensamos que un análisis descriptivo con este enfoque permite identificar cómo los hablantes emplean esta partícula en diversos contextos, y además puede aportar información valiosa sobre el dominio de la competencia pragmática en aprendientes de ELE.

Para ello, la aproximación al contenido gramatical seleccionado se apoya en el uso combinado de dos corpus, ESLORA y SPLLOC, con el fin de contrastar el empleo de *un poco* registrado en hablantes nativos y hablantes no nativos de español. A su vez, consideramos necesario hacer un análisis del tratamiento que otorgan a la atenuación los documentos base para la enseñanza de lenguas extranjeras, con especial atención a la partícula *un poco*.

## *2.1. Descripción de los corpus*

El corpus ESLORA de español oral (https://eslora.usc.es/) nació como parte del macroproyecto PRESEEA. En su versión actual (la 2.3), contiene 60 horas de entrevistas semidirigidas y 20 horas de conversaciones de hablantes de Galicia grabadas entre los años 2007 y 2015, y alcanza 768 053 palabras ortográficas, lematizadas y anotadas morfosintácticamente. Su interfaz de consulta permite obtener datos mediante un sistema de consultas simples y combinadas que incluye variables sociales (grupo de edad, nivel de estudios y sexo) junto a categorías lingüísticas (lemas, clases de palabras y categorías morfológicas). Además, es posible acceder inmediatamente al audio desde la transcripción ya que los registros sonoros se transcribieron ortográficamente con alineación texto-voz.

El recurso SPLLOC está constituido por producciones orales de aprendientes de español/LE con L1 inglés y ha sido realizado en dos fases, SPLLOC1 y SPLLOC2, entre 2008 y 2010. SPLLOC 1 contiene 40 horas de entrevistas semidirigidas y 221 horas de grabaciones en clase con tres grupos de edad. SPLLOC 2 contiene 30 horas de entrevistas semidirigidas y 240 horas de grabaciones de 120 aprendientes de tres grupos de edad y de distintos niveles de dominio de lengua: principiante (de 13 a 15 años), intermedio (de 17 a 18 años) y avanzado (de 20 a 23 años).

Además, se realizó un rastreo del contenido seleccionado en documentos base para la enseñanza del español como lengua extranjera (PCIC 2006; MCER 2002; MCER VC 2020), con la finalidad de comprobar cómo abordan su descripción gramatical.

## *2.2 Procedimiento de análisis*

A tenor de las opciones de búsqueda y recuperación que ofrece la aplicación de consulta de los dos corpus manejados, para llevar a cabo esta investigación se habilitó la opción de búsqueda simple de elementos gramaticales en la que se introdujo la forma *un poco*. Los corpus ESLORA y SPLLOC arrojaron 1 113 y 250 casos, respectivamente, que se filtraron de manera manual, con el fin de prescindir de aquellos casos en que *un poco* funciona como cuantificador en la construcción pseudopartitiva *un poco de* (*Necesito un poco de pan*). Asimismo, dada la elevada cantidad de casos registrados y dado que este es un estudio aproximativo de una muestra de corpus, se circunscribió el análisis a un total de 200 ocurrencias extraídas aleatoriamente, 100 casos por cada uno de los corpus.

El análisis de los usos de *un poco* se ha realizado teniendo en cuenta su uso contextual, lo que, como se verá, nos

ha llevado a distinguir cinco contextos de uso para la forma analizada. Estos contextos permiten identificar distintas nociones de tipo pragmático, como, por ejemplo, la atenuación o la intensificación.

## *2.3.* Un poco *en los documentos base para la enseñanza-aprendizaje de lenguas extranjeras*

La información que se recoge en los documentos base es de carácter diferente debido a que son de distinto tipo: el MCER (2002 y 2020) no es un documento curricular, sino una base común para tomar decisiones metodológicas y desarrollar herramientas de aprendizaje, mientras que el PCIC (2006) es un documento que proporciona información básica y orientaciones generales y prácticas sobre los distintos componentes curriculares: objetivos, contenidos, metodología y evaluación.

De acuerdo con lo anterior, la información que puede aportar el MCER a nuestro trabajo no remite directamente a los usos y valores del elemento *un poco*, sino que está relacionada con un aspecto general que tiene que ver con las competencias comunicativas de la lengua y, dentro de estas, con la competencia pragmática, concretamente con dos de las escalas de los descriptores ilustrativos que se asocian a dos de los componentes de esa competencia: la flexibilidad y la precisión[3].

| **Componentes** | **Escalas de descriptores ilustrativos** | |
|---|---|---|
| | **MCER (2002)** | **MCER VC (2020)** |
| Competencia discursiva (organizar, estructurar y ordenar mensajes) | **Flexibilidad** | = |
| | Turnos de palabra | = |
| | Desarrollo de descripciones y narraciones | Desarrollo temático |
| | Coherencia y cohesión | = |
| Competencia funcional (realizar funciones comunicativas) | **Precisión** | = |
| | Fluidez oral | = |
| Competencia organizativa (secuenciar mensajes) | | |

*Tabla 1: Competencia pragmática: MCER (2002) y MCER VC (2020).*

Algunos de los conceptos clave recogidos en las escalas de flexibilidad y de precisión implican que el usuario active, entre otros mecanismos lingüísticos, el uso de operadores discursivos con el fin de que pueda hacer (*can do*), esto es, llevar a cabo acciones como las siguientes (MCER VC 2020: 153-154 y 156):

[3] Los descriptores ilustrativos indican lo que el usuario y/o aprendiente puede hacer (*can do*) cuando alcanza uno de los niveles comunes de referencia, y su función principal consiste en ayudar a alinear el diseño curricular, la enseñanza y la evaluación. Concretamente, las escalas de flexibilidad y de precisión se refieren a la habilidad de adaptar la lengua aprendida a nuevas situaciones y formular los pensamientos de diferentes maneras, y a la habilidad de describir con exactitud lo que uno quiere expresar, respectivamente (cf. MCER 2002 y 2020).

| | |
|---|---|
| | **Flexibilidad** |
| C1 | — Logra un efecto positivo en unos destinatarios concretos variando de modo efectivo la manera de expresarse.<br>— Modifica la manera de expresarse para transmitir diferentes grados de compromiso o indecisión, seguridad o certidumbre. |
| B1 | — Utiliza una amplia variedad de elementos lingüísticos sencillos con flexibilidad para expresar gran parte de lo que quiere |
| | **Precisión** |
| C1 | — Realiza un uso efectivo de la modalización lingüística para señalar la fuerza de una reclamación, de un argumento o de su postura ante una cuestión. |

*Tabla 2: Escalas de flexibilidad y precisión (MCER VC 2020).*

Frente al MCER (2002, 2020), el PCIC (2006) como documento curricular, sí aporta información concreta sobre los usos y valores de *un poco* que remiten a usos y valores que hemos documentado en los corpus analizados. Esa información se asocia con tres de los cinco componentes en los que se estructura y se presenta en cuatro de los doce inventarios que contempla el documento, en los que se recogen las descripciones del material necesario para realizar las actividades comunicativas que se especifican en las escalas de descriptores de los niveles comunes de referencia del MCER (2002):

| **Componente gramatical** | |
|---|---|
| **Inventario: apartado(s)** | **Descripción de** *un poco* |
| 1. Gramática:<br>1.1. Clases de palabras (clases nominales) | Recurso gramatical |
| **Componente pragmático-discursivo** | |
| **Inventario: apartado(s)** | **Descripción de** *un poco* |
| 2. Funciones:<br>2.2. Modalidad | Recurso pragmático-discursivo |
| 2. Funciones:<br>2.3. Deseos, planes y sentimientos (es decir, la volición y las emociones) | Recurso pragmático-discursivo |
| 2. Funciones:<br>2.4. Influencia en la imagen del otro (amenazar, prohibir, sugerir, etc.) | Recurso pragmático-discursivo |
| 3. Tácticas y estrategias pragmáticas:<br>3.1. Construcción e interpretación del discurso | Recurso pragmático-discursivo |
| **Componente nocional** | |
| **Inventario: apartado(s)** | **Descripción de** *un poco* |
| 4. Nociones:<br>4.1. Nociones generales (cuantitativas y evaluativas) | Recurso semántico-gramatical |

*Tabla 3:* Un poco *en el PCIC (2006)*[4].

[4] Los contenidos recogidos en esta tabla se desarrollan pormenorizadamente en el apéndice que figura al final de este trabajo (cf. Tabla 8).

De acuerdo con lo que se recoge en la Tabla 3, *un poco* se describe como recurso gramatical, recurso pragmático-discursivo y recurso semántico-gramatical (cf. PCIC 2006: 6. Tácticas y estrategias pragmáticas. Introducción; apartado 8. Nociones generales. Introducción):

a. Como recurso gramatical, forma parte de la estructura de la lengua y sigue ciertas reglas de construcción. Concretamente, *un poco* es un cuantificador decreciente y constituye un medio morfológico mediante el cual el hablante atenúa o minimiza total o parcialmente el contenido proposicional de su enunciado: *Lee un poco; He estudiado un poco*.

b. Como recurso pragmático-discursivo, se integra entre las estrategias que los hablantes utilizan para que los mensajes resulten adecuados y eficaces según los destinatarios a los que se dirigen y el contexto en que tienen lugar:

— *Un poco* puede funcionar como un modalizador del discurso mediante el cual el hablante presenta su actitud respecto a sus propios enunciados y a su interlocutor cuando, por ejemplo, expresa opiniones, actitudes o conocimientos (*El ejercicio es un poco difícil*; *¿No conoces un poco el juego?*) y también gustos, deseos o sentimientos (*Últimamente estoy un poco deprimido*).

— Asimismo, el empleo del cuantificador *un poco* puede ser una táctica para atenuar o mitigar la fuerza ilocutiva del acto amenazador para la imagen pública del otro cuando, por ejemplo, se da una orden, se pide permiso, se propone y sugiere o se aconseja y anima (*Baja un poco el volumen, por favor*; *¿Podrías darme un poco más de agua?*; *Sería mejor que te fijaras un poco más*; *Deberías dormir un poco más*; *¡Come un poco más, hombre!*).

c. Como recurso semántico-gramatical, *un poco* soporta algunas de las nociones referidas a conceptos abstractos como espacio, ubicación, cantidad, cualidad o evaluación, que un hablante puede necesitar cualquiera que sea el contexto en el que tenga lugar un acto comunicativo. Por ejemplo, la noción cuantitativa de grado (*Soy un poco tímido*; *Estos vaqueros me quedaban un poco apretados, pero han cedido bastante desde que los compré*), o la noción evaluativa (*Lo veo un poco pequeño*; *Recibimos unas instrucciones un poco confusas*).

Si nos centramos ahora en la descripción que se hace en el PCIC (2006) de *un poco* como recurso pragmático discursivo y la comparamos con la que se formula en la bibliografía comprobamos lo siguiente (cf. infra Tabla 4):

a. Algunas de las funciones señaladas en el Plan Curricular se ajustan a las atribuidas a *un poco* en los estudios específicos. Por

una parte, se indica un valor de modalizador del discurso (asociado a funciones comunicativas) o atenuador (asociado a actos de habla), mediante el cual el hablante minimiza, mitiga o debilita el contenido proposicional de lo enunciado.

La bibliografía presta asimismo atención a otras funciones que no se contemplan en el Plan Curricular: *un poco* como intensificador, mediante el cual el hablante refuerza la verdad de lo expresado; *un poco* como aproximativo o aproximador usado para suavizar la exactitud de una afirmación, introduciendo una dimensión de imprecisión deliberada que se podría considerar clave en ciertos contextos comunicativos; *un poco* como modal-enunciativo, esto es, «aquel que modifica el registro de una situación interactiva introduciendo un grado de confianza que no corresponde a la situación real» (Mariottini 2012: 185); y *un poco* como reformulador empleado para «definir de manera más o menos amplia un concepto o una expresión cuando el hablante considera que pueden ser ambiguos para el oyente» (Repede 2023).

| **PCIC (2006)** | **Bibliografía** |
|---|---|
| **1. Modalizador de discurso**:<br>a. Expresar opiniones, actitudes y conocimientos<br>b. Expresar gustos, deseos y sentimientos.<br>c. Atenuar o mitigar la fuerza ilocutiva del acto amenazador de dar una orden, pedir permiso, proponer y sugerir, aconsejar y animar. | **1. Atenuador** (Fuentes Rodríguez y Alcaide Lara 2002; Fuentes Rodríguez 2009; Mariottini 2012; Albelda y Briz 2013; Repede 2023)<br>De actos de habla asertivos (de opinión y de información), expresivos, directivos y comisivos |
| | **2. Intensificador** (Fuentes Rodríguez y Alcaide Lara 2002; Mariottini 2012; Repede 2023) |
| | **3. Aproximativo** (Fuentes Rodríguez y Alcaide Lara 2002; Mariottini 2012; Albelda y Briz 2013) o **aproximador** (Repede 2023) |
| | **4. Modal-enunciativo** (Mariottini 2012) |
| | **5. Reformulador** (Repede 2023, 2024) |

*Tabla 4: Descripción de las funciones de* un poco *en el PCIC (2006) y en la bibliografía.*

## 3. Resultados

Basándonos en las descripciones que ofrecen el PCIC y los estudios específicos, en el análisis de los datos proporcionados por los corpus ESLORA y SPLLOC hemos distinguido cinco funciones de *un poco* como recurso pragmático-discursivo, a las que hemos añadido una sexta, la de rellenador del discurso, que no se menciona en las fuentes consultadas:

1. Modalizador o atenuador para:

a. Expresar opiniones, actitudes y conocimientos: incluye los usos de *un poco* en enunciados que reflejan juicios personales, evaluaciones y declaraciones de conocimiento.

b. Expresar gustos, deseos y sentimientos: *un poco* se emplea para suavizar o matizar la expresión de preferencias, deseos y emociones, actuando como un moderador en la intensidad del discurso afectivo.

c. Dar una orden, pedir permiso, proponer y sugerir, aconsejar y animar: los usos de *un poco* cumplen la función de mitigar la fuerza ilocutiva de actos directivos y comisivos, disminuyendo el carácter impositivo o autoritario de órdenes, sugerencias, propuestas, consejos y peticiones de permiso.

2. Intensificador: abarca los usos en los que *un poco* no atenúa, sino que, por el contrario, intensifica la afirmación realizada, a veces con carácter eufemístico, reforzando la verdad o veracidad de lo dicho.

3. Aproximativo o aproximador: remite a los usos en los que *un poco* suaviza la exactitud de la información.

4. Modal-enunciativo: da cuenta de los usos en los que *un poco* introduce un grado de cercanía que no se corresponde a la situación en curso.

5. Reformulador: los usos de *un poco* se asocian con operaciones de organización discursiva como, por ejemplo, la precisión, la ampliación del enunciado o la rectificación.

6. Rellenador del discurso: responde a los usos en los que *un poco* actúa como un elemento lingüístico de relleno que le permite al hablante disponer de tiempo para organizar su discurso.

En la tabla 5 se recogen las frecuencias absolutas de *un poco* en los dos corpus orales analizados: el de aprendientes de español (SPLLOC) y el de hablantes nativos (ESLORA).

| Funciones de un poco | SPLLOC | ESLORA |
|---|---|---|
| **1. Modalizador o atenuador** | | |
| a. Expresar opiniones, actitudes y conocimientos | 77 | 39 |
| b. Expresar gustos, deseos y sentimientos | 12 | 10 |
| c. Dar una orden, pedir permiso, proponer y sugerir, aconsejar y animar | 0 | 6 |
| **2. Intensificador** | 1 | 3 |
| **3. Aproximativo o aproximador** | 6 | 36 |
| **4. Modal-enunciativo** | 0 | 0 |
| **5. Reformulador** | 0 | 0 |
| **6. Rellenador del discurso** | 4 | 6 |
| **Total** | **100** | **100** |

*Tabla 5: Frecuencias absolutas de* un poco *en SPLLOC y ESLORA.*

Los datos ponen de manifiesto lo siguiente:

a. El uso de *un poco* como modalizador o atenuador para expresar opiniones, actitudes y conocimientos en SPLLOC es el más frecuente, pues se han documentado 77 de los 100 casos (3a-b). Se observan, por tanto, diferencias cuantitativas con relación a su uso en el corpus ESLORA, donde se localizan tan solo 39 casos con este valor (4a-b):

(3) a. mmm Y por final tengo organizar viajes a España porque es una cosa a mi parece **un poco** eh imposible (SPLLOC).
b. (que) que estaban ahí ah porque Pancho es **un poco** tonto (SPLLOC).

(4) a. Pero no era justa justo la edad exacta como hay ahora que a mí me parece **un poco** apretado para los niños pequeños porque a veces uno madura <pausa/> de distinta forma (ESLORA).
b. Y no dices nada pero fff <pausa/> no <pausa/> la gente anda **un poco** loca a veces con el coche (ESLORA).

Asimismo, comprobamos que en este contexto los aprendientes de español utilizan casi exclusivamente (63 casos) *un poco* + adjetivo en actos comunicativos para expresar una opinión negativa de forma atenuada (5a); frente a esto, solo registramos 14 casos en los que esta partícula atenúe la expresión de conocimientos o informaciones (5b). Por otro lado, los datos de ESLORA también muestran que en los actos asertivos es más frecuente el uso de *un poco* para dar una opinión (27 casos) (6a) que para dar una información (12 casos) (6b), si bien esta distribución está más equilibrada[5].

(5) a. creo que uhm al principio en mi primer año fue **un poco** difícil porque en la escuela los profesores son ingleses y nos hablan la mayoría del tiempo en inglés (SPLLOC).
b. para mí creo que es eso algo que (que) falta **un poco** y es que no tenemos (la) la oportunidad (de) de hablar con nativos sólo con (pro) los profesores y para mí eso sería uh (SPLLOC).

(6) a. eso aquí el tiempo nos lo permite en Irlanda es **un poco** más complicado eso porque el tiempo no lo permite tanto (ESLORA)
b. supongo que el alemán <pausa/> no sabe lo que significa azul en este país <pausa/> ¿no? <pausalarga/> le pasa **un poco** como a la empresa esta <pausa/> Arriva <pausa/> ¿no? <pausa/> que la filial inglesa <pausa/> perdón <pausa/> la filial española que le quería poner Arriva España (ESLORA)

Estas diferencias cuantitativas (77 casos en SPLLOC frente a los 39 de ESLORA) relativas a la atenuación cuando se expresan opiniones, actitudes o conocimientos podrían estar indicando que los aprendientes recurren al empleo de *un poco* posiblemente debido al escaso manejo de otras estructuras gramaticales más complejas diferentes de la estructura del cuantificador *un poco* + adjetivo. En contraste, los hablantes nativos parecen tener un repertorio más amplio de recursos para expresar opiniones, lo que puede explicar su menor frecuencia de uso (cf. Albelda *et al.* 2014).

b. Como modalizador o atenuador para expresar gustos, deseos y sentimientos, observamos una baja frecuencia de uso en ambos corpus (12 casos en SPLLOC (7a-b) y 10 casos en ESLORA (8a-b):

[5]

| Acto de habla asertivo | SPLLOC | ESLORA |
|---|---|---|
| De opinión | 63 | 27 |
| De información | 14 | 12 |
| Total | 77 | 39 |

*Tabla 6: Frecuencias absolutas de* un poco *en actos asertivos*

(7) a. Nada pues había una cosa que (mol) me molestaba **un poco** es que la inmigración es algo (muy) pues no (muy) muy nuevo pero bastante nuevo (SPLLOC).
b. Y (encon) encontró a un perro uh se (ponó) **un poco** nervioso (SPLLOC).

(8) a. A veces <pausa/> sobre todo cuando lo aprietan los primeros días <pausa/> molesta **un poco** <pausa/> ¿sí? (ESLORA).
b. Pues <pausa/> siempre te sientes **un poco** más arropada claro están allí (ESLORA).

c. En el corpus de aprendientes no se localizan casos en los que *un poco* se utilice como modalizador o atenuador para dar órdenes, pedir permiso, proponer y sugerir, aconsejar y animar, que sí se registran, en cambio, en ESLORA en 6 ocasiones (9a-c):

(9) a. ¿Abriré **un poco** la ventana? porque está pillando mucho vaho (ESLORA).
b. Sí sí bueno <pausa/> pues véndeme **un poco** Barcelona (ESLORA).
c. ¿cómo ves **un poco** el futuro así profesional de tu generación? (ESLORA).

d. Respecto a la función de *un poco* como intensificador, los datos aportados por los corpus reflejan una frecuencia de uso muy baja (SPLLOC 1 caso y ESLORA 3 casos), que podría estar indicando que se trata de un uso más específico y, por lo tanto, menos común tanto entre aprendientes como entre hablantes nativos. Este uso que puede ser interpretado como eufemístico (Mariottini 2012: 184) sirve para «maximizar o imprimir mayor fuerza a las acciones y puntos de vista, a la vez que realza el papel o afectación del yo o del yo y el tú, con el fin de lograr la meta prevista» (Briz, 2017: 39, en Repede (2023)) y puede ser sustituido por los adverbios *intensamente*, *fuertemente*, *profundamente*, *vivamente* y otros similares, así como por sus antónimos (*NGLE* 2009: §30.4l):

(10) ehm escuchar y es **un poco** demasiado creo (SPLLOC).

(11) me ha <risa_inicio/> costado <solapado_fin/> **un** <risa_fin/> **poco** pero (ESLORA).

e. Con valor aproximativo o aproximador, *un poco* se aleja de su función primaria de cuantificador que modifica a un sustantivo, un adjetivo, un verbo o un adverbio, y se acerca a la de operador argumentativo que se inserta en secuencias más o menos fijas como *(ser) un*

*poco para*, *(ser) un poco como*, *un poco a*, etc. En estos contextos, *un poco* puede ser sustituido por otros aproximadores como *casi*, *una especie de*, *en cierta medida*, etc., indicando cercanía a un punto que se toma como referencia, preferentemente entre un estado de cosas y su negación (cf. Mariottini 2012: 195). Se trata de un uso apenas documentado en SPLLOC (6 casos) (12a-b) que contrasta con un empleo común en ESLORA (36 caos) (13a-c), lo que pone de manifiesto una diferencia significativa entre ambos grupos de hablantes:

(12) a. también (fui a) fui a Cuba **un poco** para practicar un español un poco diferente (SPLLOC).
b. sí eso es **un poco** um como la mía um (pro) (SPLLOC).

(13) a. nos tiramos una hora hora y pico leyendo <pausa/> y es **un poco** para que coja <pausa/> el hábito (ESLORA).
b. me fui a Inglaterra básicamente a a cam a cambiar de aires a **un poco** a <ruido tipo="chasquido dedos"/> <risa/> a romper sí <pausa/> llevas toda la vida estudiando (ESLORA).
c. que es **un poco** como lo que dice Klemperer <alargamiento>en</alargamiento> en esto de la lengua del Tercer Reich (ESLORA).

f. Como modal-enunciativo y como reformulador, no hemos documentado casos en los corpus consultados, puesto que no se dan contextos que requieran un mayor grado de cercanía entre los interlocutores ni operaciones discursivas de rectificación, respectivamente.

g. El último de los valores que hemos registrado, y del que no hemos encontrado referencias en la bibliografía consultada, es el de *un poco* como un rellenador del discurso u operador periférico. En este contexto, comprobamos que *un poco* se posiciona normalmente al final del discurso o acompaña a otras partículas, expresiones de control de contacto con el interlocutor o fórmulas apelativas (*así*, *¿sabes?*, *¿no?*, *no sé*, *eh*, etc.), por lo que cualquiera de los valores anteriormente mencionados (atenuador, intensificador, aproximativo, etc.) se pierden o difuminan. Así, pasa a funcionar como un rellenador discursivo, esto es, como un elemento lingüístico de relleno que le permite al hablante disponer de tiempo para organizar su discurso. En estos casos, *un poco* no llega a modificar el sentido de la secuencia, aunque puede añadir matices de moderación, inexactitud o imprecisión, y muestra un comportamiento similar a otras partículas (*así*, *y tal*, *no sé (qué)*, *bueno*, etc.) que han sido denominadas en la bibliografía como expresiones debilitadoras del significado y minimizadoras de la intención (Albelda *et al.* 2014). Se trata de un uso poco documentado en los corpus (SPLLOC 4 casos y ESLORA 6 casos):

(14) estas cosas que te te suponen así **un poco** ¿sabes? (ESLORA).

(15) eso no me gusta nada (pero) si están solamente persiguiéndole me parece no sé es **un poco** pues no sé es una tema bastante polémico eh (SPLLOC).

Como hemos podido constatar hasta ahora, aunque existen diferencias cuantitativas significativas entre los dos corpus, tanto los aprendientes como los hablantes nativos emplean *un poco* como recurso pragmático-discursivo con distintos valores (cf. Tabla 7). Ahora bien, los datos referidos a los hablantes nativos muestran mayoritariamente un uso equilibrado del valor aproximador (36 % de los casos) y del atenuador para expresar opiniones, actitudes y conocimientos (39 % de los casos), siendo este último empleo el más frecuente entre los aprendientes (77 % de los casos).

| PCIC (2006) | Bibliografía | Otros usos registrados | SPLLOC | ESLORA |
|---|---|---|---|---|
| **1. Modalizador del discurso** | **1. Atenuador** | | | |
| a. Expresar opiniones, actitudes y conocimientos | a. Atenuador de actos de habla asertivos de opinión y de información | | ✓ | ✓ |
| b. Expresar gustos, deseos y sentimientos | b. Atenuador de actos de habla expresivos | | ✓ | ✓ |
| c. Dar una orden, pedir permiso, proponer y sugerir, aconsejar y animar | c. Atenuador de actos de habla directivos | | | ✓ |
| | d. Atenuador de actos de habla comisivos | | | |
| | **2. Intensificador** | | ✓ | ✓ |
| | **3. Aproximativo o aproximador** | | ✓ | ✓ |
| | **4. Modal-enunciativo** | | | |
| | **5. Reformulador** | | | |
| | | **7. Rellenador del discurso** | ✓ | ✓ |

*Tabla 7: Usos de* un poco *en los corpus y en las fuentes consultadas*

## 5. Conclusiones

El examen basado en corpus del uso oral contextualizado de *un poco* nos permitió observar su productividad de forma comparativa entre aprendientes de español y hablantes nativos. Estos últimos muestran un uso equilibrado de las distintas funciones de *un poco*, ya que lo emplean frecuentemente como atenuador y aproximador y también, aunque en menor medida, como intensificador y rellenador del discurso, lo que refleja un dominio sólido en el empleo de estrategias pragmáticas de la oralidad, así como una capacidad para usar la lengua de manera flexible. Frente a esto, observamos que los aprendientes parecen tener un conocimiento más limitado de la multifuncionalidad de *un poco*, y los usos que hacen de esta partícula remiten fundamentalmente a la función de atenuador, junto a otras más residuales como la de aproximador y rellenador. Esto podría deberse a diversos factores relacionados con el proceso de enseñanza-aprendizaje del español como lengua extranjera como, por ejemplo, una exposición insuficiente al habla conversacional o una falta de instrucción explícita del componente pragmático-discursivo.

Consideramos que, a pesar de los avances que han experimentado los estudios de la pragmática de la interlengua en español/LE, sería necesario, de una parte, ampliar las áreas de investigación, limitadas fundamentalmente a categorías como los actos de habla, los implícitos y la deixis, y, de otra, seguir desarrollando materiales de enseñanza-aprendizaje que incorporen contenidos pragmáticos y que los ilustren con muestras reales de lengua (cf. Koike y Pearson 2019; Pearson y Hasler-Baker 2021). De esta manera se introducirían nuevos conocimientos empíricos en la práctica de la enseñanza de lenguas, que responderían así de forma más adecuada a las condiciones y exigencias comunicativas reales.

Debido a su carácter exploratorio, las conclusiones que se desprenden del análisis realizado deben ser vistas con cautela. En futuros estudios convendría ampliar el análisis de *un poco* a una muestra mayor, atendiendo además a variables extralingüísticas relativas al tipo de discurso y a los perfiles sociolingüísticos de los informantes, así como a su alteración morfológica *un poquito*.

### bibliografía

Albelda, Marta, y Antonio Briz (2013), «Una propuesta teórica y metodológica para el análisis de la atenuación lingüística en español

y portugués: la base de un proyecto en común (es.por.atenuación)», *Onomázein*, (28): 288-319. DOI: 10.7764/onomazein.28.21.

Albelda, Marta, Antonio Briz, Ana M.ª Cestero, Dorota Kotwica, y Cristina Villalba (2014), «Ficha metodológica para el análisis pragmático de la atenuación en corpus discursivos del español (es.por.atenuación)», *Oralia*, (17): 7-62. DOI: https://doi.org/10.25115/oralia.v17i1.

ESLORA = Corpus para el estudio del español oral. https://eslora.usc.es/.

Fuentes Rodríguez, Catalina (2009), *Diccionario de conectores y operadores del español*, Madrid, Arco/Libros.

Fuentes Rodríguez, Catalina, y Esperanza R. Alcaide Lara (2002), *Mecanismos lingüísticos de la persuasión*, Madrid, Arco/Libros.

Koike, Dale A., y Lynn Pearson (2019), «Pragmática (Pragmatics)», en Javier Muñoz-Basols, Elisa Gironzetti y Manel Lacorte (eds.), *The Routledge handbook of Spanish language teaching: metodologías, contextos y recursos para la enseñanza del español L2*, Londres/Nueva York, Routledge: 384-361.

Mariottini, Laura (2012), «Modalidad y atenuación: análisis de un poco y de sus alternaciones morfológicas en las conversaciones coloquiales», *Oralia*, (15): 177-203. DOI: https://doi.org/10.25115/oralia.v15i1.8061.

Matte Bon, Francisco (2001), *Gramática comunicativa del español. Tomo 2 De la idea a la lengua*, (nueva edición revisada), Madrid, Edelsa.

MCER = Consejo de Europa (2002), *Marco Común Europeo de Referencia para las Lenguas: Aprendizaje, Enseñanza, Evaluación, Madrid, MECD y Anaya*. Disponible en: https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/default.htm.

MCER VC = Consejo de Europa (2020), *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación. Volumen complementario*, Estrasburgo, Servicio de Publicaciones del Consejo de Europa. Disponible en: https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco_complementario/default.htm.

NGLE = RAE y ASALE (2009), *Nueva gramática de la lengua española*, Madrid, Espasa Libros.

PCIC = Instituto Cervantes (2006), *Plan curricular del Instituto Cervantes. Niveles de referencia para el español*, Madrid, Instituto Cervantes, Biblioteca Nueva. Disponible en:

https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/default.htm.

Pearson Lynn y Maria Hasler-Barker (2021), «Second language acquisition of Spanish pragmatics», en Dale A. Koike y J. César Félix-Brasdefer (eds.), *The Routledge handbook of Spanish pragmatics: foundations and interfaces*, Londres/Nueva York, Routledge: 423-439.

PRESEEA = *Proyecto para el Estudio Sociolingüístico del Español de España y América*. Disponible en: http://preseea.linguas.net/.

Repede, Doina (2023), «La locución *un poco* en el corpus oral PRESEEA-Sevilla: funciones discursivas y distribución social», *Forma y Función*, 36 (1). DOI: https://doi.org/10.15446/fyf.v36n1.97379.

Repede, Doina (2024), «Análisis sociopragmático de *un poco* en las entrevistas semidirigidas». *Onomázein*, 63: 01-19. DOI: https://doi.org/10.7764/onomazein.63.01.

Sedano, Mercedes (2009), «*Poco*/*Un poco* + adjetivo: diferencias semánticas y consecuencias distribucionales». *Núcleo*, 21 (26): 151-179. Disponible: https://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0798-97842009000100006&lng=es&nrm=i&tlng=es.

SPLLOC = Spanish Learner Language Oral Corpora. https://web-archive.southampton.ac.uk/www.splloc.soton.ac.uk/.

# APÉNDICE. TABLA 8: *UN POCO* EN EL PCIC (2006)

| **1. Componente gramatical** | |
|---|---|
| **Inventario: apartado(s)** | **Descripción de *un poco*** |
| 1. Gramática:<br>1.1. Clases de palabras (clases nominales) | Recurso gramatical: cuantificador decreciente (atenuación/minimización)<br>Nivel B1<br>*Lee* ***un poco***; *He estudiado* ***un poco***. |
| 1. Gramática:<br>1.2. Sintagma | - |
| 1. Gramática:<br>1.3. Oración | - |
| **2. Componente pragmático-discursivo** | |
| **Inventario: apartado(s)** | **Descripción de *un poco*** |
| 2. Funciones:<br>2.1. Información | - |
| 2. Funciones:<br>2.2. Modalidad | Recurso pragmático-discursivo: expresar opiniones, actitudes y conocimientos<br>Nivel A1<br>2. Expresar opiniones, actitudes y conocimientos. 2.4. Valorar: *Es (un poco)* + adj: *El ejercicio es* ***un poco*** *difícil*<br>Nivel A2<br>2. Expresar opiniones, actitudes y conocimientos. 2.19. Preguntar por el conocimiento de algo: *¿Conoces* (+ *un poco*) + SN?: *¿No conoces* ***un poco*** *el juego*? 2.20. Expresar conocimiento: *Conozco un poco* + SN: *Conozco* ***un poco*** *el país.* |
| 2. Funciones:<br>2.3. Deseos, planes y sentimientos (es decir, la volición y las emociones) | Recurso pragmático-discursivo: expresar gustos, deseos y sentimientos<br>Nivel B1<br>3. Expresar gustos, deseos y sentimientos. 3.13. Expresar tristeza y aflicción: *Estoy* (+cuantif.) + *triste/deprimido/mal*...: *Últimamente estoy* ***un poco*** *deprimido*. 3.18. Expresar miedo, ansiedad y preocupación: *Estoy* (+cuantif.) + *preocupado/asustado*...: *Estaban* ***un poco*** *asustadas*. 3.19. Expresar nerviosismo: *Estoy* (+cuantif.) + *estresado/histérico*...: *Estamos* ***un poco*** *nerviosos*. 3.30. Expresar sensaciones físicas: *Me duele* +cuantif.+SN: *Les duele* ***un poco*** *la garganta.* |

| | |
|---|---|
| 2. Funciones:<br>2.4. Influencia en la imagen del otro (amenazar, prohibir, sugerir, etc.) | Recurso pragmático-discursivo: influir en el interlocutor (dar una orden y pedir objetos de forma atenuada; pedir permiso, proponer o sugerir, aconsejar y animar)<br>Nivel B1<br>4. Influir en el interlocutor. 4.1. Dar una orden o instrucción de forma atenuada: Imperativo + atenuador: *Baja* ***un poco*** *el volumen, por favor*. *¿Podrías* + inf.*? ¿Podría bajar* ***un poco*** *la radio, por favor?* 4.3. Pedir objetos de forma atenuada: *¿Puedes / Podrías darme / dejarme / prestarme / pasarme / traer(me)* + SN*? ¿Podrías darme* ***un poco*** *más de agua?*. 4.8. Pedir permiso: *¿Te molesta/Te importa*...+ *si* + pres. indic.*?*: *¿Os molesta que abra* ***un poco*** *la ventana? Hace mucho calor*. 4.13. Proponer o sugerir: *Podrías/Podríamos* +inf.: *Podrías pensarlo* ***un poco*** *más. Es una decisión difícil*. 4.18. Aconsejar: *Podrías* +inf: *Podrías trabajar* ***un poco*** *menos*; *Debes/Deberías* + inf.: *Deberías dormir* **un poco** *más*. 4.25. Animar: Imperativo (*hombre/mujer*): *¡Come* **un poco** *más, hombre!*<br>Nivel B2<br>4. Influir en el interlocutor. 4.18. Aconsejar: *Sería mejor/necesario/importante*...: *Sería mejor que te fijaras* ***un poco*** *más*<br>Nivel C1<br>4. Influir en el interlocutor. 4.13. Proponer o sugerir: *Te sugiero*...+ *que*...: T*e sugiero que instales un programa* ***un poco*** *más moderno*<br>Nivel C2<br>4. Influir en el interlocutor. 4.13. Proponer o sugerir: *No estaría de más que*..: *No estaría de más que limpiaras esto* ***un poco*** *de vez en cuando*. 4.18. Aconsejar: (*Yo que tú / Yo en tu lugar*) +imperf. indic. *Yo que tú se lo decía. También podías trabajar* ***un poco*** *menos.* |
| 2. Funciones:<br>2.5. Relaciones sociales y cortesía formulaica o convencional | - |
| 2. Funciones:<br>2.6. Forma de estructurar el discurso | - |
| 3. Tácticas y estrategias pragmáticas:<br>3.1. Construcción e interpretación del discurso | Recurso pragmático-discursivo: valores ilocutivos de los enunciados interrogativos (petición / permiso)<br>Nivel B1<br>1.6. Valores ilocutivos de los enunciados interrogativos. 1.6.1. Interrogativos neutros Petición Permiso: *¿Te importa / importaría si abro* ***un poco*** *la ventana?*<br>Nivel C2<br>1.8. Significados interpretados. 1.8.2. Indicadores de la ironía: Interrogativas retóricas que expresan una queja o reproche: *¿No podías haber encontrado uno* ***un poco*** *peor?* |

| | |
|---|---|
| 3. Tácticas y estrategias pragmáticas:<br>3.2. Modalización | - |
| 3. Tácticas y estrategias pragmáticas:<br>3.3. Conducta interaccional | - |
| 4. Géneros discursivos y productos textuales:<br>4.1. Géneros orales y escritos | - |
| 4. Géneros discursivos y productos textuales:<br>4.2. Muestras de género | - |
| 4. Géneros discursivos y productos textuales:<br>4.3. Macrofunciones (descriptiva, narrativa y expositiva) | - |
| **3. Componente nocional** | |
| **Inventario: apartado(s)** | **Descripción de *un poco*** |
| 5. Nociones.<br>5.1. Nociones generales (cuantitativas y evaluativas) | Recurso semántico-gramatical: noción cuantitativa de grado y noción evaluativa<br>Nivel A1<br>2. Nociones cuantitativas. 2.5. Grado: *mucho, bastante, un poco*: *Me gusta mucho. Soy* ***un poco*** *tímido*<br>Nivel B1<br>6. Nociones evaluativas. 6.1. Evaluación general: *encontrar* (*algo*), *ver* (*algo*): *Lo veo* ***un poco*** *pequeño*<br>Nivel C1<br>6. Nociones evaluativas. 6.16. Facilidad: *(in)asequible, (in)comprensible, confuso, problemático*: *Recibimos unas instrucciones* ***un poco*** *confusas*.<br>Nivel C2<br>2. Nociones cuantitativas. 2.6. Medidas. 2.6.3. Tamaño: *menguar, ceder, dar(se) de sí*: *Estos vaqueros me quedaban* ***un poco*** *apretados, pero han cedido bastante desde que los compré* |
| 5. Nociones.<br>5.2. Nociones específicas | - |
| **4. Componente cultural** | |
| **5. Componente de aprendizaje** | |

Florencio del Barrio de la Rosa(ORCID: 0000-0002-6008-1004), «*Ni tan mal*: un «operador argumentativo» en el español del siglo xxi». *Biblioteca de Babel: Revista de Filología Hispánica*, vol. extra. 2 (2024), pp. 171-194.
**DOI**: 10.15366/bibliotecababel2024.extra2.007 · **ISSNe**: 2695-6349 · Recibido: 01/10/2024 · Aceptado: 21/10/2024

# *Ni tan mal*: un «operador argumentativo» en el español del siglo xxi

Florencio del Barrio de la Rosa
*Università Ca' Foscari Venezia*
fbarrio@unive.it

**Resumen**: La presente investigación se ocupa de la construcción *ni tan mal* en el español europeo del siglo xxi y la describe como un operador argumentativo. Este operador invierte la orientación discursiva de la argumentación y, consideradas las circunstancias, presenta la conclusión como la mejor opción posible. En este movimiento argumentativo, el denominado *complemento de proporción* (*para el precio, la ubicación no está ni tan mal*) desempeña una función crucial. Además de la caracterización gramatical y discursiva de *ni tan mal*, el trabajo defiende la transformación incipiente de esta construcción en un marcador conversacional y la relaciona con otros operadores modales de afirmación y negación descritos en estudios recientes. Los datos están extraídos de corpus representativos del registro coloquial escrito y se analizan dentro del marco de la sintaxis discursiva o tética y la Teoría de la Argumentación.

**Palabras clave**: *ni tan mal*, operador argumentativo, inversor argumentativo, suficiencia argumentativa, sintaxis tética, cooptación, español del siglo xxi.

## *Ni tan mal*: an «argumentative operator» in 21st century Spanish

**Abstract**: The present research focuses on the construction *ni tan mal* in 21st century European Spanish and describes it as an argumentative operator. This operator reverses the orientation of the argumentation and, given the circumstances, presents the conclusion as the best possible option. In this argumentative movement, the so-called «complemento de proporción» (*for*-complement) (*Para el precio, el hotel no está ni tan mal* 'for the price, the hotel isn't so bad') plays a crucial role. In addition to the grammatical and discursive characterization of *ni tan mal*, the study supports its emerging transformation into a conversational marker and relates it to other modal affirmation and negation

operators described in recent studies. Data are drawn from corpus representative of the written colloquial register and are analyzed within the framework of Thetical Grammar and Argumentation Theory.

**Keywords**: *ni tan mal*, argumentative operator, argumentative inversion, argumentative sufficieny, thetical grammar, cooptation, 21st century Spanish.

## 1. Presentación de *ni tan mal*

En este artículo me ocuparé de *ni tan mal* (a partir de ahora: NTM), un «operador en proceso» (Fuentes Rodríguez 2020) en el español del siglo XXI, con el objetivo principal de ofrecer una descripción preliminar de sus características gramaticales y de sus funciones discursivas. En concreto, defenderé la hipótesis de que NTM funciona como un *operador argumentativo*[1] que refuerza un argumento frágil y lo reorienta para presentar la conclusión como la mejor opción posible consideradas las circunstancias. De acuerdo con esta hipótesis, NTM formaría parte de la clase de los «inversores argumentativos» (Portolés 2016), categoría, por lo demás, poco investigada en los estudios sobre marcación discursiva en español. En el primer ejemplo con el que ilustramos las funciones argumentativas de NTM (1)[2], el resultado obtenido se manifiesta, vistas las complicadas condiciones en las que el golfista ha de golpear la bola, como algo inesperado. En el microdiálogo reproducido en (2), por su parte, se advierte cómo los argumentos puestos en juego (<la lluvia> y <el cielo nublado>) se coorientan hacia la decisión de no ir a la playa. Considerando que la lluvia supone un escenario meteorológico probable en las localidades marineras del norte de España, el operador reorienta el argumento débil <estar nublado> —la menos mala de las opciones— hacia la conclusión de pasar un buen día de playa.

[1] En el marco de la Teoría de la Argumentación (Anscombre y Ducrot 1994; Portolés 1998a, 1998b), se define operador argumentativo como el elemento con capacidad para modificar el potencial argumentativo de un enunciado:

> Un morphème X est un opérateur argumentatif s'il y a au moins une phrase P telle que l'introduction de X dans P produit une phrase P', dont le potentiel d'utilisation argumentative est différent de celui de P, cette différence ne pouvant pas se déduire de la différence entre la valeur informative des énoncés de P et de P' (Ducrot 1983: 10).

Para los propósitos actuales, adoptaré los términos de *operador* y de *marcador conversacional* en su definición primaria (Martín Zorraquino y Portolés 1999). En este sentido, cualquier otra denominación (*partícula discursiva*, *marcador discursivo*, *operador pragmático* y otras semejantes) ha de entenderse, a falta de indicación expresa, como variantes estilísticas o abarcadoras.

[2] Los ejemplos se reproducen respetando la grafía original. En el caso de las conversaciones de Whatsapp empleo iniciales para sustituir los nombres de los participantes y marco con un número el orden de su intervención. Además del nombre del corpus, ofrezco la fuente y la fecha.

(1) He tenido opciones en los siguientes hoyos y en el 7 he ido justo de palo, he jugado agresivo y se me ha quedado junto a un árbol, la he tenido que jugar a zurdas así que un bogey **ni tan mal** (EsTenTen18, madridiario.es, 28/06/2014).

(2) H1: A cudillero!!  Ayer espectacular Y playita de espasa q estaba hasta arriba Hoy muy nublado y cudillero  .
M1: Diafrutaddd Q aunq haga nublado sino llueve **ni tan mal** (Whatsapp, 27/07/2023).

En ambos ejemplos, NTM expresa una conclusión inesperada e imprevista en la medida en que no concuerda con las premisas discursivas de las que se parte. La anteposición de *pero*[3], en un movimiento discursivo concesivo-adversativo que con frecuencia envuelve la aparición de nuestra partícula, resalta la naturaleza inopinada y, de algún modo, sorpresiva de la conclusión, como se nota en (3). En este fragmento, las fórmulas apelativas (*oye*) y confirmativas (*¿eh?*) sugieren un discurso polifónico y anticipan el empleo conversacional del operador.

(3) Pero nada, mucho ruido y pocas nueces, y el 22 de diciembre me pilló con un búnker construido, la nevera vacía, la cuenta en números rojos y la casa por barrer. Pero oye, **ni tan mal**, ¿eh? (CORPES, S. Villegas Saurí, *Marketingdencias*, 2014).

Efectivamente, el español contemporáneo brinda la ocasión de examinar la transformación de esta unidad, que se desempeña esencialmente en el nivel monológico, en un «marcador conversacional» (Martín Zorraquino y Portolés 1999) operativo en el diálogo. Este salto sería equiparable en buena medida al trazado por *bien* —adverbio léxico valorativo, originariamente— en su función de «operador modal» (Fuentes Rodríguez 2009: *s.v. bien*$_4$, *DPDE s.v. bien*$_3$). En este recorrido, NTM se integraría en un paradigma de operadores —encabezados igualmente por la partícula *ni*— de afirmación (*ni que decir tiene*) (Fuentes Rodríguez 2021) y de negación (*ni soñarlo*, *ni loco*, *ni de broma*) (Padilla Herrada 2023). La diferencia con el resto de miembros (re)afirmativos (Brenes 2020) del paradigma estriba en que NTM no refuerza tanto lo dicho cuanto lo que el hablante supone implicado por el enunciado de su interlocutor. En la conversación digital de (4) el participante C ratifica, en su segunda intervención, la elección de Santander como destino para una estancia breve. Frente a las muestras de (1-3), en este

[3] La relación entre *pero* y la introducción de expectativas contrarias a las premisas se desarrolla en Rodríguez Rosique (2023). El estudio de NTM merecerá, sin duda, explorarse desde la perspectiva de la miratividad (categoría gramatical de lo inesperado), senda que no recorreré —por sugerente que resulte— en estas páginas.

ejemplo no se aportan indicios que impidan considerar la capital cántabra como una alternativa óptima o que relativicen su elección como meta turística. En su intervención, C valora positivamente la decisión de su interlocutor y corrobora un contenido no expreso y, en principio, argumentativamente neutro. Cabría, por tanto, en línea con los estudios mencionados, calificar NTM de operador modal confirmativo.

(4) C1: Te animas el 17 o no? Dime hoy si eso
J1: No, no puedo, muchas gracias. Pero el proximo fin de semana nos vamos a santander.
C2: **Ni tan mal** Santander (Whatsapp, 28/11/2023).

Para cumplir con los objetivos mencionados en esta breve presentación, organizo el trabajo de acuerdo con el siguiente guion. En primer lugar, presento el corpus empírico (§ 2) elaborado a partir de las ocurrencias de NTM en bancos de datos representativos del «registro coloquial escrito». En segundo lugar, me detendré en la caracterización gramatical de este elemento (§ 3) como unidad tética y en su descripción de operador de inversión argumentativa (§ 4). Cerraré el estudio con una sintética sección de conclusiones (§ 5) y las obligadas referencias bibliográficas.

## 2. Distribución de *ni tan mal* en el «registro coloquial escrito»

Para componer el conjunto de datos objeto de análisis, se han realizado búsquedas en distintos corpus electrónicos sincrónicos de lengua hablada, coloquial y subestándar. Además, he elaborado un corpus *ad hoc* de conversaciones digitales (Whatsapp) recopiladas gracias a las contribuciones voluntarias de amigos y parientes (dejo constancia aquí de mi agradecimiento sincero). Salvo este último corpus de confección artesanal, el resto está compuesto por un contingente léxico inmenso; sin embargo, a pesar de contar con el apoyo de los *big data*, el conjunto de datos apenas sobrepasa los dos centenares (en concreto: 228). El volumen exiguo de ocurrencias del operador evidencia, sin duda, las fases iniciales de su difusión, así como el estado embrionario de su polifuncionalidad[4].

[4] Cianca y Gavilanes (2018) incluyen la partícula entre las voces empleadas por los jóvenes madrileños. En otro tenor, la Real Academia, a través de sus cuentas electrónicas, ha respondido a las dudas de los usuarios acerca del empleo de NTM a partir de 2019. Agradezco a uno de los revisores anónimos del trabajo que me haya llamado la atención sobre estas referencias que inciden en el carácter reciente e innovador de la expresión. En esta fase de la investigación, las búsquedas se limitan al español europeo, si bien futuros trabajos habrán de extender el estudio a las variedades americanas.

Las ocurrencias de NTM como operador argumentativo se obtienen de corpus representativos del denominado «registro coloquial escrito» (Felíu y Pato 2019), favorecedor de la emergencia de rasgos no estándares. La partícula NTM se documenta en géneros digitales monológicos (páginas electrónicas, blogs) o interactivos (foros, conversaciones electrónicas) y parece difundirse no antes de 2010. En concreto, las documentaciones proceden de:

a) Corpus del español: las búsquedas se limitan al subcorpus Web/Dialects (www.corpusdelespañol.org) restringiendo las muestras al español europeo. El volumen léxico concerniente a esta variedad roza el medio billón de palabras y proviene de textos fechados a partir de 2015. Este corpus arroja 41 casos de la secuencia <ni tan mal> de los cuales 33 corresponden al marcador discursivo objeto de estudio (en una proporción de 0,8).

b) Corpus del español del siglo XXI (CORPES): el corpus académico (www.rae.es/banco-de-datos/corpes-xxi) recoge muestras de habla datadas entre 2000 y 2024 de todos los países hispanohablantes con una amplia variedad de ámbitos temáticos y textuales. La búsqueda, circunscrita a España, devuelve nueve resultados de <ni tan mal>, de los cuales siete cumplen la función de marcador discursivo, localizados, sobre todo, en obras de divulgación sobre temas de la vida cotidiana o las ciencias sociales y en reportajes de la prensa digital. No se encuentra ningún caso en el nivel dialógico (ni siquiera en el diálogo novelado).

c) EsTenTen18: subcorpus de la plataforma Sketchengine (Kilgarriff y Renau 2013) con más de 15 billones de palabras recogidas de una heterogénea tipología de textos sacados de Internet entre 2011 y 2018. La consulta se limita a las páginas electrónicas con dominio .es. De una muestra aleatoria de 1017 registros de <ni tan mal> 156 se ajustan a nuestra partícula (proporción: 0,15).

d) Whatsapp: Se trata de un corpus de conversaciones digitales elaborado *ad hoc* para el presente estudio[5]. La fecha de los intercambios abarca desde finales del año 2019 hasta finales de 2023. Este corpus ofrece el más alto porcentaje de apariciones de NTM en el nivel dialógico. Por más que se trate de diálogos escritos, este dato apunta a la extensión del marcador discursivo en la lengua conversacional cotidiana.

[5] En este subcorpus no es posible establecer una proporción entre *ni tan mal* como operador y *ni tan mal* como sintagma libre, ya que todos los casos corresponden a la partícula discursiva. El recurso a las conversaciones digitales —«textos escritos oralizados» (Yus 2020)— recogidas a través de la aplicación de mensajería instantánea Whatsapp está cada vez más difundido en los estudios —sobre todo— de marcadores pragmáticos, como puede verse, entre otros trabajos recientes, en Padilla Herrada (2023).

| Corpus | Oración | Discurso | Monólogo | Diálogo | Total |
|---|---|---|---|---|---|
| Corpus del español | 7 (21,2 %) | 26 (78,8%) | 32 (97,0%) | 1 (3,0%) | **33** |
| CORPES | 1 (14,3%) | 6 (85,7%) | 7 (100,0%) | 0 | **7** |
| EsTen-Ten18 | 13 (8,3%) | 143 (91,7%) | 153 (98,1%) | 3 (1,9%) | **156** |
| Whatsapp | 2 (6,3%) | 30 (93,7%) | 19 (59,4%) | 13 (40,6%) | **32** |
| **Total** | **23 (10,5%)** | **204 (89,5%)** | **211 (92,5%)** | **17 (7,5%)** | **228** |

*Tabla 1. Distribución de* NTM *por corpus, sintaxis y nivel conversacional.*

En la tabla 1 expongo de manera sintetizada los datos clasificándolos por la estructura sintáctica que acoge el operador discursivo (oración vs. discurso) y por el nivel conversacional (monólogo vs. diálogo).

Para ampliar el conjunto de datos, he consultado además otros corpus de lengua hablada como el del Proyecto de Estudio Sociolingüístico del Español de España y América (PRESEEA: preseea.uah.es/corpus-preseea), el Corpus Oral y Sonoro del Español Rural (COSER: corpusrural.es) y, por último, el Corpus Oral de Lenguaje Adolescente (COLA: clarino.uib.no/korpuskel). En los corpus orales del español no se registra ningún caso de la secuencia <ni tan mal> en función de marcador pragmático. El resultado negativo de esta consulta corrobora que NTM se halla en las fases inaugurales de su expansión y sugiere una difusión determinada por el tenor vernáculo de los registros digitales. En estos corpus orales se documenta únicamente el sintagma de grado *tan mal* en construcciones comparativas elípticas y se destaca su aparición dependiente de predicados negativos. La estructura oracional <*no* + verbo + *tan mal*> ilustrada en los fragmentos siguientes ha de tomarse como el punto de arranque de nuestro operador (cf. § 3.2 abajo).

(6) Pasas las dos manos pertinentes y cubre la pintura y eso tampoco va a quedar **tan mal**… claro, claro, claro, pero bueno, siempre es, probar (PRESEEA, Las Palmas de Gran Canaria, 14/08/2007).

(7) Pues a mí no me parecían **tan mal** las novias. Con los zapatos un poco altos, negros… y las medias… (COSER, 1902, Canredondo (Guadalajara), 10/05/2003).

(8) De cinco a ocho estás tirado en la estación no tienes nada mejor que hacer y no se duerme **tan mal** o sea no yo también estuve en Granada he dormido varias veces en la estación de tren (COLA, Madrid, 2002-2007).

## 3. Caracterización gramatical de *ni tan mal*

### *3.1. Ntm como unidad tética*

Dentro del esquema argumentativo prototípico <argumento(s) – base argumentativa – conclusión> (cf., entre otros, Fuentes Rodríguez 2003), el operador ntm, en el que persiste la semántica léxica del adverbio primitivo, expresa el resultado final del razonamiento en consonancia con el contenido asertivo propio de los marcadores de modalidad deóntica (Martín Zorraquino y Portolés 1999: 4161). El estatuto de consecuente ejercido por ntm queda explicitado por conectores consecutivos como *así que* (9) y por comentadores como *pues* (10), dos de los introductores habituales[6] de nuestra construcción.

(9) Según he entendido por el texto, es como un «complemento» a las cartas de evento, y la calidad del dibujo (de la muestra al menos) es bastante buena, así que, **ni tan mal** (EsTenTen18, darkstone.es, 24/01/2018).

(10) Al final, consiguió que le devolviesen la obra y recuperó su burrito querido. Pues **ni tan mal** (EsTenTen18, Telemadrid.es, 26/06/2017).

Desde la perspectiva de la estructura informativa, la información se distribuye según una ordenación bimembre de tópico-comentario, en la que nuestra fórmula aporta aquello que se dice o concluye sobre un miembro discursivo. Para este tipo de ordenación, ntm no requiere de elemento introductorio o anclaje discursivo explícito, tal y como pone de manifiesto el microtexto de (11). En su articulación más elemental y escueta —no faltan, como sugiere alguna de las muestras de habla expuestas hasta ahora, estructuras más complejas—, la información se reparte en un esquema semejante al reflejado en la figura 1. En este reparto, ntm realiza, sin necesidad de conectores, la función de comentario y evidencia, prueba de la autonomía sintáctica que va adquiriendo, la naturaleza extrapredicativa y periférica de su misión discursiva.

(11) Porque si llego allí y me encuentro que es un perranco enorme me da un infarto con el miedo que los tengo...Unos gatitos **ni tan mal** (EsTenTen18, traviajar.es, 01/04/2016).

[6] El conector *así que* y el comentador *pues* aparecen en más de la mitad de los casos (56/96) en los que un elemento explícito ancla ntm al discurso precedente. Se documentan otros conectores conclusivos cuales *o sea que*: «Con un uso que tampoco voy a decir que sea exhaustivo sino más tirando a moderado, pero con rato para todo - o sea que ni tan mal» (EsTenTen18, miui.es, sin fecha). Otros conectores frecuentes son los polifuncionales *y* (16 veces), el marcador *bueno* (7 veces) y *que* (6 veces).

| Tópico | Comentario |
|---|---|
| unos gatitos | ni tan mal |

*Figura 1. Distribución de las funciones informativas con* ntm *(ejemplo 11).*

Como prueba adicional de la independencia sintáctica de ntm, se añade su autonomía prosódica. Efectivamente, la locución se caracteriza por un contorno entonativo propio y por aparecer entre pausas, símbolo de su emancipación respecto del contexto discursivo. Incluso en sus ocurrencias como constituyente oracional de predicados verbales (12) es posible intuir una pausa de separación: «no estamos | ni tan mal»[7].

(12) Es fácil caer en la dinámica de compararnos con los que tienen más, pero no nos damos cuenta que la mayoría de nosotros, incluso en la peor de las situaciones, no estamos **ni tan mal** (Corpus del español, blog, personalizaciondeblogs.blogspot.com, 03/09/2013).

Por más que, en consonancia con su función de consecuente y remedo de su papel sintáctico de adjunto adverbial, la posición final del enunciado resulte la preferida, no faltan ejemplos de una creciente movilidad posicional con ocurrencias en anteposición enfática (13) propia, por lo demás, de la partícula *ni* (Martí Sánchez 1998, Sánchez López 1999, Porroche Ballesteros 2000, RAE-ASALE 2009, Albelda y Gras 2011, Conti Jiménez 2020, entre otros). La máxima señal de libertad posicional la alcanza ntm cuando aparece formando un enunciado propio con entonación exclamativa y plena autonomía sintáctica (14).

(13) Y eso fue maravilloso. Por aquí lo tenemos claro desde hace algún tiempo: «El soldado y la muerte», la versión de Henson de un cuento ruso bastante jodido, fue el episodio que hizo que la infancia se asomase por primera vez al abismo. Y oye, **ni tan mal** hemos salido (CORPES, D. Cuevas, «¿Cuál es la mejor serie de antología de la televisión?», jotdown.es, 06/2019).

(14) Un restaurante chino en España podría ser comparado con un restaurante de comida rápida española en China... ¡**Ni tan mal**! Lo que el cerdo es en España, el pato lo es en China.

[7] Como sugiere acertadamente un revisor, el análisis de este ejemplo podría corresponder a la siguiente estructura informativa: *no estamos* (tópico) / *ni tan mal* (comentario). El ejemplo refleja la vuelta a la sintaxis oracional de la partícula una vez construccionalizada, tal y como se bosqueja en la figura 3 más abajo.

¡Del pato hasta los andares! (EsTenTen18, motociclismo.es, 25/06/2017)

Las propiedades indicativas de la autonomía sintáctica de NTM encajan en las que definen, dentro del marco de la «gramática tética» (Kaltenböck, Heine y Kuteva 2011) o la «macrosintaxis» (Fuentes Rodríguez y Gutiérrez Ordóñez 2019, Fuentes Rodríguez 2024), a las unidades téticas o parentéticos[8]. Estos elementos adquieren su significado en la situación comunicativa con funciones textuales, subjetivas e interactivas y surgen por medio de un mecanismo de cooptación que explico en la sección § 3.3.

## *3.2. Origen de NTM: escalaridad y evaluación*

Las oraciones con negación interna <*no* + verbo + *tan mal*> —como las extraídas de los corpus orales (cfr. 6-8)— son generales en español y en ellas ha de situarse el arranque de la construccionalización[9] de NTM. La presencia de la negación en el predicado principal guía la interpretación comparativa de las secuencias encabezadas por el demostrativo *tan*, en tanto en cuanto invita a la recuperación anafórica del referente de *tan* y favorece la elisión de la coda comparativa (*NGLE* 2009: § 17.10p-q, Bosque y Sáez 2017: 125). La coda comparativa, por ejemplo —entre otras posibles (*como el anterior*, *como a Juan*, *como las prácticas*)—, *como pensaba* en (15a), marca el límite máximo en el resultado del examen. Este grado máximo, negado en las oraciones negativas ('el examen no me ha salido mal hasta ese extremo'), se transforma en el valor ponderativo propio de las oraciones afirmativas (15b)[10] y permite la continuación con una coda consecutiva intensiva (*que me van a suspender*). Asumiendo que las comparativas de igualdad implican una determinada orientación creciente, en caso de incrustarse en oraciones afirmativas, o decrecientes, si la estructura base es negativa (*NGLE*

[8] Dentro del marco de la «thetical grammar», una «unidad tética» se define como «syntactically unattached pieces of discourse» (Heine *et al.* 2017: 817), mientras que «parentético» es una estructura extraproposicional (Fuentes Rodríguez 2018: 21). La «gramática tética» —adaptable a lingüística española como «macrosintaxis»— diferencia el nivel de la sintaxis oracional (microsintaxis) del de la sintaxis tética o discursiva (macrosintaxis) y propone la interacción dinámica entre ambos.

[9] El diálogo entre la «gramática de construcciones» (baste citar Trousdale 2014) y la «gramática tética» (cfr., por ejemplo, Heine 2013) o «macrosintaxis» (véanse los trabajos recogidos en Fuentes Rodríguez 2020) se está revelando altamente fructífero. La creación «por construccionalización» de NTM cumple —como se verá en breve— con los requisitos de (a) productividad, (b) esquematicidad y (c) falta de composicionalidad postulados en el trabajo citado de Trousdale (2014).

[10] Reparemos en que la negación del predicado facilita la continuación con *como para* (*El examen no me ha salido tan mal como para que me suspendan*), mientras que esta secuencia parece más restringida con oraciones afirmativas (#*El examen me ha salido tan mal como para que me suspendan*). Sobre la negación, la coda consecutiva y *como para*, véase Sánchez López (2006: 64). Considerando que el complemento con *como para* sirve para identificar un grado determinado en una escala (Sánchez López 2006: 62), *no… tan mal* permite localizar ese grado (el máximo posible en estas construcciones) dentro de una escala y negarlo. Esto favorece la producción de inferencias semánticas (y argumentativas) que veremos a continuación.

2009: § 45.8i), podemos concluir que una oración como la versión afirmativa de (15a) orientaría hacia una escala de «maldad» ascendente y admitiría los encadenamientos de (15c) (cf. *peor* 'más mal') y, del mismo modo pero en progresión menguante, (15a) orienta hacia 'menos mal' (es decir: *mejor*). Las oraciones comparativas de (15a) y (15c) admiten encadenamientos argumentativos opuestos, como ponen de manifiesto los contrastes de (15d) y (15e), respectivamente (cf. Anscombre 1975, 1976, acerca de las repercusiones argumentativas de las comparativas de igualdad).

(15) a. El examen no me ha salido tan mal como pensaba.
b. El examen me ha salido tan mal…
c. El examen me ha salido tan mal como pensaba {si no peor ~ e incluso peor}
d. El examen no me ha salido tan mal como pensaba {creo que aprobaré / #creo que suspenderé}
e. El examen me ha salido tan mal como pensaba {#creo que aprobaré / creo que suspenderé}

En definitiva, la negación[11] del sintagma encabezado por *tan mal* permite excluir los grados de «maldad» situados por encima del punto individualizado por la coda comparativa, que se niega, mientras que favorece las inferencias hacia grados inferiores que, a la postre, confluyen en la zona del antónimo *bien*. Esta interpretación escalar de *no… tan mal* se refleja en la figura 2. En el caso de NTM, no resultan posibles —o, al menos, no se registran en los corpus investigados— las codas comparativas: *El examen no me ha salido ni tan mal *(como pensaba)*. Esta restricción, además de redundar en la atrición categorial del sintagma adverbial origen de la construcción, obliga a buscar en el contexto previo la recuperación del valor del demostrativo *tan*. Sea como fuere, la contribución de la partícula *ni* a la construcción consiste en negar el valor establecido —ahora implícitamente— por *tan mal* y excluir, además, cualquier valor superior (Sánchez López 1999: 2591, Conti Jiménez 2020: 100-102). Al excluir los valores superiores de la escala, *ni* está indicando que se niega el grado más alto de la escala (Albelda y Gras 2011: 25) y deja, en consecuencia, abierta la puerta a la interpretación positiva (*ni tan mal* ≈ *bien*)[12].

[11] En las escasas ocurrencias como constituyente oracional NTM actúa de término de polaridad negativa, pues en todos los casos aparece con un verbo negado: «… Si ya nos vinieran dos o tres como Landa no estaríamos ni tan mal, ¿no?», señala el corredor del Movistar» (EsTenTen18, diariodeltriatlon.es, 25/10/2017).

[12] Podríamos analizar el significado de la construcción NTM como la convencionalización de implicaturas generadas por la máxima de Modo griceana o el principio M de Levinson (2004: 74, 224-229), de suerte que el hablante que emplea *ni tan mal* no estaría en grado de afirmar *bien* y por lo tanto evocaría una zona intermedia entre el «mal» y el «bien», precisamente el significado procedimental al que apunta NTM.

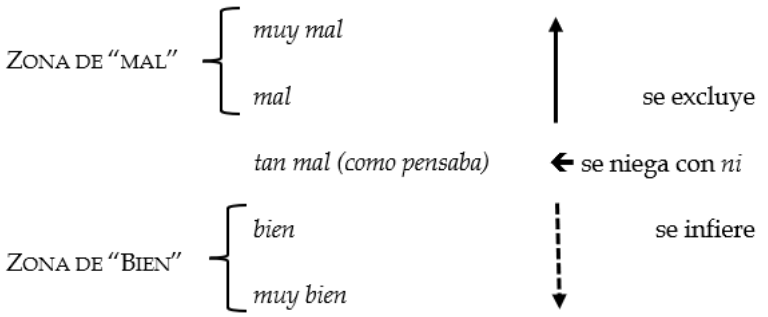



*Figura 2. Escala de* (no)… tan mal *y contribución de* ni *al significado de* NTM.

La comparación —contenido primitivo y básico de la expresión *(no)…tan mal*— implica la evaluación del hablante. En nuestra construcción, se compara, contrastándola con una nueva base argumentativa, el estado de cosas real con el resultado esperable si se hubiera aplicado una norma de referencia convencional. Además, el adverbio *mal* comporta una valoración subjetiva del hablante, para concluir que la conclusión efectiva, sin alcanzar tal vez lo esperado, ha de tenerse —cuando menos, mínimamente— por satisfactoria.

En definitiva, el operador NTM amplía «el [extenso] grupo de expresiones y locuciones» encabezadas por *ni* (*NGLE* 2009: 3714) compuesto, entre otros elementos, por expresiones minimizadoras (*ni un alma*, *ni un duro*) (Sánchez López 1999: 2614), estructuras concesivas (*ni que*) (Gras 2007) y, claro está, operadores modales de afirmación (*ni que decir tiene*) (Fuentes Rodríguez 2021) y de negación (*ni de broma*) (Padilla Herrada 2023). La incorporación de NTM a este conjunto testimonia la productividad del esquema construccional <*ni* + constituyente>. Además, la construcción NTM ha perdido la composicionalidad, en cuanto no admite la alternancia o conjunción con otros elementos (*ni tan mal ni tan bien*), la enumeración de alternativas (*ni tan mal como yo pensaba, ni tan mal como decía el profesor*) y no implica a otros elementos (*no habla ni con su padre ni con sus hermanos*). Esta pérdida de composicionalidad constituye un rasgo característico de los elementos que convocan escalas absolutas como ha explicado Portolés (2007: 214-215). En efecto, la fijación del conjunto de valores no admite alternativas como sucede cuando <*ni* + constituyente> configura un sintagma libre en estructuras discontinuas como (16a-b). Cuando invoca una escala relativa, donde se ordenan las alternativas posibles que la conforman, *ni* acepta la inserción de *siquiera* (16c-d). El rechazo de este adverbio confirma que *ni* en NTM apela a una escala absoluta, en la que, considerado un conjunto de valores, se señala un límite (Portolés 2000: 216). En calidad de operadores, las estructuras con *ni* se comportan como unidades (semi-)lexicalizadas de significado maximizador (16e). Este

valor maximizador caracteriza, igualmente, la locución NTM y resulta de la progresiva desemantización y decategorización.

(16) a. El examen no me ha salido ni tan mal como creía ni tan bien como deseaba mi madre.
b. No me habla ni en serio ni de broma.
c. No habla ni (siquiera) con su padre.
d. No me habla ni (siquiera) de broma, ¿cómo me va a hablar en serio?
e. No habla ni con su padre (≈> 'no habla con nadie en absoluto').

### *3.3. Ntm y el mecanismo de la cooptación*

La autonomía sintáctica, la construccionalización y la adquisición de funciones discursivas que caracteriza la evolución de NTM sugieren la intervención, en el surgimiento de esta partícula, del mecanismo de cooptación (*cooptation*)[13]. El operador se crea a partir de la construccionalización de las estructuras oracionales *no… tan mal* (microsintaxis) con la contribución de la partícula *ni*. Una vez construccionalizada, el operador se trasplanta a la esfera discursiva (primera fase de la cooptación), donde desarrolla sus tareas argumentativas. Como operador argumentativo vuelve a saltar al ámbito oracional (segunda fase de la cooptación). Las raras documentaciones de NTM dependiente de verbos (apenas un décimo de sus registros, cf. Tabla 1) apoyan este retorno a la oración desde la macrosintaxis. En el ámbito macrosintáctico, continúa desarrollando funciones discursivas en el campo de la marcación conversacional, con un trasvase sucesivo del nivel monológico al dialógico (cf. § 4.3 *infra*). La dirección del mecanismo de cooptación (microsintaxis > macrosintaxis > microsintaxis) y la adquisición de funciones discursivas en los niveles monológico y dialógico se esbozan en la figura 3.

[13] «Cooptation is a cognitive-communicative operation whereby some fragment of linguistic discourse is transferred from one domain of discourse to another» (Heine *et al.* 2017: 813). La introducción del concepto de cooptación está motivada por la necesidad de explicar los procesos de creación de los marcadores del discurso que ha obligado a reexaminar los fundamentos básicos de la gramaticalización «in a narrow sense» (véase Heine 2013 para una revisión crítica del estado de la cuestión). La creación de estos elementos requiere de dos procesos: la transferencia —espontánea— de una pieza gramatical del nivel sintáctico al discursivo (cooptación) y la fijación —gradual— de esta pieza dentro de una nueva categoría (gramaticalización).

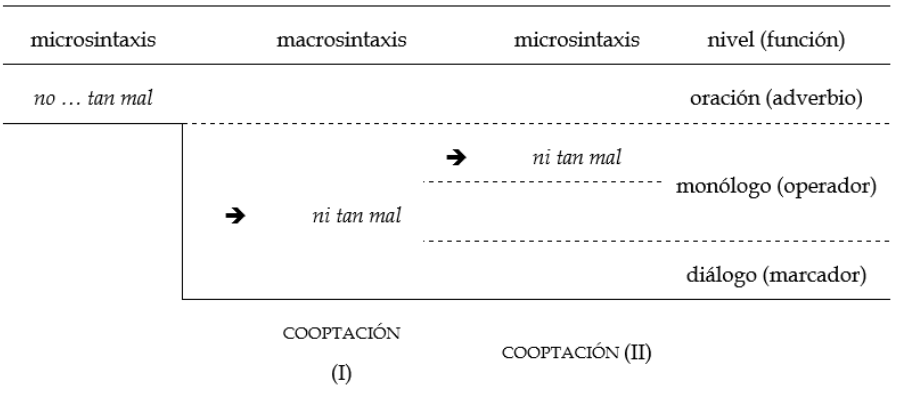



*Figura 3. Dirección del mecanismo de cooptación.*

## 4. La construcción NTM como «inversor argumentativo»

### *4.1. Esquemas argumentativos*

El operador NTM participa en varias estructuras contraargumentativas. La más frecuente se corresponde al esquema de «contraargumentación indirecta» (Portolés 1995: 247). En este esquema se presentan dos argumentos (A y B) encaminados hacia conclusiones opuestas. La reorientación argumentativa vehiculada por este esquema suele estar explicitada mediante el conector *pero* (17a). Del primer enunciado (A: *llevamos desde hace muchos años siendo engañados*) se extrae la conclusión (C) de que los ciudadanos deberíamos rebelarnos contra el sistema; sin embargo, el hecho expreso en el segundo enunciado (B) de que las cosas —sería deseable, a pesar de todo, que funcionaran mejor— fueran bien prevalece y justifica la conclusión (C'), condescendiente y acomodaticia, de que no merece la pena actuar. A través de NTM, se invierte el peso de los argumentos según un movimiento contraargumentativo prototípico de *pero* (Anscombre y Ducrot 1977, Portolés 1995: 244), si bien puede inferirse del contenido léxico de las palabras del enunciado implícitamente contrapuestas (18): las connotaciones negativas de la voz *crisis* se atenúan por un efecto positivo: el protagonismo de la vida de los pueblos.

(17) a. Llevamos desde hace muchos años siendo engañados, pero como la cosa no iba mal del todo, pues **ni tan mal**. Pan y circo, oiga, no necesitamos más (EsTenTen18, perdidoeneldesierto.es, sin fecha).
b. A: Llevamos desde hace muchos años siendo engañados →

C: Deberíamos rebelarnos **pero** B: la cosa no iba mal del todo → C': no hemos creído necesario rebelarnos.

(18) a. Con ésta crisis los pueblos van a volver a cobrar protagonismo.... **ni tan mal** (EsTenTen18, wersemei.es, sin fecha).
b. A: crisis → C: consecuencias negativas [**pero**] B: protagonismo de los pueblos → C': una consecuencia positiva.

(19) a. Rober, yo envie un mail a atencion del cliente de estos hijos de puta, me respondieron diciendome como lo tenia que hacer, asi que por lo menos **ni tan mal**, ellos mismos me han recordado mis datos (EsTenTen18, tencuidado.es, 18/06/2011).
b. B: me explicaron cómo tenía que hacer (argumento mínimamente suficiente) [**y además**] B': me recordaron mis datos (argumento mínimamente suficiente en apoyo de B).

(20) Antes, no obstante, pueden sufrir humillaciones de distinta intensidad, como que se les prohíba la participación en desfiles militares y en las procesiones de Semana Santa, dos de los momentos cumbres para los legionarios cada año. Y hasta ahí, **ni tan mal**, porque en el siguiente nivel las amenazas del alto mando pasan a no dejarles participar en cursos, no pagarles el complemento de dedicación especial, excluirlos de recompensas, felicitaciones y premios o, incluso, no dejar que participen en misiones internacionales (EsTenTen18, diariodepontevedra.es, 08/01/2018).

(21) **BV**: Los DVD son mas difíciles de cuidar y de conseguir! Aunque la verdad me parece molesto tener que cambiar de disco durante el juego :S
**YS**: Bueno parece ser que solo habrá que cambiar una vez de disco, **ni tan mal**! (EsTenTen18, borntoplay.es, 02/10/2012).

(22) Los de pañales y llantos **ni tan mal** pero cuando nos ponemos a contar nuestros partos, a veces parecen hilos gore, ja,ja,ja... pero al menos una se mantiene (EsTenTen18, spaniards.es, sin fecha).

Repárese cómo en (17) el hablante añade un argumento más («no necesitamos más que pan y circo») con el objeto de reforzar el que, por más que suficiente, considera débil. La suma de argumentos para consolidar el consecuente se observa en varios ejemplos del corpus. La dinámica aditiva de (19a) está organizada (19b), de modo que, una vez propuesta una razón de suficiencia mínima, el hablante la sostiene con argumentos adicionales de fuerza equiparable, pero

igualmente coorientados: B *y además* B'. Frente a la falta de argumentos cualitativamente fuertes, el hablante recurre a la acumulación de argumentos mínimamente suficientes para justificar la conclusión. Esta justificación puede darse, asimismo, por medio de causales (20). En definitiva, el argumento B, que orienta la argumentación, contiene la suficiencia mínima para garantizar la mejor conclusión posible. Este grado mínimo de suficiencia argumentativa se pone de manifiesto a través de expresiones, como *solo* (21), de límite inferior. Por supuesto, el hablante, consciente de la existencia de argumentos con mayor fuerza argumentativa, puede continuar el razonamiento con premisas que matizan la conclusión mediante una estructura argumentativa del tipo: B (*así que* NTM) *pero* A (22).

Con relativa frecuencia, el argumento B —de suficiencia mínima para garantizar la prosecución del discurso— se expresa bajo la forma de una prótasis condicional (23-25). En el primer ejemplo de la batería, *solo* —de nuevo— indica el polo más bajo de una escala <mocos, bronquiolitis>, contraponiendo la mucosidad a una situación más grave. La no documentación del adverbio *entonces* en estas estructuras condicionales sugiere que la relación entre prótasis y apódosis no es del tipo implicativo (<*si* p, *entonces* q>), sino que más bien pertenecen a las «condicionales de actos de habla» (Montolío 2000: 158-159), confirmando la enunciación como campo de funcionamiento de NTM. En la tanda de microtextos se observa, además, cómo los emisores, en apoyo de la menos mala de las conclusiones posibles, aportan causas (la difusión de la bronquiolitis entre los niños), argumentos complementarios (la insustancialidad de un personaje) o circunstancias atenuantes (la situación económica).

(23) Bueno, pues si solo son mocos **ni tan mal**. Porque fíjate cómo están con la bronquiolitis Tenemos que tener mucho cuidado con estos peques (Whatsapp, 22/11/2022).

(24) Bueno, si la palma el colega pues **ni tan mal**, nos olvidaremos de él dentro de dos capítulos (EsTenTen18, pirateking.es, 30/08/2017).

(25) De como pintaba a estas alturas a como quedó al final desde febrero hubo una notable diferencia, pero lo justito para seguir viviendo. Fíjate que este año, viendo lo que hay, si sale así **ni tan mal**», cuenta Fernando Dimas Andrés, de Dimas Ski (EsTenTen18, eldiariomontanes.es, 03/01/2017).

Menos habitual que los anteriores resulta el esquema coincidente con el patrón de «contraargumentación directa» (Portolés 1995: 244). De acuerdo con este patrón se atenúan las consecuencias del argumento

más fuerte (A) que no conduce a la conclusión esperable (26), sino a una sorprendentemente satisfactoria (26b). El conector *pero* subraya este matiz de contraexpectativa (Rodríguez Rosique 2023) connotado por NTM[14].

(26) a. Y, de despedida, el típico dulce japonés con dos bizcochos de forma redonda rellenos, en este caso con judía roja. Un poco sequetes, como todos los dorayaki que solemos probar, pero **ni tan mal** (EsTenTen18, eatandlovemadrid.es, 30/12/2013).
b. A: los dorayaki estaban secos → C: no estaban gustosos **pero** C': estaban sorprendentemente buenos.

## *4.2. Las escalas complejas y el complemento de proporción*

NTM requiere una «escala compleja» (Portolés 2007: 203-204). En efecto, los enunciados sobre los que interviene están basados en dos conjuntos de argumentos ordenados en sendas escalas y pertenecientes a clases argumentativas (Ducrot 1980: 16). Estas clases argumentativas están sustentadas sobre *topoi* o «formas tópicas» inversas, en la medida en que el argumento *q* se orienta hacia la conclusión *r*, mientras que *p* encamina la argumentación hacia *~r*. Consideremos el siguiente contexto. Para visitar una ciudad podemos reservar un hotel en función de dos criterios: la ubicación (escala U) y el precio (escala P). Cualquiera de las dos escalas en juego es susceptible de aportar argumentos para reservar una habitación, pero cada una lo hace de manera inversa (<+U, -P>): si *q* es un argumento en favor de *r* en la escala U (27a), *p* vale como un argumento para *~r* en la escala P (27b)[15].

(27) a. El hotel es caro, pero está céntrico, así que reservamos la habitación (*p* pero *q*, entonces *r*)
b. El hotel está céntrico, pero es caro, así que no reservamos la habitación (*q* pero *p*, entonces *~r*)

[14] Menudean los casos en los que NTM va acompañado de marcadores de contraexpectativa como *la verdad es que* (Soler Bonafont 2017):

(i) Sí, ya he terminado mi primera sesión con Ángel y la verdad es que **ni tan mal** (CORPES, C. Pradas Gallardo, *Todo saldrá (bien)*, 2022)

o *al final* (Pardo Llibrer 2017):

(ii) porque las clases en tu casa, que las ves por el ordenador, no tienes que ir a Almería que tú ya sabes que yo no podría ir a clases desde el pueblo, y los libros, como los he fotocopiao pues mira, al final, **ni tan mal** (Corpus del español, trabajosocialytal.blogspot.com, 01/08/2013).

[15] Por supuesto, los argumentos pueden poseer grados de fuerza divergentes en sus respectivas escalas (U: *q*'' > *q*' > *q* / P: *p*'' > *p*' > *p*), donde «>» equivale a 'es argumentativamente más fuerte para *r* que'.

(28) La ubicación, por el precio **ni tan mal**, en relacion a lo que uno se espera de Paris. La cama me parecio comoda (aunque no se si fue porque llegaba cansadisima cada dia) (EsTenTen18, hotelmix.es, 09/2023).

El esquema argumentativo desarrollado por NTM no encaja exactamente en los patrones de (27), ya que no conecta dos valores situados en posiciones directamente proporcionales, sino que el operador argumentativo lleva a cabo, más bien, un mecanismo de compensación a fin de alcanzar el valor más alto en la escala U con relación a un valor considerado mínimamente suficiente en la escala P. Este mecanismo compensatorio, esquematizado en la figura 4, se ilustra en (28): la ubicación del hotel ha de juzgarse como la mejor posible considerando el precio del alojamiento en una ciudad como París.

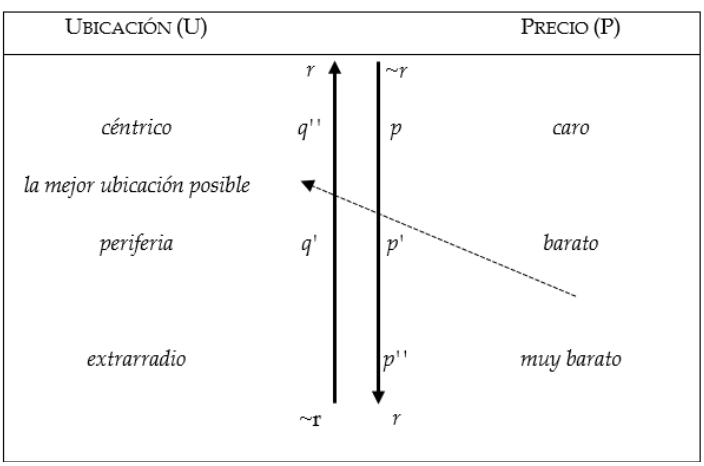



*Figura 4. Escala compleja de* NTM (r = *reservar una habitación)*

El mecanismo compensatorio apenas formulado exige la introducción de coordenadas argumentativas extraordinarias que inviten a evaluar el peso y la suficiencia de los argumentos dentro de una nueva norma de referencia a la que se adecue la argumentación. Esta misión compete al denominado «complemento de gradación o proporción» (Fuentes Rodríguez 2003: 296), generalmente introducido por la preposición *para* (29)[16].

[16] El concepto de complemento de proporción (o de suficiencia) se debe a Salvá (1850: 255) quien lo ilustra con un ejemplo (*Para ser nuevo en las tablas, no lo hace del todo mal*) en el que bien podría emplearse hoy, 200 años después, nuestro NTM. Han sido bien estudiadas las lecturas proconcesivas, de contraexpectación e intensificación generadas por este sintagma preposicional con *para*, clave en el surgimiento de construcciones contraargumentativas. Por motivos de espacio me limito a citar el reciente estudio de Grande Alija (2022) y remito al completo estado de la cuestión que presenta el autor. En los ejemplos se encuentran también otras formas de expresar el marco argumentativo:

(i) Viendo la histórica desidia de este ayuntamiento con la cultura y el casi nulo compromiso con la juventud, que hayan tenido esta iniciativa creo que no está ni tan mal (Corpus del español, diariodevurgos.com, 25/08/2009).

(29) Pues al final genial, las pistas **ni tan mal** para el tiempo que lleva sin caer una nevada (EsTenTen18, amigosmadrid.es, 05/11/2016).

El complemento de proporción constituye, por tanto, un componente central en los esquemas argumentativos en los que actúa NTM y canaliza así su significado procedimental básico: «dadas las circunstancias, la conclusión alcanzada supone la mejor opción posible entre las consideradas». Este significado está fundado sobre el significado comparativo originario del sintagma *tan mal*, pues la conclusión óptima se alcanza después de haber comparado entre sí las opciones disponibles.

### *4.3. De operador argumentativo a marcador conversacional*

Las muestras de NTM recogidas en el español oralizado del siglo XXI consienten examinar la transformación de un operador argumentativo en un marcador conversacional. Esta transformación culmina en el nivel dialógico, donde el marcador aparece como reacción a un enunciado del interlocutor. En las conversaciones digitales de (30-31), advertimos los tipos argumentativos descritos en § 4.1 en los que NTM actúa como operador argumentativo rebajando el peso de un argumento mediante el esquema <*si* B, NTM> (30) o adecuando las premisas a un nuevo marco argumentativo expresado en el complemento de proporción (31). El uso dialógico de NTM se encuentra también en otros géneros interactivos como las conversaciones digitales realizadas a través de los foros electrónicos (32). Lo característico del empleo de NTM en estos microdiálogos consiste en que la operación argumentativa se distribuye en las intervenciones de hablantes diferentes.

(30) **S1**: Holaa q tal la vuelta al cole? Jajajaj
**M1**: Joder de lunes total ☹☹☹
**S2**: Jajaja si solo es eso **ni tan mal**
**M2**: Jajaja (WhatsApp, 02/10/2023)

(31) **J1**: Dura??
**C1**: Un sufrimiento. Una recta de 5 km, ida y vuelta…
**J2**: Hombre, para tus zapatillas nuevas, **ni tan mal**
C2: Por eso decía 😀😀😀 (Whatsapp, 21/02/2024)

(32) **Usuaria 1**: La verdad es que si…fuera de un par de veces que decidió que en el cine se habla fuerte y me contaba la peli de monstruos el resto ¡fue total! 🙂un besote desmadroso
**Usuaria 2**: **Ni tan mal** amiga…. te diré que yo quiero probar

este año y pensaba esperarme al estreno de los pitufos.
(Corpus del español, web, desmadreando.com, 01/07/2013).

En las intervenciones reactivas, el hablante refuerza el acuerdo con su interlocutor mediante la confirmación de los contenidos implícitos inferibles de las intervenciones previas. De esta forma, se pone en acción una estrategia social de refuerzo de la imagen positiva del interlocutor acorde con el tipo de intensificación «alo-reafirmativa» (Briz 2017a, 2017b)[17].

## 5. A modo de conclusión

En las páginas anteriores he realizado una caracterización preliminar y exploratoria de NTM. He definido este elemento como una unidad parentética y he delineado, estableciendo un vínculo con el sintagma primitivo *(no)…tan mal*, su proceso de construccionalización. Además, he examinado la operación de inversión argumentativa que realiza. En esta operación se ha revelado fundamental el denominado complemento de proporción, elemento clave en el surgimiento de construcciones contraargumentativas. Finalmente, he esbozado su transformación (incipiente) en un marcador conversacional de acuerdo (o, si se prefiere: operador modal confirmativo).

Aparte de afinar, tanto en lo que se refiere a sus aspectos gramaticales cuanto, especialmente, en lo relativo a sus funciones argumentativas y discursivas, la caracterización de NTM presentada en este estudio, investigaciones futuras deberán explorar nuevas hipótesis (la miratividad, por ejemplo) para comprender su comportamiento, delimitar sus espacios diastráticos —tal vez—, dialectales y —en general— variacionales, trazar con exactitud sus vías de difusión y, en definitiva, realizar un seguimiento —¿longitudinal?— de su desarrollo. Las perspectivas de estudio que ofrece NTM son merecedoras de la mayor atención por parte de la lingüística diacrónica, pues pocas etapas de la historia de nuestra lengua conceden, como lo hace el español del tercer milenio, la oportunidad nítida de examinar la emergencia de un marcador discursivo.

## Bibliografía

Albelda Marco, Marta (2014), «Escalaridad y evaluación: rasgos caracterizadores de la intensificación pragmática», en Elisa Putska y

[17] Los componentes de escalaridad y evaluación configuradores de la semántica de NTM (§ 3.2) forman los rasgos definitorios de la intensificación pragmática (Albelda 2014).

Stephanie Goldschmitt (eds.), E*motionen, Expressivität, Emphase*, Berlín, Erich Schmidt Verlag: 79-94.

Albelda Marco, Marta, y Pedro Gras Manzano (2011), «La partícula escalar *ni* en español coloquial», en Ramón González y Carmen Llamas (eds.), *Gramática y discurso: nuevas aportaciones sobre partículas discursivas del español*, Pamplona, EUNSA: 15-38.

Anscombre, Jean-Claude (1975), «Il était une princesse aussi belle que bonne (I)», *Semantikos*, 1 (1): 1-28.

Anscombre, Jean-Claude (1976), «Il était une princesse aussi belle que bonne (II)», *Semantikos*, 1 (2): 1-26.

Anscombre, Jean-Claude, y Oswald Ducrot (1977), «Deux *mais* en français?», *Lingua*, 43: 23-40.

Anscombre, Jean-Claude, y Oswald Ducrot (1994), *La argumentación en la lengua*, Madrid, Gredos.

Bosque Muñoz, Ignacio, y Luis Sáez (2017), «La naturaleza composicional de *tan(to)* y los contextos antiasertivos», en Ángel J. Gallego Bartolomé, Yolanda Rodríguez Sellés y Javier Fernández Sánchez (eds.), *Relaciones sintácticas: homenaje a Josep M. Brucart y M. Lluïsa Hernanz*, Barcelona, Universitat Autònoma de Barcelona: 121-140.

Brenes Peña, Ester (2020), «De construcciones a operadores: la alusión al decir», en Catalina Fuentes Rodríguez (ed.), *Operadores en proceso*, Múnich, Lincom: 74-110.

Briz Gómez, Antonio (2017a), «Una propuesta funcional para el análisis de la estrategia pragmática intensificadora en la conversación coloquial», en Marta Albelda y Wiltrud Mihatsch (eds.), *Atenuación e intensificación en géneros discursivos*, Madrid / Frankfurt am Main, Iberoamericana / Vervuert: 43-67.

Briz Gómez, Antonio (2017b), «Otra vez sobre las funciones de la intensificación en la conversación coloquial», *Boletín de Filología*, 52 (2): 37-58.

Cianca Aguilar, Elena y Emilio Gavilanes Franco (2018), «Voces y expresiones del argot juvenil madrileño actual», *Círculo de Lingüística Aplicada a la Comunicación*, 74: 147-168.

COLA = *Corpus del lenguaje adolescente* [en línea]. Disponible en: https://clarino.uib.no/korpuskel/corpora [Fecha de consulta: junio de 2024].

Conti Jiménez, Carmen (2020), «¿Coordinadores discontinuos en español? Problemas de análisis de los correlativos

disyuntivos y copulativos», *Onomázein*, 49: 88-114. DOI: 10.7764/onomazein.49.05.

CORPES = Real Academia Española (en línea), *Corpus del Español del Siglo XXI* [en línea]. Disponible en: https://rae.es/corpes [Fecha de consulta: junio de 2024].

COSER = Fernández-Ordóñez, Inés (dir.) (2005), *Corpus oral y sonoro del español rural* [en línea]. Disponible en: http://www.corpusrural.es [Fecha de consulta: junio de 2024].

*DPDE* = Briz, Antonio, Salvador Pons, y José Portolés (coords.) (2008), *Diccionario de partículas discursivas del español* [en línea]. Disponible en: https://dpde.es [Fecha de consulta: septiembre de 2024].

Ducrot, Oswald (1980), *Les échelles argumentatives*, París, Éditions de Minuit.

Ducrot, Oswald (1983), «Opérateurs argumentatifs et visée argumentative», *Cahiers de Linguistique Française*, 5: 7-36,

EsTenTen = *SketchEngine: EsTenTen* [en línea]. Disponible en: https://auth.sketchengine.eu/> [Fecha de consulta: junio de 2024].

Felíu Arquiola, Elena y Enrique Pato (2019), «¿Realmentes existen?: la "pluralización" de los adverbios en *-mente* en español actual», ***Onomázein***, 44: 166-190. DOI: 10.7764/onomazein.44.08.

Fuentes Rodríguez, Catalina (2003), «Factores argumentativos y correlatos sintácticos», *Estudios de Lingüística de la Universidad de Alicante*, 17: 289-304. DOI: 10.14198/ELUA2003.17.16.

Fuentes Rodríguez, Catalina (2009), *Diccionario de conectores y operadores del español*, Madrid, Arco/Libros.

Fuentes Rodríguez, Catalina (ed.) (2018), *Parentéticos*, Madrid, Arco/Libros.

Fuentes Rodríguez, Catalina (2021), «*Ni que hablar / ni que decir*: ¿Construcciones u operadores escalares?», *Pragmalingüística*, 29: 149-172.

Fuentes Rodríguez, Catalina (ed.) (2020), *Operadores en proceso*, Múnich, Lincom.

Fuentes Rodríguez, Catalina (2024), *Macrosintaxis del español*, Berlín, De Gruyter. DOI: 10.1515/9783111315454

Fuentes Rodríguez, Catalina y Salvador Gutiérrez Ordóñez (eds.) (2019), *Avances en macrosintaxis*, Madrid, Arco/Libros.

Grande Alija, Francisco Javier (2022), «Escalas argumentativas y estructuras con *para* de valor comparativo-intensificador», en Catalina Fuentes Rodríguez (ed.), *Operadores argumentativos*, Madrid, Arco/Libros: 211-246.

Gras Manzano, Pedro (2007), «Gramática y pragmática de construcciones. Subordinadas introducidas por *ni que*: un enfoque construccionista», en Pablo Cano López *et al.* (coords.), A*ctas del VI Congreso de Lingüística General*, vol. 2, Madrid, Arco/Libros: 1609-1620.

Heine, Bernd (2013), «On discourse markers: grammaticalization, pragmaticalization, or something else?», *Linguistics*, 51 (6): 1205-1247.

Heine, Bernd, Gunther Kaltenböck, Tania Kuteva, y Haiping Long (2017), «Cooptation as a discourse strategy», *Linguistics*, 55: 1-43.

Kaltenböck, Gunther, Bernd Heine, y Tania Kuteva (2011), «On thetical grammar», *Studies in Language*, 35 (4): 848-893.

Kilgarriff, Adam, e Irene Renau (2013), «esTenTen, a vast web corpus of Peninsular and American Spanish», *Procedia - Social and Behavioral Sciences*, 95: 12-19. DOI: 10.1016/j.sbspro.2013.10.617.

Levinson, Stephen C. (2004), *Significados presumibles: la teoría de la implicatura conversacional generalizada*, Madrid, Gredos.

López Serena, Araceli (2018), «Hacia una revisión de la caracterización semántica y discursiva de la locución *y eso que* en español actual», *Estudios de Lingüística de la Universidad de Alicante*, 32: 195-217.

Martí Sánchez, Manuel (1998), «Recorrido por *ni*», *Lingüística Española Actual*, 20 (1): 79-108.

Martín Zorraquino, María Antonia, y José Portolés Lázaro (1999), «Los marcadores del discurso», en Ignacio Bosque y Violeta Demonte (eds.), *Gramática descriptiva de la lengua española*, Madrid, Espasa-Calpe: 4051-4214.

Montolío Durán, Estrella (2000), «On affirmative and negative complex conditional connectives», en Elizabeth Couper-Kuhlen y Bernd Kortmann (eds.), *Cause - Condition - Concession - Contrast: cognitive and discourse perspectives*, Berlín, De Gruyter: 143-171.

*NGLE* = Real Academia Española y Asociación de Academias de la Lengua Española (2009), *Nueva gramática de la lengua española*, Madrid, Espasa-Calpe.

Padilla Herrada, María Soledad (2023), *La negación reactiva en el español actual: una aproximación desde la macrosintaxis*, Lausana, Peter Lang.

Pardo Llibrer, Adrià (2017), «Sobre *al final*», *Estudios Interlingüísticos*, 6: 132-152.

Porroche Ballesteros, Margarita (2000), «Aspectos de *ni* como marcador discursivo», en José Jesús de Bustos Tovar (coord.), *Lengua, discurso, texto*, Madrid, Visor: 669-682.

Portolés Lázaro, José (1995), «Diferencias gramaticales y pragmáticas entre los conectores discursivos *pero*, *sin embargo* y *no obstante*», *Boletín de la Real Academia Española*, 75 (265): 231-270.

Portolés Lázaro, José (1998a), «El concepto de suficiencia argumentativa», *Signo y Seña*, 9: 199-224.

Portolés Lázaro, José (1998b), «La teoría de la argumentación en la lengua y los marcadores del discurso», en María Antonia Martín Zorraquino y Estrella Montolío Durán (eds.), *Los marcadores del discurso: teoría y análisis*, Madrid, Arco/Libros: 71-92.

Portolés Lázaro, José (2007), «Las escalas informativas convocadas por *ni* y *ni siquiera*», *Revista Internacional de Lingüística Iberoamericana*, 10: 199-219.

Portolés Lázaro, José (2016), «*Razón de más* como inversor argumentativo», *Revista Internacional de Lingüística Iberoamericana*, 27 (1): 157-172.

PRESEEA = *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América* [en línea]. Disponible en: http://preseea.linguas.net> [Fecha de consulta: junio de 2024].

Rodríguez Rosique, Susana (2023), «Más allá de la suficiencia argumentativa: *pero* y la miratividad», *Biblioteca de Babel: Revista de filología hispánica*, vol. extra. 1: 163-178.

Salvá, Vicente (1850), *Gramática de la lengua castellana según ahora se habla*, París, Librería de los hermanos Garnier.

Sánchez López, Cristina (1999), «La negación», en Ignacio Bosque y Violeta Demonte (eds.), *Gramática descriptiva de la lengua española*, Madrid, Espasa-Calpe: 2561-2634.

Sánchez López, Cristina (2006), *El grado de adjetivos y adverbios*, Madrid, Arco/Libros.

Soler Bonafont, María Amparo (2017), «*La verdad (es que)*: significado nuclear y atenuante», *Revista Signos: Estudios de Lingüística*, 50 (95), 430-452.

Trousdale, Graene (2014), «On the relationship between gramaticalization and constructionalization», *Folia Lingüística*, 48 (2): 557-577.

Yus, Francisco (2020), «La comunicación en la era digital», en María Victoria Escandell-Vidal, José Amenós Pons y Aoife Kathleen Ahern (eds.), *Pragmática*, Madrid, Akal: 608-623.

Mar Capilla Martín (ORCID: 0009-0002-3979-1527), «Entre lo oral y lo escrito: un estudio de corpus sobre los reformuladores *es decir*, *o sea* y *en plan* en español». *Biblioteca de Babel: Revista de Filología Hispánica*, vol. extra. 2 (2024), pp. 195-212.
**DOI**: 10.15366/bibliotecababel2024.extra2.008 · **ISSNe**: 2695-6349 · Recibido: 03/10/2024 · Aceptado: 15/11/2024

# Entre lo oral y lo escrito: un estudio de corpus sobre los reformuladores *es decir*, *o sea* y *en plan* en español

Mar Capilla Martín
*Centro de Estudios de la Real Academia Española*
mcapilla@rae.es

**Resumen**: Esta investigación tiene como objetivo identificar si existen diferencias contextuales en el uso de *es decir*, *o sea* y *en plan* entre los hablantes de español. Con la aparición de material procedente de Internet, existe un nuevo paradigma que ubica ciertos usos de los marcadores discursivos en la intersección entre lo oral y lo escrito. En este contexto, resulta pertinente un estudio que tome como referencia un corpus específico, como el CORPES XXI, caracterizado por una notable representación de este tipo de fuentes, así como por facilitar datos sobre la distribución tipológica y temática del texto. Los resultados del análisis revelan la presencia de diferencias determinadas por el tipo de texto en el que los hablantes emplean cada uno de los marcadores discursivos examinados en este estudio. En este sentido, el artículo se presenta como un trabajo que evidencia los nuevos intereses y enfoques en el análisis del discurso en la actualidad.

**Palabras clave**: lingüística de corpus, corpus oral, lenguaje coloquial, reformuladores, marcadores discursivos.

## Between spoken and written language: a corpus study on the reformulators *es decir*, *o sea*, and *en plan* in Spanish

**Abstract**: This research aims to identify whether there are contextual differences in the use of *es decir*, *o sea*, and *en plan* among Spanish speakers. With the emergence of material from the Internet, a new paradigm has arisen that positions certain uses of discourse markers at the intersection between oral and written forms. In this context, a study that takes as a reference a specific corpus, such as CORPES XXI, characterized by a remarkable representation of this type of sources, as well as by providing data on the typological and thematic distribution of the text, is pertinent. The results of the analysis reveal the presence

of differences determined by the type of text in which speakers use each of the discourse markers examined in this study. In this sense, the article constitutes a work that highlights new interests and approaches in contemporary discourse analysis.

**Keywords**: corpus linguistics, oral corpus, colloquial language, reformulators, discourse markers.

## 1. Introducción

En el vasto campo de la lingüística contemporánea, el uso de corpus ha revolucionado la manera en que investigamos y comprendemos el lenguaje. El corpus lingüístico, ya sea oral o escrito, constituye el recurso más efectivo y natural para el análisis del lenguaje. Por ello, cualquier investigación lingüística debe fundamentarse en un conjunto amplio y bien delimitado de materiales que ofrezcan datos fiables (Briz 2012). Entre los diferentes tipos de corpus, los corpus que contienen textos escritos y orales ocupan un lugar destacado al capturar el lenguaje en diferentes medios, lo que permite un análisis profundo de fenómenos lingüísticos (Bolaños 2015; Rojo 2021). Los corpus lingüísticos se caracterizan por tres aspectos distintivos: un diseño, la recuperación selectiva de datos y la codificación textual. Por otro lado, también la calidad de los textos incluidos es un elemento fundamental, estrechamente ligada al proceso de codificación y anotación. En este sentido, resulta esencial subrayar la relevancia de la selección y codificación del material para garantizar la calidad de los datos que serán analizados en fases científicas posteriores.

Teniendo esto en cuenta, el presente artículo pretende explorar la importancia y relevancia de los corpus que contienen textos escritos y orales en la investigación lingüística del habla coloquial. A través de ejemplos y estudios de caso, se examinará la presencia y la naturaleza de tres marcadores discursivos propios de este tipo de lenguaje en diferente tipología y temática textual con el fin de arrojar luz sobre qué contexto lingüístico es más propenso a presentar estos marcadores. El trabajo se estructurará en tres secciones principales. En primer lugar, se presentarán las características fundamentales que debe tener un corpus, o subcorpus, compuesto por textos escritos y orales, así como los beneficios de su utilización para la investigación lingüística. En segundo lugar, se ofrecerá un resumen de los rasgos más relevantes de los reformuladores *es decir*, *o sea* y *en plan*, con el propósito de vincularlos a los datos extraídos del corpus empleado para nuestra investigación, el CORPES XXI. Finalmente, se analizará la información que este corpus

proporciona y su distribución tipológica y temática, con el fin de identificar en qué contextos, ya sean orales o escritos, se emplean estos marcadores del discurso.

## 2. Los corpus, o subcorpus, orales como fuente de investigación

El empleo de corpus lingüísticos en el campo de la lingüística ha transformado el estudio de la lengua en las últimas décadas. La lingüística de corpus ha subrayado la importancia de fundamentar la descripción lingüística en un estudio exhaustivo de esta en su uso natural, proporcionando una perspectiva de análisis más amplia y datos de frecuencia más precisos que facilitan la interpretación y la veracidad de los resultados. En consecuencia, podemos afirmar que el uso de corpus lingüísticos ha revolucionado la investigación del lenguaje (Rojo 2015, 2021). Los corpus lingüísticos son un conjunto de textos, tanto orales como escritos, obtenidos en condiciones naturales y pertenecientes a una lengua o variedad lingüística específica, que se almacenan en formato electrónico y se codifican con el propósito de ser sometidos a análisis. La elaboración de un corpus implica la integración de textos conforme a un diseño previamente establecido. En este sentido, cada corpus adopta una estructura general que responde a los objetivos para los cuales ha sido construido (Rojo 2021). En cuanto a su tipología, los corpus lingüísticos pueden ser compilados de diversas maneras y en variados tamaños, abarcando desde colecciones de textos de un único autor hasta extensas recopilaciones de textos de distintos géneros y épocas. Cada corpus se crea con unos objetivos específicos que lo diferencian de otros corpus, por lo que se pueden agrupar y clasificar según distintos criterios. Dado que nuestra investigación se centra en la tipología y temática textual, abarcando tanto textos escritos como orales, es necesario basarnos en corpus que distingan el medio en el que se encuentra el texto, ya sea escrito u oral. En este sentido, existen corpus de textos escritos, conocidos como *written corpora* o *text corpora*, que consisten en textos originalmente escritos. Por otro lado, tenemos los corpus que contienen textos orales, que consisten principalmente en la transcripción ortográficamente convencional de grabaciones. El tratamiento y análisis del corpus se efectúa a partir de dicha transcripción, denominada *texto hablado* (*spoken text*). La transcripción ortográfica se enriquece con diversos elementos que reflejan el proceso de producción del habla, adaptándose a los objetivos y aplicaciones del corpus. El objetivo es proporcionar una representación simbólica del uso oral espontáneo y natural de la lengua, sin incluir la variación fonética (Sinclair 1996).

En este contexto, es fundamental entender que la distinción entre oral y escrito suele estar asociada con la diferencia en los registros: los textos orales se vinculan comúnmente con el lenguaje coloquial, mientras que los textos escritos tienden a reflejar una orientación más literaria y cercana a la variedad estándar en el ámbito lingüístico correspondiente. A pesar de que esta asociación puede ser adecuada en gran medida, es importante reconocer que la naturaleza de la distinción se centra en el soporte material del texto y no en otras características. Una prueba de ello es la aparición y evolución de los medios electrónicos y las redes sociales, que ha dado lugar a nuevos géneros en los cuales el soporte escrito se combina frecuentemente con la lengua coloquial, como es el caso de los blogs y los tuits (Rojo 2021). Así, es importante destacar que, incluso en textos escritos, es posible identificar patrones propios de la oralidad. Esto se debe a que, en numerosos casos, dichos textos buscan emular la oralidad mediante el uso de diálogos, la transcripción de entrevistas y otros recursos similares.

En español, existe una amplia disponibilidad de corpus que contienen textos escritos como el CORDE, el CREA, el CORPES XXI, el CE o el CDH; y textos orales, entre los cuales destacan el corpus PRESEEA, Val.Es.Co., COSER, ESLORA, AMERESCO, COVJA o CREA, entre otros. Estos corpus, de distintas dimensiones y características, permiten obtener una visión general de fenómenos lingüísticos, proporcionando una base esencial para el estudio del habla espontánea y coloquial. Su relevancia es particularmente significativa para los análisis pragmalingüísticos y sociopragmáticos del español, facilitando la investigación en áreas como la atenuación, la cortesía o los marcadores discursivos (Briz y Carcelén 2019). Además, es relevante destacar la accesibilidad de estas herramientas, ya que están disponibles a través de plataformas en línea que permiten filtrar las búsquedas. Algunas de estas plataformas incluso facilitan la creación de subcorpus específicos, lo que posibilita la extracción de información más detallada y precisa, optimizando así su utilización en la investigación lingüística. Uno de estos corpus es el Corpus del Español del Siglo XXI (CORPES XXI), un corpus de referencia que incluye tanto un subcorpus oral como información textual escrita, lo cual permite la comparación de datos, enfocando el análisis en rasgos del habla espontánea y coloquial en diferentes medios. Este corpus se caracteriza por su rigor metodológico: los textos están cuidadosamente anotados con metadatos relevantes, como el tipo de texto, el país de origen, el año, el medio, el soporte, la tipología textual o el área temática. En el caso de los textos orales, se incorporan además datos sociolingüísticos específicos, como el sexo, la edad, el nivel educativo y la profesión de los hablantes. Asimismo, los textos están marcados con información morfosintáctica, lo que facilita la búsqueda y el análisis de palabras, frases o estructuras gramaticales. Por otro lado, la calidad del CORPES XXI radica en su representatividad, su actualización

constante y su diseño equilibrado, que garantiza la fiabilidad de los datos. Este enfoque asegura una selección rigurosa de los datos para garantizar una distribución equilibrada en relación con países, zonas geográficas, tipologías textuales y áreas temáticas. Debido a su diseño, codificación y a la calidad de los datos que ofrece, hemos seleccionado este corpus lingüístico para la investigación de tres partículas discursivas del español, *es decir*, *o sea* y *en plan*, con el objetivo de determinar los contextos más favorables para su uso. Con este enfoque, además, buscamos demostrar la utilidad de los corpus como herramientas para la investigación lingüística y, en particular, del habla coloquial.

## 3. Los reformuladores *es decir*, *o sea* y *en plan*

Como señala Portolés (1998), los marcadores del discurso no presentan la misma distribución en textos escritos y en conversaciones. De este modo, existen marcadores que son más frecuentes en el discurso oral y otros que, por el contrario, suelen limitarse al discurso escrito. Según este mismo autor, la diferencia principal en la distribución se debe al contexto del discurso oral y del discurso escrito. En la comunicación oral, los interlocutores tienen acceso directo a una abundante información contextual; mientras que el discurso escrito se caracteriza por una menor riqueza contextual, por lo que es necesario que se explicite lingüísticamente toda la información contextual necesaria para conseguir una comunicación efectiva (Portolés 1998). En el primer caso, se incluirían *es decir* y *o sea*, partículas vinculadas al lenguaje escrito; mientras que en el segundo caso, más propia de la oralidad, se encontraría *en plan* (Briz, Pons y Portolés 2008; Ciapuscio 2001). No obstante, como se ha señalado previamente, la aparición de material proveniente de Internet (blogs, redes sociales, etc.) ha dado lugar a nuevos géneros en los que el soporte escrito se combina frecuentemente con el lenguaje coloquial. En este sentido, será interesante investigar, a través de la documentación del corpus seleccionado, CORPES XXI, en qué contextos son comunes estos marcadores, aparentemente divididos en marcadores característicos del discurso escrito y/u oral.

En español, los conectores discursivos *es decir*, *o sea* y *en plan* se categorizan dentro del grupo de reformuladores, cuya función es introducir una nueva formulación en el discurso. La reformulación es una característica intrínsecamente vinculada a la oralidad conversacional. Esta se lleva a cabo, según Martín Zorraquino y Portolés (1999), de dos formas: reexpresando de manera más precisa lo mencionado anteriormente o expresando directamente las conclusiones que deberían derivarse del primer segmento, introduciendo un nuevo tema con un carácter conclusivo. En este sentido, aunque todos son clasificados como reformuladores, se evidencia un comportamiento contextual

diferenciado para *es decir*, *o sea* y *en plan*, dependiendo de la naturaleza del discurso o del texto, que será examinado detenidamente en el análisis de nuestro trabajo.

Los reformuladores explicativos introducen una idea que explica la anterior (*NGLE* § 30.13; Casado 1991; Portolés 1998). En concreto, existe una relación de equivalencia en distintos grados, lo que los hace intercambiables (Briz 2001). De manera particular, el marcador explicativo *es decir* se caracteriza principalmente por su función parafrástica, al expresar de manera diferente lo que se ha mencionado previamente, estableciendo una equivalencia entre ambas formulaciones. A través de esta operación reflexiva, el emisor ejerce un control metadiscursivo que demuestra su preocupación por el discurso y la comprensión del mismo (Portolés 1998; Núñez, Muñoz y Mihovilovic 2006).

(1) Por ello, los cabildos y el ejército ocuparían las funciones de gobierno y administración de las instituciones del Imperio Español, **es decir**, Capitán General, Real Audiencia, Intendencia de Real Hacienda y Gobernaciones de las distintas provincias (CORPES XXI, *Rebeliones, alzamientos y movimientos preindependentistas en Venezuela* [Ven.] 2001).

Otros autores como Ciapuscio (2001) defienden que *es decir* también participa en procedimientos de reformulación no parafrástica, implicando un cambio en la perspectiva enunciativa (García Negroni 2009).

(2) Si deseo ver obras de varios autores, en cada búsqueda tengo que cambiar la opción. **Es decir**, tengo que esforzarme el doble (CORPES XXI, *No me hagas pensar: Una aproximación a la usabilidad en la Web de Steve Krug* [Esp.] 2001).

Tiene un sentido parcialmente equivalente el conector *o sea*, cuando funciona como reformulador explicativo, que implica que el segmento donde aparece funciona como una explicación, clarificación o corrección de todo o parte de lo mencionado anteriormente.

(3) Lo primero que lleva la papa es una semilla apropiada y, desde luego, no le puede faltar la nutrición, que en nuestro caso lo hacemos con biofertilizantes, **o sea**, productos biológicos logrados aquí mismo en la finca (CORPES XXI, *La papa, bien atendida, vale por dos* [Cuba] 2024).

El siguiente, y último, reformulador es *en plan*, que se posiciona como una alternativa a otros marcadores que se utilizan en discursos

más formales, como el ya mencionado *es decir*. Las funciones atribuidas a *en plan*, según los estudios existentes, incluyen el papel de atenuador, intensificador, reformulador, ejemplificador, introductor de estilo referido y prolongador (Méndez Orense 2016; Borreguero Zuloaga 2020). Para nuestro estudio comparativo con los marcadores *es decir* y *o sea*, nos enfocaremos en su función como reformulador explicativo. En este contexto, *en plan* actúa principalmente como un conector entre diferentes partes del discurso, facilitando la interpretación del mensaje por parte del interlocutor (Briz 2001). Al desempeñar el papel de reformulador explicativo, algunas autoras como Méndez Orense (2016: 135-136), plantean la posibilidad de que *en plan* haya empezado a desplazar a *o sea* como el reformulador más frecuente en el lenguaje coloquial. No obstante, *en plan* se distingue de *o sea* en que el primero de ellos funciona exclusivamente como reformulador explicativo y no como rectificativo, como ocurre con *o sea*.

Con el fin de investigar los contextos textuales y la extensión del uso de estos tres reformuladores con valores explicativos, ya sea en términos tipológicos y temáticos, procedemos a analizar el uso documentado de *en plan*, *o sea* y *es decir* en el corpus, y subcorpus, seleccionado para nuestra investigación: el CORPES XXI.

## 4. Resultados y análisis de *en plan*, *o sea* y *es decir* en el CORPES XXI

El análisis llevado a cabo a través de la documentación de corpus se ha consolidado como una metodología clave en la investigación lingüística. Este enfoque permite no solo identificar y catalogar patrones recurrentes en el uso del lenguaje, sino también entender las circunstancias y contextos en los que tales fórmulas son más propensas a surgir. En este apartado, se llevará a cabo un análisis exhaustivo de *es decir*, *o sea* y *en plan*, empleando documentación procedente de corpus, con el objetivo de establecer su naturaleza y determinar los contextos lingüísticos en los que se manifiestan con mayor frecuencia estos marcadores. Con el fin de alcanzar este objetivo establecido, se empleará como fuente de datos el CORPES XXI. Este corpus es un corpus de referencia con 25 millones de formas por año y una distribución de textos del 70 % provenientes de América y del 30 % de España. Los textos se seleccionan conforme a criterios específicos: el 90 % corresponde a la lengua escrita y el 10 % a la oral. Los materiales escritos se dividen en libros (40 %), publicaciones periódicas (40 %), material de Internet (7,5 %) y miscelánea (2,5 %). Geográficamente, asigna el 30 % de los textos a España y el 70 % a América, incluyendo por primera vez textos de Guinea Ecuatorial y Filipinas. La clasificación temática uniforme de los textos, organizada en ficción y no ficción, y su subdivisión en

áreas temáticas, junto con la categorización según género textual (novelas, relatos, noticias, reportajes, entrevistas, etc.), crea una tipología textual compleja que mejora la precisión en la recuperación de información para los investigadores. En resumen, el CORPES XXI es una herramienta representativa y equilibrada diseñada para el estudio del español contemporáneo, con una estructura meticulosa que garantiza la fiabilidad de los datos.

Atendiendo a la búsqueda de *es decir*, *o sea* y *en plan* en el CORPES XXI, encontramos la presencia de estos tres reformuladores en textos de distinta naturaleza: prensa y libros, orales e Internet. En particular, podemos hacer una distinción clave según el origen del texto: escrito u oral.

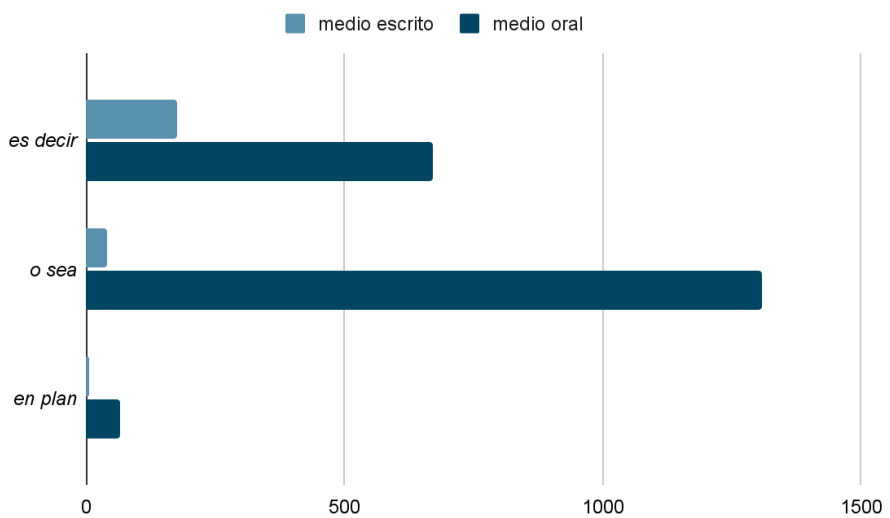



*Gráfico 1. Distribución de* en plan, es decir *y* o sea *según el medio.*

La información presentada en el Gráfico 1, y a continuación en la Tabla 1, muestra que estos reformuladores son característicos del lenguaje oral. Sin embargo, no podemos ignorar la presencia de estos elementos en textos escritos, donde, en algunos casos, se registra un número relevante de concordancias.

| | **Frecuencia normalizada** | |
|---|---|---|
| | **Escrito** | **Oral** |
| *Es decir* | 177,6 | 664,81 |
| *O sea* | 41,43 | 1 353,41 |
| *En plan* | 4,44 | 69,81 |

*Tabla 1. Distribución de* en plan, es decir *y* o sea *según el medio.*

Al analizar con detalle cada marcador, observamos que, en el caso de *es decir*, la mayoría de los ejemplos provenientes de textos orales, como el ejemplo 1a, corresponden a autores con un alto nivel educativo. Esto sugiere que, aunque aparezca en textos orales, su uso tiende a estar asociado a discursos más formales, alejados de la coloquialidad propia de la oralidad. Del mismo modo, podría considerarse que el tipo de discurso oral puede ejercer influencia, dado que en el ámbito radiofónico, es común encontrar textos oralizados. En cuanto a los ejemplos de textos escritos, se pueden distinguir casos procedentes de Internet, como el ejemplo 3b, que proviene de un blog de cocina de una marca conocida, donde se infiere un uso más cuidado del lenguaje. Esta tendencia se ha observado en los últimos años en textos de Internet, como blogs, tuits o cuentas de Instagram, los cuales pertenecen a instituciones o cuentas verificadas con una importante cantidad de seguidores, donde el lenguaje es más cuidado.

(3) a. eso no quiere decir que no haya méritos, **es decir**, evidentemente, el mérito tiene que estar y a partir de ahí (CORPES XXI, *Cadena SER:Las entrevistas de Aimar | Baltasar Garzón*, 2022).
b. Siempre debemos bridar el pavo, **es decir**, amarrar las piernas y acomodar los alones para que no pierdan forma, así mismo, cubrir la orilla de las piernas para evitar que se quemen (CORPES XXI, *Thermomix Blog - Thermomix México*, 2023).

En particular, y en línea con el objetivo de esta investigación, los datos del CORPES XXI muestran una frecuencia normalizada mayor en textos orales (664,81 por millón) que en textos escritos (177,6 por millón). Como se observa en la Tabla 2 y la Tabla 3, en las cuales se presentan los tres casos tipológicos más frecuentes, el marcador *es decir* es característico de textos escritos formales y de textos orales en los que se emplea un lenguaje más formal y elaborado, mostrando características similares a las de los textos escritos.

| Tipología textual | Frecuencia absoluta | Frecuencia normalizada |
|---|---|---|
| Académico | 11 670 | 379,84 |
| Manual de instrucciones | 22 | 365,05 |
| Libro de texto | 122 | 333,16 |

*Tabla 2. Distribución de* es decir *en textos escritos según la tipología textual.*

| Tipología textual | Frecuencia absoluta | Frecuencia normalizada |
|---|---|---|
| Tertulia | 249 | 1 175,65 |
| Debate | 101 | 823,83 |
| Discurso | 40 | 736,87 |

*Tabla 3. Distribución de* es decir *en textos orales según la tipología textual.*

Algo similar ocurre con *o sea*, que se documenta con frecuencia en textos escritos, pero sobre todo orales. Así, el CORPES XXI evidencia una frecuencia normalizada de 1 353,41 por millón en textos orales y 41,43 por millón en textos escritos. En términos generales, este marcador se ha catalogado dentro del lenguaje hablado; sin embargo, su presencia en textos escritos proporciona información significativa. Se observa que en la mayoría de los ejemplos orales, así como en muchos de los escritos, se utiliza un lenguaje coloquial característico de la oralidad. Por ejemplo, en 4a se presenta una entrevista transcrita con un tono relajado y menos formal, y en 4b, un texto extraído de Internet muestra el reformulador en una entrevista de un blog, en la que el lenguaje se acerca a rasgos propios de la oralidad. Siguiendo a Brown y Gillman (1960), en textos orales en los que se da una situación en la que no existen relaciones de poder y sí de solidaridad entre los conversadores, y en la que hay además afinidad generacional y, quizá también, de sexo, se propicia el habla coloquial. Por consiguiente, en el ejemplo 4a la relación entre iguales, en este caso mujeres de edades similares, favorece la coloquialidad.

(4) a. [...] esto ni me lo creo de que pude hacer yo ese trabajo / eehh / **o sea** no / **o sea** nunca nunca soñé que íbamos a poder a a lograr tantas cosas (CORPES XXI, *A solas: Sofía Vergara y Vicky Martín Berrocal | A SOLAS CON: Capítulo 11 | Podium Podcast*, 2024).

b. Con Armando fue muy curioso, me acuerdo que dije "quiero trabajar con Armando", **o sea** yo quería hacer la obra con alguien con quien yo quisiera trabajar, hice una lista de tres actores con el perfil de personaje con los que quería trabajar y lancé mi primera carta, que era Armando (CORPES XXI, *El blog de FilminLatino*, 2023).

De esta manera, a partir de la documentación extraída del CORPES XXI y tal como se detalla en la Tabla 4, se concluye que *o sea* aparece frecuentemente en contextos comunicativos orales espontáneos, donde este marcador discursivo facilita al hablante iniciar sus intervenciones y organizar la estructura del discurso.

| Tipología textual | Frecuencia absoluta | Frecuencia normalizada |
|---|---|---|
| Entrevista | 5 139 | 3 099,27 |
| Tertulia | 520 | 2 455,18 |
| Sorteos y concursos | 4 | 1 416,43 |

*Tabla 4. Distribución de* o sea *en textos orales según la tipología textual.*

En cuanto a *en plan*, este marcador se diferencia notablemente de los contextos formales. Tal como se muestra en la Gráfica 1, se trata de una partícula predominantemente oral, aunque también se encuentra en textos escritos. En este sentido, el CORPES XXI registra una frecuencia normalizada de 4,44 por millón en textos escritos y 69,81 por millón en textos orales. Teniendo en cuenta su uso reciente y la cantidad de documentación ya recopilada en los corpus, es posible que este reformulador sea considerado uno de los más frecuente en el lenguaje coloquial (v. Tabla 5).

| Tipología textual | Frecuencia absoluta | Frecuencia normalizada |
|---|---|---|
| Mensaje en redes sociales | 2 | 32,97 |
| Entrevista | 363 | 28,69 |
| Tertulia | 4 | 19,04 |

*Tabla 5. Distribución de* en plan *según la tipología textual.*

Además de encontrarse en textos orales (5a), *en plan* aparece en textos escritos como prensa o libros, como se observa en 5b.

(5) a. normalmente este este tipo de personas tiene doble cara lo dice con tanta seguridad / y tanta frialdad que hasta yo misma dudo de mí misma y digo **en plan** puede ser que yo (CORPES XXI, *Luc Loren | Somos Estupendas: "LO PEOR FUE EL MALTRATO PSICOLÓGICO" - Jessica Goicoechea #NoEstamosLocas*, 2023).
b. Te he traído una película —dice—. El extraño caso de Angélica, de Manoel de Oliveira. No la has visto, ¿verdad? —Dolores niega con la cabeza—. Da un poco de bajón, pero me ha recordado un poco a lo que está pasando, **en plan** un fotógrafo al que lo lían para unas fotos de difuntos. Lo que pasa es que el tío luego se enamora de la muerta (CORPES XXI, *Anoxia*, 2023).

En este contexto, lo más notable es la presencia de estos marcadores, que inicialmente se consideran propios de la lengua hablada, en textos escritos. Por lo tanto, resulta pertinente examinar su comportamiento en este medio, donde, *a priori*, se presume la existencia de un lenguaje más elaborado y formal. De acuerdo con las observaciones realizadas, procedemos a clasificar los datos recuperados del CORPES

XXI primero según la tipología textual (Gráfico 2) y, en segundo lugar, según la temática del texto (Gráfico 3). De acuerdo con la exposición de este estudio, se observa un mayor número de casos de *es decir* en textos de carácter formal con un estilo más elaborado, tales como textos académicos, libros de texto, noticias, reportajes, opiniones y prospectos, donde existe una planificación y se evita la espontaneidad. Entre *es decir* y *en plan*, se sitúa el reformulador *o sea*, que combina características de los dos extremos del espectro formal-coloquial. En textos escritos, este marcador se emplea habitualmente con un tono formal, en contextos donde el lenguaje es cuidado y planificado; mientras que, en textos orales o provenientes de Internet, parece favorecer un tono informal, la inmediatez, el diálogo y la espontaneidad, rasgos propios de la conversación. Por otro lado, *en plan* prevalece en textos de carácter informal, donde el lenguaje se aproxima más a la oralidad, como entrevistas, divulgación, blogs y redes sociales. Según Briz (1995, 2001), existen rasgos situacionales o coloquializadores que promueven el uso del registro coloquial. En relación con esto, en los ejemplos extraídos de la documentación del CORPES XXI, encontramos que factores como la igualdad entre los interlocutores, la relación de proximidad de los mismos y el marco discursivo familiar, determinado por el espacio físico y la relación específica de los participantes con ese espacio, entre otros, influyen en la preferencia por este registro.

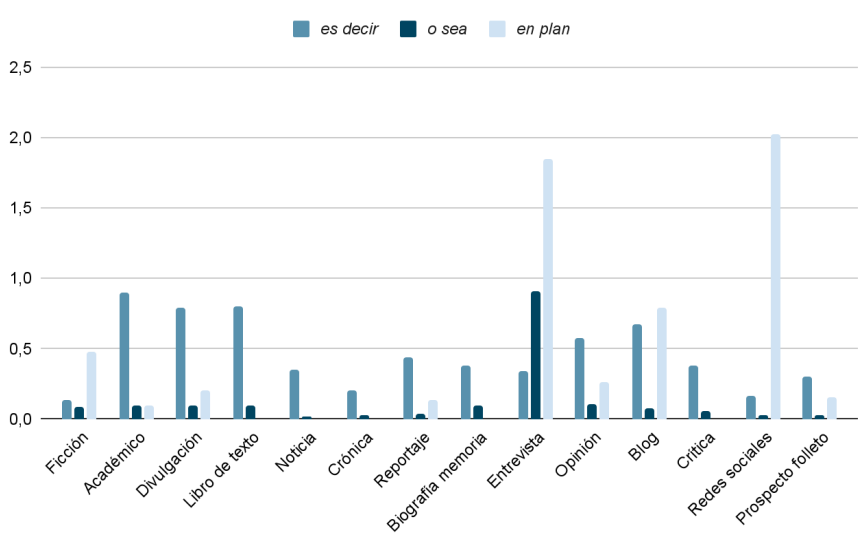



*Gráfico 2. Distribución de* en plan, es decir *y* o sea *según la tipología textual.*

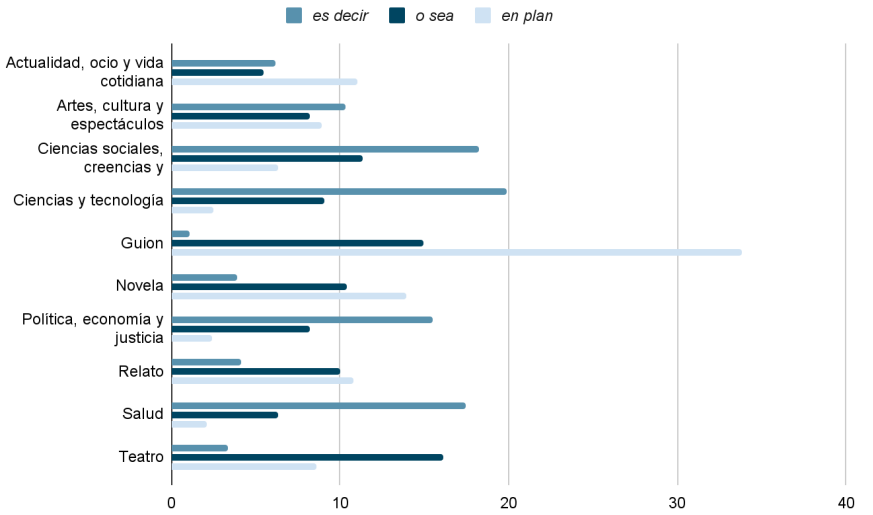



*Gráfico 3. Distribución de* en plan*,* es decir *y* o sea *según la clasificación temática.*

Considerando la información presentada en el Gráfico 2 y la información subsecuente del Gráfico 3, resulta notable la documentación proveniente de bloques como ficción para *en plan* y, en menor medida, *o sea*. Según lo planteado, en este tipo de documentación textual debería prevalecer el uso del reformulador *es decir*; sin embargo, al revisar los ejemplos, se descubre que en ficción los casos de *en plan* son, en realidad, casos en discursos que imitan la oralidad, ya que la mayoría aparecen en diálogos que simulan conversaciones entre personajes. En este tipo de textos, es fundamental la temática no especializada, la cotidianidad pretendida, la aparente ausencia de planificación, el tono informal, el dinamismo y el diálogo. Por lo tanto, este tipo de contenido textual puede ser considerado como un escrito coloquial que, en realidad, encubre una oralidad coloquial.

En este punto, resulta pertinente clasificar la información sobre estos tres marcadores, de acuerdo con los datos del corpus, en categorías de *+formal* o *-formal*, considerando las características típicas de cada tipo de texto. Con ello, se busca identificar qué reformulador predomina en cada contexto textual para determinar los rasgos distintivos de cada marcador y los tipos de textos en los que se emplean, según la información proporcionada por el CORPES XXI.

En este sentido, la Tabla 6 muestra, a través de la frecuencia normalizada —también presentada en el Gráfico 2—, que en textos generalmente formales (académicos, de divulgación, libros de texto, noticias, crónicas, reportajes, biografías, textos de opinión, críticas o prospectos) el marcador discursivo *es decir* es el más frecuentemente utilizado, marcando una considerable diferencia respecto a los otros

dos marcadores, *o sea* y *en plan*. En contraste, en textos de ficción, donde se incluyen diálogos, entrevistas, blogs o textos provenientes de redes sociales, el marcador *en plan* es el más utilizado, superando al marcador *es decir*. No obstante, se observa una excepción en los blogs, donde *es decir* también está presente en una proporción significativa. Esto se debe a que, en los blogs, las características de la escritura y la oralidad tienden a difuminarse. Así, se identifican blogs de marcas comerciales, instituciones o periódicos nacionales que se adhieren a normas de redacción establecidas por sus respectivos manuales de estilo; en contraste, también existen blogs que emplean un lenguaje más cercano a la oralidad y a la coloquialidad, lo que facilita el predominio del marcador *en plan*.

| **Tipología textual** | **+formal** | **-formal** | **Marcadores discursivos** | | |
|---|---|---|---|---|---|
| | | | *es decir* | *o sea* | *en plan* |
| Ficción | + | + | 0,14 | 0,09 | 0,48 |
| Académico | + | - | 0,9 | 0,1 | 0,1 |
| Divulgación | + | - | 0,79 | 0,1 | 0,2 |
| Libro de texto | + | - | 0,8 | 0,1 | 0 |
| Noticia | + | - | 0,35 | 0,02 | 0 |
| Crónica | + | - | 0,2 | 0,03 | 0 |
| Reportaje | + | - | 0,44 | 0,04 | 0,14 |
| Biografía memoria | + | - | 0,38 | 0,1 | 0 |
| Entrevista | + | + | 0,34 | 0,91 | 1,85 |
| Opinión | + | - | 0,58 | 0,11 | 0,26 |
| Blog | + | + | 0,67 | 0,08 | 0,79 |
| Crítica | + | - | 0,38 | 0,06 | 0 |
| Redes sociales | + | + | 0,17 | 0,03 | 2,02 |
| Prospecto folleto | + | - | 0,3 | 0,03 | 0,16 |

*Tabla 6. Distribución de* en plan, es decir *y* o sea *según una propuesta de clasificación textual.*

## 5. Conclusiones

A partir del estudio realizado, podemos concluir que es fundamental contar con datos empíricos y fiables que hayan sido previamente codificados siguiendo criterios establecidos. La recuperación de textos, tanto orales como escritos, a través del uso de corpus en general, y de subcorpus orales en particular, permite desarrollar una investigación capaz de generar conclusiones significativas en el análisis lingüístico de marcadores discursivos como *es decir*, *o sea* y *en plan*. Así, sin una codificación previa del corpus utilizado, en este caso el CORPES XXI, cuyo diseño y codificación consideran tanto la tipología textual como la temática, no habría sido posible alcanzar las conclusiones que se presentan a continuación.

Este artículo ha presentado, en primer lugar, un análisis basado en corpus sobre los marcadores discursivos previamente mencionados. Los resultados indican que, aunque los tres marcadores son típicos de la oralidad, existe una clara diferenciación en su uso según el medio y la tipología textual en la que se emplean. De acuerdo con los datos obtenidos, *es decir* se caracteriza principalmente por su presencia en textos orales y, en menor medida, en textos escritos. En ambos casos, eso sí, es propio de discursos más formales, incluso en textos orales, que parecen tener rasgos propios frente a los escritos. El marcador *o sea*, por su parte, ocupa una posición intermedia entre *es decir* y *en plan*, ya que su campo de uso no parece estar claramente delimitado. En general, su frecuencia normalizada es similar en los distintos tipos de textos analizados, aunque destaca en el género de la entrevista, donde su frecuencia es mayor, seguida por los textos de opinión, que tienden a ser menos coloquiales y más formales. En contraste, el marcador *en plan* tiene un campo de uso bien definido, predominando en textos coloquiales cercanos a la oralidad, como blogs, entrevistas, redes sociales y textos de ficción con diálogos.

En cuanto al uso de estos marcadores según la clasificación temática de los textos, se observa una notable presencia de *es decir* en textos de no ficción que abordan temas como Salud, Política, economía y justicia, y Ciencias y tecnología. Por otro lado, *en plan* se concentra mayoritariamente en textos de ficción, como guiones, novelas, relatos y obras teatrales, así como en textos de no ficción relacionados con temas de Actualidad, ocio, vida cotidiana, y Artes, cultura y espectáculos. Cabe destacar también la presencia de *o sea* tanto en textos con rasgos formales como en aquellos más coloquiales. Como se ha señalado en este estudio, este marcador parece compartir características de ambos registros, reflejando la delgada línea que existe entre ellos. Por ello, la aparición de nuevas tipologías textuales híbridas, que combinan elementos propios de distintos medios orales y escritos, dificulta la categorización de estos marcadores en una sola clasificación. Teniendo esto en cuenta, consideramos relevante destacar que la principal diferencia en la distribución de estos marcadores está relacionada con el contexto *+formal* o *-formal*, más que con la distinción entre discurso oral y escrito.

En resumen, este artículo ha tenido como propósito analizar la frecuencia de uso y los contextos en los que se emplean los marcadores *es decir*, *o sea* y *en plan*, presentando los datos proporcionados por el CORPES XXI en relación con su distribución tipológica y clasificación temática. Con ello, además, se ha enfatizado la relevancia de incorporar datos de corpus, los cuales aportan mayor fiabilidad a las investigaciones gracias a la calidad y cantidad de información disponible. Finalmente, este trabajo busca ser una invitación a utilizar estos

datos como base para futuras investigaciones que profundicen en el comportamiento de los marcadores discursivos *es decir*, *o sea* y *en plan* teniendo en cuenta otras variables.

## Bibliografía

AMERESCO = Albelda, Marta, y Estellés, María (dirs.), *Corpus Ameresco*. Disponible en: https://corpusameresco.com.

Bolaños, Sergio (2015), «La lingüística de corpus: perspectivas para la investigación lingüística contemporánea», *Forma y Función*, 28 (1): 31-54. DOI: 10.15446/fyf.v28n1.49564.

Borreguero Zuloaga, Margarita (2020), «Los marcadores de aproximación (en el lenguaje juvenil): esp. *en plan* vs. it. *tipo*», en Miguel Ángel Cuevas Gómez, Fernando Molina Castillo y Paolo Silvestri (coords.), *España e Italia: un viaje de ida y vuelta: studia in honorem Manuel Carrera Díaz*, Sevilla, Editorial de la Universidad de Sevilla: 53-78.

Briz Gómez, Antonio (1995), «La conversación coloquial (materiales para su estudio)», en Antonio Briz (coord.), *Anejo XVI de la Revista Cuadernos de Filología*, Valencia, Universitat de València: 239-240.

Briz Gómez, Antonio (2001), *El español coloquial en la conversación: esbozo de pragmagramática*, Barcelona, Ariel.

Briz, Antonio (2002), «Apuntes para la definición lexicográfica de *o sea*», en Bernhard Pöll y Franz Rainer (eds.), *Vocabula et vacabularia: études de lexicologie et de (méta)lexicographie romanes en l´honneur du 60e anniversaire de Dieter Messner*, Frankfurt am Main, Peter Lang: 45-52.

Briz, Antonio, Salvador Pons, y José Portolés (coords.) (2008), *Diccionario de partículas discursivas del español*, [versión en línea]. Disponible en: www.dpde.es.

Briz, Antonio (2012), «Los déficits de los corpus orales del español (y de algunos análisis)», en Tomás Eduardo Jiménez Juliá *et al.* (coords.), *Cum corde et in nova grammatica: estudios ofrecidos a Guillermo Rojo*, Santiago, Universidad de Santiago de Compostela: 115-137.

Briz, Antonio y Andrea Carcelén Guerrero (2019), «El futuro iberoamericano del español: la investigación del español oral y en español», en Instituto Cervantes (ed.), *El español en el mundo*:

a*nuario del Instituto Cervantes*, Madrid, Bala Perdida/Instituto Cervantes: 189-217.

Brown, Roger, y Albert Gilman (1960), «The pronouns of power and solidarity», en Pier Paolo Giglioli (ed.), *Language as social context: selected readings*, Nueva York, Penguin Brooks: 252-281.

Camargo, Laura, y Ana María Grimalt (2022), «Nuevas y viejas funciones de *en plan*: estudio microdiacrónico en corpus orales y digitales del castellano de Mallorca en el siglo xxi», *Revista de Investigación Lingüística*, 25, e537931. DOI: 10.6018/ril.537931.

Casado Velarde, Manuel (1991), «Los operadores discursivos *es decir*, *esto es*, *o sea* y *a saber* en español actual: valores de lengua y funciones textuales», *Lingüística Española Actual*, 13 (1): 87-116.

Ciapuscio, Guiomar Elena (2001), «Los conectores reformulativos: el caso de es decir», en Elvira Narvaja de Arnoux y Ángela Di Tullio (comps.), *Homenaje a Ofelia Kovacci*, Eudeba: 157-172.

CORPES XXI = Real Academia Española, *Corpus del Español del Siglo XXI (CORPES)*. Disponible en: https://www.rae.es/corpes. [Fecha de consulta: 8 de junio de 2024].

COSER = Fernández-Ordóñez, Inés (dir.), *Corpus Oral y Sonoro del Español Rural (COSER)*. Disponible en: https://coser.lllf.uam.es.

Costa Outeiro, Susana Luisa (2021), *Los marcadores del discurso* en plan, o sea *y* rollo *en WhatsApp: funciones textuales*, trabajo de Fin de Grado, Universidade da Coruña.

COVJA = *Corpus oral de la variedad juvenil universitaria del español de Alicante*; integrado en el *Corpus oral para el estudio del lenguaje juvenil y del español hablado en Alicante*.

CREA = Real Academia Española, *Corpus de Referencia del Español Actual (CREA)*. Disponible en: https://www.rae.es/crea.

ESLORA = *Corpus para el estudio del español oral*. Disponible en: http://eslora.usc.es.

Garcés, María Pilar (2008), *La organización del discurso: marcadores de ordenación y de reformulación*, Iberoamericana Vervuert.

García Negroni, María Marta (2009), «Reformulación parafrástica y no parafrástica y *ethos* discursivo en la escritura académica en español: contrastes entre escritura experta y escritura universitaria avanzada», *Letras de Hoje*, 44 (1): 46-56.

Martín Zorraquino, María y José Portolés (1999), «Los marcadores del discurso», en Ignacio Bosque y Violeta Demonte (coords.), *Gramática descriptiva de la lengua española*, Madrid, Espasa Calpe: 4052-4213.

Méndez Orense, María (2016), «Valores pragmático-discursivos de la construcción lingüística *en plan*: ¿formación de un nuevo marcador?», *Philologia Hispalensis*, 30 (1): 123-144.

*NGLE* = RAE y ASALE (2009), *Nueva gramática de la lengua española*, Madrid, Espasa.

Núñez Paulina, Astrid Muñoz, y Estenka Mihovilovic (2006), «Las funciones de los marcadores de reformulación en el discurso académico en formación», *Revista Signos*, 39 (62): 471-492.

Portolés, José (1998), *Marcadores del discurso*, Barcelona, Ariel.

PRESEEA = *Proyecto para el estudio sociolingüístico del español de España y América*. Disponible en: https://preseea.linguas.net.

Pons Bordería, Salvador (2013), «Un solo tipo de reformulación», *Cuadernos AISPI*, 2, 151-170.

Pons Bordería, Salvador (2016), «Historia de *o sea*», *Boletín de la Real Academia Española*, 96 (313): 5-30.

Rojo, Guillermo (2015), «Los corpus textuales del español», en Javier Gutiérrez Rexach (coord.), *Enciclopedia lingüística hispánica*, Londres, Routledge: 285-296.

Rojo, Guillermo (2021), *Introducción a la lingüística de corpus en español*, Londres, Routledge.

Sinclair, John McHardy (1991), *Corpus, concordance, collocation*, Oxford, Oxford University Press.

Sinclair, John McHardy (2004), *Trust the text: language, corpus and discourse*, Londres, Routledge.

Teubert, Wolfgang, y Anna Cermáková (2004), *Corpus linguistics: a short introduction*, London, Continuum.

Val.Es.Co. = Pons Bordería, Salvador (dir.), *Corpus Val.Es.Co*. Disponible en: http://www.valesco.es.

Adrián Cabedo Nebot (ORCID: 0000-0002-3881-9308) y Noelia Ruano Piqueras (ORCID: 0000-0001-9513-9600), «Un acercamiento al estudio paralingüístico de las emociones en el corpus PRESEEA Valencia».
*Biblioteca de Babel: Revista de Filología Hispánica*, vol. extra. 2 (2024), pp. 213-237.
**DOI**: 10.15366/bibliotecababel2024.extra2.009 · **ISSNe**: 2695-6349 · Recibido: 05/09/2024 · Aceptado: 28/10/2024

# Un acercamiento al estudio paralingüístico de las emociones en el corpus PRESEEA Valencia

Adrián Cabedo Nebot
*Universitat de València*
adrian.cabedo@uv.es

Noelia Ruano Piqueras
*Universitat de València*
noelia.ruano-piqueras@uv.es

**Resumen**: Este artículo presenta una propuesta de análisis de las emociones presentes en la conversación. El estudio destaca la importancia de la prosodia (tono, intensidad, velocidad) en la expresión emocional, revelando que ciertas combinaciones prosódicas están asociadas a emociones específicas como alegría, tristeza, enfado o miedo. El análisis se centra en entrevistas sociolingüísticas obtenidas del corpus PRESEEA-Valencia y explora cómo estas emociones se manifiestan en el habla espontánea. Este trabajo apunta desafíos metodológicos en la identificación de emociones en el discurso espontáneo, como la importancia de considerar el contexto interactivo y cultural. Los resultados sugieren que la prosodia no solo comunica emociones, sino que también estructura las interacciones conversacionales. El estudio concluye que la prosodia es esencial para entender la relación entre emociones y discurso y sienta las bases para futuros análisis con mayor profundidad estadística.

**Palabras clave**: emociones, PRESEEA, Oralstats.

## An approach to the paralinguistic study of emotions in the PRESEEA Valencia corpus

**Abstract**: This paper presents an approach to analyzing emotions present in conversation. The study highlights the importance of prosody (tone, intensity, speed) in emotional expression, revealing that certain prosodic combinations are associated with specific emotions such as happiness, sadness, anger, or fear. The analysis focuses on sociolinguistic interviews obtained from the PRESEEA-Valencia corpus and explores how these emotions manifest in spontaneous speech.

This work points out methodological challenges in identifying emotions in spontaneous discourse, such as the importance of considering the interactive and cultural context. The results suggest that prosody not only communicates emotions but also structures conversational interactions. The study concludes that prosody is essential to understanding the relationship between emotions and discourse and lays the foundation for future analyses with greater statistical depth.

**Key words**: emotions, PRESEEA, Oralstats.

## 1. Introducción

El estudio de las emociones en el discurso ha despertado una relevancia creciente en la investigación sociolingüística, dado que permite entender no solo el contenido verbal de las expresiones orales de los hablantes a nivel interactivo, sino también la manera en la que las emociones influyen y moldean sus enunciados. En este contexto, consideramos que el corpus PRESEEA (Proyecto para el Estudio Sociolingüístico del Español de España y América, https://preseea.uah.es/, coordinado actualmente por Ana María Cestero y originalmente ideado y desarrollado por Francisco Moreno Fernández) se presenta como un instrumento útil para analizar la expresión emocional en contextos de habla real (Moreno Fernández 2021). Al mismo tiempo, este artículo se integra en el marco del proyecto de investigación Estudio de los condicionantes sociales del español actual en el centro y norte de España, nuevos retos, nuevas soluciones (ECOS-C/N) (PID2023-148371NB-C42), coordinado por Adrián Cabedo Nebot y Cristina Illamola Gómez y financiado por el Ministerio de Ciencia, Innovación y Universidades.

El corpus PRESEEA, compuesto por entrevistas sociolingüísticas, nos ofrece un caudal de datos orales muy interesante para explorar cómo los hablantes manifiestan sus emociones a través del lenguaje en situaciones espontáneas y de qué modo estas emociones se relacionan con variables sociales como la edad, el género y el nivel de instrucción. En el caso del presente estudio, presentamos una primera aproximación al estudio cualitativo de varias entrevistas del corpus PRESEEA-Valencia (Gómez Molina 2001, 2005, 2007 [en la actualidad coordinado por Adrián Cabedo Nebot, en la Universitat de València]). A diferencia de otros corpus de habla, que integran datos de grabaciones secretas y recogen interacciones con varios hablantes, hemos seleccionado el corpus PRESEEA para este estudio aproximativo porque nos permite analizar audios de mejor calidad y porque aboga por la naturalidad de

las interacciones del hablante con el entrevistador (Moreno Fernández 2021), dada la amplia duración de las grabaciones (en torno a 45-50 minutos por grabación).

En todo caso, debemos señalar que las emociones atraen actualmente la mirada de la comunidad científica por su vinculación ineludible con el día a día discursivo y con la propia personalidad de los hablantes. Como señala Plutchik (2001: 346): «emotions are behavioral homeostatic processes, playing a crucial role in regulating social interactions and ensuring the survival of the individual». En general, podemos destacar la importancia de conceptualizar las emociones no solo como estados internos, sino como cadenas de eventos que incluyen estímulos, cambios psicológicos y comportamientos dirigidos a objetivos específicos; la conducta humana ampara estas manifestaciones emocionales (Plutchik 2001; Bänziger y Scherer 2005; Cowie y Cornelius 2003). Entendemos, por tanto, que estudiar la relación entre emociones y prosodia puede devenir un hecho constructivo porque puede ayudarnos a entender la manera en la que las emociones contribuyen a la adaptación, supervivencia y convivencia discursivas; en tal sentido, las emociones son recursos generales y observables en la naturaleza, tanto en seres humanos como en otros animales, tal y como las entiende Plutchik (2001).

A nuestro juicio, el análisis de las emociones en el marco del proyecto PRESEEA genera diversos desafíos metodológicos, de los que destacan, entre otros, la identificación y categorización de las emociones en el habla espontánea; cuestión complicada en cualquier entorno de habla, no únicamente el espontáneo (Garrido Almiñana y Chica Sabariego 2018). A diferencia de los estudios experimentales donde las emociones son inducidas o simuladas, como señalábamos previamente, en un corpus de entrevistas sociolingüísticas las emociones emergen de manera natural y están vinculadas con las dinámicas del discurso y las características individuales de los hablantes. Esto requiere que utilicemos un método de estudio que no solo sea dependiente de variaciones prosódicas y morfosintácticas o léxicas, sino que también considere el contexto interactivo en el que se produce el discurso (Hidalgo 2020a, 2020b).

Además, este tipo de análisis, inicialmente de base cualitativa y valorativa, nos permite examinar cómo los hablantes utilizan diferentes recursos lingüísticos para expresar emociones en función de su contexto social y, en futuros estudios, podremos acercarnos con mayor propiedad a una variación potencial incluso entre diferentes comunidades de habla. En relación con los recursos lingüísticos que puede utilizar el hablante, el estudio de las inflexiones prosódicas, la intensidad, el rango tonal, las pausas y la elección de determinadas palabras

o frases puede revelar patrones en la manera en que se comunican emociones como la alegría, la tristeza, la ira o la sorpresa (Quilis 1993; Chafe 2002). Estudios clásicos de la entonación ya definieron algunas relaciones significativas entre el carácter o emoción del hablante y el recurso prosódico empleado. Por ejemplo, Navarro Tomás (1960: 210) indicaba que la alegría y la cólera generan inflexiones más variadas y tonos más agudos, mientras que la tristeza y el abatimiento se expresan con entonaciones bajas y monótonas; también señalaba, entre otros valores, que el carácter animado de la interacción discursiva produce una prosodia más dinámica, a diferencia del carácter apático, que se asocia con monotonía, o que los niños y las personas exaltadas muestran inflexiones más amplias y dramáticas que ancianos o, también, que personas melancólicas.

## 2. Sobre emociones y su expresión prosódica

En general, todos los listados de emociones que hemos observado en la bibliografía, *grosso modo*, incluyen emociones como la alegría, la sorpresa, el enfado, la tristeza o el miedo. Plutchik (1980, 2001) sumaba a esta lista emociones como la aversión, la confianza y la anticipación y las desglosaba todas en múltiples matices emocionales, tales como serenidad, júbilo, admiración, horror, tedio, aceptación, odio, pena, etc.

A partir de su configuración y taxonomía, reconocer las emociones, prestando especial atención al comportamiento verbal y paralingüístico del hablante, parece ser un hecho inevitable para una correcta determinación emocional (Padilla 2020, 2022; Hidalgo 2020a, 2020b; Schuller y Batliner 2014). Concretamente, podemos observar que el papel de la prosodia en el reconocimiento de emociones ha sido señalado por diferentes autores en los últimos tiempos; generalmente, se han destacado como indicadores emocionales en el habla las variaciones en el tono, la duración, la intensidad y los patrones de entonación (Navarro Tomás 1960; Quilis 1993; Cao *et al.* 2014; Schuller y Batliner 2014). A partir de la bibliografía, podemos especificar varios parámetros prosódicos que parecen relacionarse con emociones de manera tanto general como individualizada.

En uno de los primeros estudios sistemáticos sobre el tema, Mozziconacci (1995) identifica que la ira se asocia típicamente con un gran aumento en el tono (F0) y un rango tonal ampliado, mientras que la tristeza se caracteriza por un descenso en el tono y un rango tonal estrecho. Por otro lado, la alegría se distingue, en modo similar a la ira, por un tono alto y un amplio rango tonal. Estos hallazgos nos permiten determinar la importancia del tono en la diferenciación de estos estados emocionales, aunque Mozziconacci (1995: 181) insiste en

que: «it is clear that high recognition performance of emotions cannot be obtained through pitch manipulation only, and that other aspects such as duration and voice quality must also be taken into consideration». Ampliando la perspectiva anterior, Mozziconacci y Hermes (1999) destacan el papel de los patrones de entonación específicos en la modulación de la percepción emocional. Por ejemplo, demuestran que «specific intonation patterns, such as the final '3C' and '12', can introduce a perceptual bias towards a particular emotion, suggesting that intonation not only conveys linguistic information but also modulates the emotional interpretation of the message» (Mozziconacci y Hermes, 1999: 2001). En este modelo, el patrón 3 se define como «a late prominence-lending rise » y «C» como «a very late non-prominence-lending fall»; por su parte, «1» sería «an early prominence-lending rise» y «2», «a very late non-prominence-lending rise». Los resultados apuntan hacia una relación entre el patrón ascendente-descendente de 3C y valores como la tristeza o el aburrimiento (aunque también alegría), mientras que la inflexión melódica ligeramente ascendente de 12 apuntaría a transmitir emociones como miedo e indignación. Los estudios de Mozziconacci y Hermes (1999) superan la concepción paralingüística de la entonación y se centran en patrones entonacionales concretos, habitualmente emparentados con valores comunes en el discurso, como la afirmación, la negación o la interrogación.

En línea similar, Padilla (2020) aborda la prosodia emocional en la conversación espontánea, proponiendo un protocolo para la identificación perceptiva de las emociones a partir de parámetros acústicos como la frecuencia fundamental (F0), la intensidad y la velocidad del habla. El estudio demuestra que emociones con alta excitación, como la alegría, tienden a aumentar la F0, mientras que emociones como la tristeza la disminuyen, destacando la importancia de la prosodia en la interpretación emocional. En Padilla (2022), el autor amplía su análisis, explorando el modo en el que la prosodia actúa como una reacción emocional en el discurso, enfatizando la influencia cultural en la expresión prosódica de emociones. Además, este estudio incluye un análisis estadístico detallado que cuantifica la relación entre diversos patrones prosódicos y las emociones expresadas. Para ello, Padilla (2022) utiliza un conjunto de datos provenientes de un corpus conversacional amplio y aplica pruebas estadísticas como las correlaciones de Pearson para determinar la significancia de las diferencias en los parámetros prosódicos según el tipo de emoción. Por ejemplo, el análisis revela que la variabilidad en la F0 y la intensidad es significativamente mayor en expresiones emocionales que presentan un perfil prosódico distintivo:

> hay tendencias con validación estadística que muestran que la F0 en las reacciones sintagmáticas sí obedece a un patrón. Hay un aumento de este valor en casi todas las reacciones emocionales descritas con el rasgo

> [+excitación], esto es, el enfado, la sorpresa y la alegría. No lo hay, sin embargo, en los ejemplos de asco, que presentan un comportamiento irregular atribuible a sus características especiales (ser un estadio intermedio entre la sorpresa y el enfado). Hay igualmente un descenso regular del valor de la F0 en las reacciones emocionales descritas con el rasgo [-excitación], es decir, la tristeza y el miedo. Es posible por tanto afirmar que sí hay algunas regularidades en la prosodia emocional (Padilla 2022: 14)

Finalmente, en Padilla (2023), el autor profundiza en cómo se construyen las emociones en la entonación coloquial, utilizando estudios de caso para ilustrar las variaciones prosódicas en distintos contextos conversacionales. Este artículo destaca por su enfoque en la entonación como un vehículo clave para la construcción emocional, respaldado por un análisis detallado de datos empíricos. La rigurosidad de todos estos estudios se manifiesta en la aplicación sistemática de metodologías mixtas, cuantitativas y cualitativas, que combinan análisis perceptivos y acústicos y un extenso análisis de datos, para ofrecer así una visión integral de la prosodia emocional.

Desde un acercamiento metodológico parecido, también basado en datos, Garrido Almiñana y Chica Sabariego (2018) exploran los marcadores prosódicos de emociones como la sorpresa, el miedo y la tristeza en el habla española; este estudio está basado en datos de laboratorio, en el que dos enunciados se pronunciaban según distintos valores emocionales por dos actores profesionales. Su investigación muestra que la sorpresa se indica a menudo mediante un aumento en el F0 y grandes rangos tonales, especialmente en movimientos tonales ascendentes-descendentes rápidos al final de los enunciados:

> global pitch range is mainly associated by listeners to specific emotions (ESCP hypothesis), as 'surprise' and 'joy', two emotions related to positive arousal but not to its highest values. However, 'anger', the emotional label associated to the highest arousal level in this paper, is not associated to maximum pitch levels. And 'disgust' and 'fear', associated usually to mid-levels of arousal, are not linked neither by listeners to high levels of global F0 range (Garrido y Chica 2018: 32).

También otros autores vinculan el comportamiento de la F0 con la expresión de algunas emociones concretas; así, Bänziger y Scherer (2005) ofrecen una descripción detallada de la prosodia específica para cada emoción, mostrando que

> For some emotional expressions— especially hot anger (HA anger), cold anger (LA anger), and elation (HA joy)—the second F0 excursion in the utterances tended to be larger than for other emotions—such as sadness (LA sad) or happiness (LA joy), which showed much smaller F0 excursions in the second part of the utterances (Bänziger y Scherer 2005: 265).

En un acercamiento distinto y más local de la entonación, Cao *et al.* (2014) proporcionan un análisis de los patrones tonales utilizados en la transmisión de las emociones; para ello, se sirven del marco de análisis de la entonación ToBI. Señalan que «anger and disgust utterances have higher occurrence of !H-L% and L- in the middle of utterance» y que el miedo y la felicidad «are characterized by a somewhat higher rate of !H-L% and H-L% boundary tones» (Cao *et al.* 2014: 133). La tristeza, en contraste, se asocia con «H-L% boundaries and less occurrence of down-stepped (!H*) pitch accents» (Cao *et al.* 2014: 134). El estudio también revela que las emociones con valencia negativa y alta excitación tienden a asociarse con enunciados que se diversifican más, «chunked into more prosodic units despite their relatively short duration» (Cao *et al.* 2014: 134). Estos hallazgos indican que las emociones pueden manifestarse en ciertos contornos tonales y tonos de frontera concretos.

A partir de un enfoque más general, Hidalgo (2020a) enfatiza la influencia de las normas culturales en el uso de la prosodia en la expresión emocional. Señala que «como código semiestable relacionado con la competencia sociocultural de un hablante; permite transmitir afectividad, pero no pertenece al idioma (sistema), sino a la comunidad de habla o comunidad cultural» (Hidalgo 2020a: 40), con lo que se añade otra capa de complejidad, la de la variación cultural, a la interpretación de marcas prosódicas de las emociones. De hecho, el mismo autor, en una publicación posterior (Hidalgo 2020b), indica que puede existir incluso un factor fisiológico en la expresión de algunas emociones concretas; esto podría llevarnos a señalar que el uso de un mayor tono e intensidad en emociones como la alegría «apunta a una mayor implicación de los músculos de la fonación en la expresión de emociones» (Hidalgo 2020b: 97).

Por todo lo visto en la bibliografía, podemos concluir que la exploración de las emociones a través de la prosodia en el habla revela que ciertos parámetros acústicos, como la frecuencia fundamental (F0), la intensidad y los patrones de entonación desempeñan un papel crucial en la expresión y reconocimiento de las emociones. Estudios como los de Padilla (2020, 2022, 2023) y Mozziconacci (1995, 1999) demuestran la importancia del tono, aunque también la velocidad de habla, para diferenciar emociones como la alegría, la tristeza, la ira y el miedo. Además, trabajos como los de Cao *et al.* (2014) utilizan marcos teóricos de la entonación, como ToBI, para caracterizar las inflexiones tonales asociadas a diferentes emociones. La influencia cultural y fisiológica en la prosodia emocional, destacada por autores como Hidalgo (2020a, 2020b), evidencia que la interpretación de las marcas prosódicas está profundamente entrelazada con la comunidad de habla y sus normas culturales, lo que plantea desafíos adicionales que deberán ser tenidos en cuenta en futuras investigaciones.

Por todo lo expuesto, consideramos que un corpus como PRESEEA, con su detallada anotación sociolingüística, ofrece un marco de estudio altamente adecuado para explorar la relación entre prosodia y emociones en el discurso. Este artículo se adhiere a esta dirección, pero el potencial del corpus para estudios futuros con mayor profundidad estadística es evidente y promete contribuir significativamente al ámbito de estudio.

## 3. Metodología de análisis

En la presente sección de metodología, detallaremos los procedimientos empleados para la evaluación cualitativa de los datos y la identificación de los parámetros acústicos clave para determinar las emociones expresadas en el discurso. Haremos uso del corpus PRESEEA-Valencia como ámbito de recogida de datos. Además, describiremos el proceso de procesamiento y análisis de estos datos utilizando la herramienta Oralstats, la cual permite una evaluación cuantitativa y sistemática de los parámetros prosódicos en las muestras de habla. Con esta metodología, pretendemos acometer un primer estudio de la prosodia emocional en un material de procedencia sociolingüística.

### *3.1. Evaluación cualitativa de los datos: PRESEEA-Valencia*

En esta sección, presentamos los datos analizados y la perspectiva cualitativa de análisis. Hemos realizado un análisis de seis entrevistas seleccionadas del corpus PRESEEA-Valencia (Gómez Molina 2001), correspondientes al nivel alto. Dado que este artículo presenta una primera aproximación al estudio emocional en un corpus de habla sociolingüístico, el acercamiento es cualitativo y pretende proponer una lista de correlaciones entre emociones y prosodia que, en futuras investigaciones y con mayor profundidad, puedan acometerse desde una perspectiva cuantitativa y con finalidad estadística de representación poblacional. Al mismo tiempo, no se ha efectuado ninguna prueba de coincidencia de jueces para la evaluación de las emociones, sino que se ha procedido a un estudio intuitivo del material sonoro recogido y, también, se ha consensuado con la catalogación en sentimientos realizada automáticamente mediante un diccionario de sentimientos (Mohammad y Turney 2013).

En el proyecto PRESEEA encontramos una iniciativa de investigación de gran envergadura que busca analizar la variación sociolingüística del español en diferentes comunidades de habla tanto en España como en América Latina (Moreno Fernández 2021). Este proyecto tiene como objetivo principal documentar, analizar y entender las variaciones en el uso del español, considerando factores como la edad, el género, el

nivel educativo, la ocupación y otros aspectos sociales y demográficos de los hablantes. Se trata de un proyecto colaborativo que involucra a numerosas universidades y centros de investigación, cada uno de los cuales se encarga de estudiar su propia comunidad de habla.

En ese ámbito de estudio, el corpus PRESEEA Valencia (Gómez Molina 2003, 2005, 2007) constituye uno de los corpus regionales que forman parte de este proyecto más amplio. Este corpus específico se centra en la variación sociolingüística del español en la ciudad de Valencia y área metropolitana. El corpus está diseñado para recopilar una muestra representativa del habla de los habitantes de Valencia y permite un análisis detallado de cómo se utiliza el español en esta región. El corpus PRESEEA Valencia fue inicialmente coordinado por José Ramón Gómez Molina hasta el año 2018. Desde 2018, el proyecto ha sido coordinado por Adrián Cabedo Nebot, quien ha continuado y expandido el trabajo iniciado previamente. El corpus incluye en el momento actual un total de más de 300 entrevistas que se encuentran actualmente en proceso de revisión.

Para realizar el estudio cualitativo y valorativo de emociones y prosodia, los archivos que hemos seleccionado, del nivel de instrucción alto (Gómez Molina 2001), han sido los siguientes:

— 1990_alto_01.mp3 (Mujer, 35-55 años)

— 1990_alto_02.mp3 (Hombre, 18-35 años)

— 1990_alto_03.mp3 (Hombre, 35-55 años)

— 1990_alto_04.mp3 (Hombre, 18-35 años)

— 1990_alto_05.mp3 (Mujer, 35-55 años)

— 1990_alto_06.mp3 (Mujer, 18-35 años)

En total, son casi seis horas de grabación donde los hablantes, tres hombres y tres mujeres, de edades comprendidas entre los 18 y los 55 años, se someten voluntariamente a una entrevista sociolingüística, que sigue además las pautas de recogida de datos establecidas en el marco del proyecto PRESEEA (Moreno Fernández 2021).

La elección de estas entrevistas para este estudio no ha seguido criterios de representación sociolingüística específica, como el equilibrio entre géneros, edades o contextos socioeconómicos. En lugar de ello, se ha priorizado un enfoque exploratorio para identificar patrones prosódicos que reflejen emociones diversas dentro de un rango controlado de edad y nivel educativo, como un primer paso hacia un futuro análisis más exhaustivo.

El objetivo de esta selección ha sido, por tanto, investigar la relación entre prosodia y emoción en el discurso, centrándonos en cómo se manifiestan diferentes emociones a través de elementos prosódicos como el comportamiento de la F0, la intensidad, la velocidad de habla y la inflexión tonal. Como señalábamos previamente, este análisis no pretende ser exhaustivo ni representativo de la población, sino que busca proporcionar un primer acercamiento a la complejidad de la expresión emocional en el habla y servir de una primera base que pueda ser extendida y mejorada, a nivel cuantitativo y estadístico, en próximos estudios. En este sentido, un modelo de análisis sistemático y al que aspiramos a seguir es el utilizado en los estudios de Padilla comentados previamente (sobre todo, el de 2022).

## *3.2. Parámetros acústicos para determinar la emoción*

La prosodia desempeña un papel crucial en la expresión y percepción de las emociones. A partir de los estudios revisados, se pueden establecer reglas de catalogación para identificar emociones específicas basadas en parámetros prosódicos como el rango tonal, el tono (*pitch* o F0), la intensidad y la velocidad de habla. A continuación, se presentan, de manera resumida, las reglas de catalogación para cada emoción (Garrido y Chica 2018; Cao *et al.* 2014; Padilla 2020, 2022; Hidalgo 2022a, 2022b):

| **Emoción** | **Rango tonal** | ***Pitch*** | **Intensidad** | **Velocidad** | **Características adicionales** |
|---|---|---|---|---|---|
| Ira | Amplio | Alto | Alta | Rápida | L+H*, !H-L% |
| Alegría | Amplio | Alto | Moderada a alta | Moderada a rápida | L+H* |
| Tristeza | Estrecho | Bajo | Baja | Lenta | H-L%, !H* |
| Sorpresa | Amplio | Alto | Moderada a alta | Variable | L+H* |
| Miedo | Variable | Alto | Variable | Rápida | L+H*, !H-L% |

*Tabla 1. Resumen de emociones y marcas prosódicas.*

Con la finalidad de que la visualización de los fragmentos de entrevistas que comentamos en este artículo sea más intuitiva, incorporamos un listado de emoticonos que transmiten valoración prosódica. Esta lista de emoticonos aparece en la versión 1.0 del programa Oralstats Furious (Cabedo 2023):

— 🦕: Velocidad baja

— 🚴: Velocidad alta

— ⬇️: Rango bajo

— ⬆️: Rango alto

— 🎷: Tono bajo

— 🎻: Tono alto

— 🔇: Intensidad baja

— 🔊: Intensidad alta

— ↘️: Inflexión descendente

— ↗️: Inflexión ascendente

— 🤏: Duración baja

— ⏳: Duración alta

### *3.3. Procesamiento de los datos: Oralstats*

Aunque la perspectiva de análisis es cualitativa, un primer tratamiento de los datos orales, tanto las transcripciones como los valores acústicos, ha sido realizados previamente con medios computacionales. Para esta tarea, hemos utilizado Oralstats (Cabedo 2022), herramienta digital gratuita programada con el lenguaje de programación R, desarrollada específicamente para el análisis cuantitativo y estadístico de datos lingüísticos provenientes de corpus orales, como los recopilados en proyectos sociolingüísticos como PRESEEA (Proyecto para el Estudio Sociolingüístico del Español de España y América). Esta herramienta permite a los investigadores realizar análisis detallados y sistemáticos de grandes volúmenes de datos orales; esto facilita el estudio de patrones de variación lingüística y otros fenómenos del habla. Las características del programa son las siguientes:

1. **Análisis estadístico**: Oralstats está diseñado para llevar a cabo una variedad de análisis estadísticos que son cruciales en estudios sociolingüísticos. Estos análisis incluyen la frecuencia de uso de ciertas formas lingüísticas, la distribución de variantes según factores sociales como la edad, el género, o el nivel educativo y la correlación entre distintas variables lingüísticas y sociales.

2. **Procesamiento de datos orales**: La herramienta está optimizada para manejar datos orales, que suelen ser más complejos y variados que los datos escritos debido a las características intrínsecas del habla, como la prosodia, las pausas y la variación en la pronunciación. Oralstats permite la codificación y análisis de estos elementos, proporcionando un enfoque cuantitativo a lo que tradicionalmente podría haber requerido un análisis más cualitativo y manual. En el caso del presente artículo, los datos prosódicos que se tienen en cuenta son aquellos que, en una escala estadística de valores Z, quedan por encima de 1,9 o por

debajo de -1,9 (Brezina 2018). Es decir, consideramos sobre todo los casos en los que los valores prosódicos están muy por encima o muy por debajo de los que manifiestan habitualmente los hablantes; esta técnica se ha demostrado efectiva para visualizar dinámicas conversacionales en situaciones de conflicto (Estellés 2023).

3. **Interfaz de usuario**: Oralstats cuenta con una interfaz que permite a los investigadores cargar y gestionar sus datos de manera eficiente. Aunque su principal función es el análisis cuantitativo, la interfaz está diseñada para ser accesible a investigadores que pueden no tener una formación profunda en estadística, facilitando su uso en estudios sociolingüísticos.

4. **Adaptabilidad y flexibilidad**: Oralstats es flexible y adaptable a diferentes tipos de estudios lingüísticos. Los investigadores pueden personalizar la herramienta según sus necesidades específicas, según sea el interés en estudiar fenómenos fonéticos, morfosintácticos, léxicos, o de cualquier otra índole en los datos orales. En este sentido, la elección del lenguaje de programación R facilita esta usabilidad.

5. **Visualización de datos**: Una de las ventajas de Oralstats es su capacidad para generar informes detallados y visualizaciones de datos, lo que facilita la interpretación de los resultados. Gráficos, tablas y otros recursos visuales pueden ser creados directamente en la plataforma.

## 4. Análisis y valoración de los fragmentos seleccionados

En esta sección de resultados, a partir de las seis entrevistas sociolingüísticas analizadas, se presenta un análisis detallado de cómo las emociones identificadas en el modelo de Plutchik (1980, 2001) se correlacionan con parámetros prosódicos específicos. Se han evaluado cuatro emociones clave: alegría, tristeza, enfado y miedo. A través del análisis de las entrevistas seleccionadas, hemos observado marcas prosódicas que aparecen recurrentemente en la expresión de algunas emociones en el discurso, con un enfoque en la relación entre el tono, el rango tonal, la inflexión melódica, la velocidad de habla y la intensidad de habla.

### *4.1. Alegría*

En el caso de la alegría, en las entrevistas hemos detectado generalmente una aparición recurrente con un tono más alto (🎻) y con rango tonal también alto (⬆️); estos mecanismos tonales pueden aparecer de forma independiente (ejemplos 1 y 2) o también de manera conjunta (ejemplo 3); en general, podemos decir de estos ejemplos que,

al manifestar alegría, enfatizan la implicación positiva y el interés del hablante por responder a preguntas que quizá no esperaban por parte del entrevistador y que vinculan en estos casos la alegría mostrada con una cierta carga de sorpresa.

Debemos recordar que en el proyecto PRESEEA los entrevistados conocen muy ligeramente la finalidad de la actividad en la que están participando, por lo que suelen verse sorprendidos muchas veces tanto por las preguntas que reciben como por sus propias respuestas, al tener estas que recoger pensamientos o vivencias que quedan en el pasado vivencial del hablante y que, en muchas ocasiones, pueden generar sorpresa y alegría al unísono. En definitiva, se genera un tono optimista y entusiasta:

(1) 00:35:17 - 1990_alto_05_e: si te tocara la lotería
00:35:19 - 1990_alto_05_i: ¿el domingo por ejemplo qué harías?
00:35:23 - 🎻 1990_alto_05_e: muchos millones muchos muchos muchos 🎻
00:35:25 - ↘️ 1990_alto_05_e: los que yo quiera ↘️
00:35:26 - 1990_alto_05_e: es que para eso necesitaría una primitiva
00:35:29 - 🎻 1990_alto_05_e: lo primero que haría 🎻
00:35:30 - 1990_alto_05_e: es
00:35:31 - 1990_alto_05_e: tomarme un año sabático
00:35:33 - 1990_alto_05_e: e irme a dar la vuelta al mundo
00:35:34 - 1990_alto_05_e: eso es lo primero que haría por eso
00:35:36 - 1990_alto_05_e: necesito muchos millones con 30 millones
00:35:39 - 1990_alto_05_e: no puedo hacer mucho
00:35:41 - ⬆️ 1990_alto_05_e: entonces sí que me gustaría el dar la vuelta al mundo ⬆️

En el ejemplo 1, la hablante se alegra ante la perspectiva introducida por la pregunta del entrevistador, en la que se alude a la posibilidad de que se gane el premio de la lotería. La combinación de tonos altos culmina con un rango tonal elevado en el que se señala que el deseo de la hablante es el de viajar por todo el mundo, para lo que se necesita una gran cantidad económica. Aunque no puede transmitirse adecuadamente en la transcripción, la audición de esta secuencia denota un cierto valor de énfasis mediante un elemento paralingüístico cercano a la risa, aunque solo acompaña a las palabras y no se articula todavía en risa completa o carcajada.

(2) 00:30:31 - 1990_alto_01_i: descríbeme tu piso como si fueras a vendérmelo
00:30:33 - ↗️ 1990_alto_01_e: ay ↗️
00:30:34 - 1990_alto_01_e: me encanta
00:30:35 - 1990_alto_01_e: y además mi piso no te lo vendo porque me encanta
00:30:37 - 🎻 1990_alto_01_e: no lo vendo 🎻
00:30:38 - 1990_alto_01_e: (RISAS) pero bueno
00:30:39 - 1990_alto_01_e: mira mi piso
00:30:40 - 1990_alto_01_e: es el típico que tiene pasillo pero a mí me gusta el pasillo

En el ejemplo 2, sucede algo similar al ejemplo 1, dado que el valor de la enunciación entre risas, o en un valor paralingüístico aproximado, aumenta la notoriedad tonal y confiere una tonalidad más aguda a las palabras de la hablante. Sin embargo, ese valor cercano a la risa culmina en el grupo de entonación posterior: *(RISAS) pero bueno*.

(3) 00:12:20 - 1990_alto_06_i: ¿la prepararías tú la cena?
00:12:21 - ↘️ 1990_alto_06_e: no ↘️
00:12:22 - 1990_alto_06_e: porque si no no cenaríamos
00:12:24 - 1990_alto_06_i: ¿no te gusta cocinar?
00:12:25 - ⬆️ 🎻 1990_alto_06_e: no es que no guste me es que no tengo arte para cocinar ⬆️ 🎻
00:12:29 - 1990_alto_06_e: (RISAS) tengo que practicar
00:12:30 - 1990_alto_06_e: algún día prepararé una cena pero de momento no (RISAS)

Nuevamente, el ejemplo 3 introduce un escenario similar a los ejemplos 1 y 2. En este caso, tanto tono alto como rango tonal amplio se acompañan con el mismo valor paralingüístico de la risa (en términos aproximados); al igual que en el ejemplo 2, la cercanía a la risa termina en risa final en el enunciado posterior *(RISAS) tengo que practicar*.

Por todo lo visto anteriormente, y más en el ámbito paralingüístico, podemos concluir que la cercanía a la risa es la que confiere una perceptibilidad mayor de los grupos de entonación en términos tonales.

## *4.2. Tristeza*

En el caso de la tristeza, se observa un uso recurrente de un tono bajo (🎷) y un rango tonal descendente (⬇️), lo que refleja un sentimiento de aceptación melancólica y una sensación de pérdida al narrar experiencias dolorosas o frustrantes. Estos patrones prosódicos,

como se muestra en los ejemplos 4 y 5, se asocian con una intensidad baja (🔇) y un tono bajo (🎷), sugiriendo un discurso reflexivo y controlado, especialmente cuando se abordan temas personales o sensibles. Este tipo de entonación destaca la carga emocional negativa en el discurso, creando un ambiente de confidencialidad y entrega de información cuidadosa.

(4) 00:26:41 - ⏳ 1990_alto_02_e: problema principal la de sociedad que es lo que está llevando a todos los demás problemas ⏳
00:26:46 - 1990_alto_02_e: es decir a la violencia
00:26:49 - 1990_alto_02_e: a la droga a
00:26:51 - 🎷 🔇 1990_alto_02_e: ver gente 🎷🔇
00:26:52 - ⬆️ 1990_alto_02_e: tirada por la calle ⬆️
00:26:55 - 1990_alto_02_e: a la pobreza
00:26:57 - 1990_alto_02_e: es el problema principal de
00:27:00 - 🔇 1990_alto_02_e: la sociedad 🔇

En el ejemplo 4, el hablante utiliza un tono grave (🎷) y baja intensidad (🔇) para expresar su percepción negativa sobre los problemas sociales. El uso de estas características prosódicas sugiere una entrega de información que resalta la gravedad y la falta de esperanza asociadas a los problemas mencionados, como la violencia, la droga y la pobreza. En ocasiones, la tristeza puede devenir en momento de enfado o desagrado, como evidencia en el ejemplo 4 el caso en el que hay un registro tonal elevado en *tirada por la calle*. La secuencia discursiva termina volviendo a bajar la intensidad en alusión a la sociedad.

(5) 00:29:56 - 1990_alto_01_e: yo muchas veces les digo mirad las cosas están muy mal
00:29:58 - 1990_alto_01_e: hay que ser el mejor
00:29:59 - 1990_alto_01_e: y tenéis que procurar ser el mejor posible
00:30:02 - ⏳ 1990_alto_01_e: cuanto mejor seáis mejor lo tendréis hay poco trabajo y va haber a para pocos ⏳
00:30:05 - 1990_alto_01_e: tenéis que ser de los mejores
00:30:07 - 1990_alto_01_e: tenéis que esforzaros por ser de los mejores
00:30:09 - 🎷 1990_alto_01_e: lo que ocurre es que todos no pueden ser de los mejores 🎷
00:30:12 - 🎷 1990_alto_01_e: también a lo mejor 🎷
00:30:14 - 🎷 1990_alto_01_e: orientarlos hacia el trabajo en equipo 🎷
00:30:18 - 🎷 1990_alto_01_e: en fin no que fueran solo tareas

competitivas 🎷
00:30:21 - ⬆️ ⏳ 1990_alto_01_e: a fin de cuentas individualistas y yo además la competencia la odio pero es que reconozco que está ahí ⬆️ ⏳
00:30:25 - 1990_alto_01_e: es que no pueden hacer otra cosa

En el ejemplo 5, la hablante, cuyo trabajo es se profesora de instituto, expresa su preocupación por el futuro de sus estudiantes en un mercado laboral cada vez más competitivo y difícil. Aunque reconoce que la competencia es una realidad inevitable, también sugiere la importancia de orientar a los estudiantes hacia el trabajo en equipo, una habilidad que podría ser igualmente valiosa en el futuro. El tono general de su mensaje es bajo (representada por el símbolo 🎷) y la melancolía transmitida se basa en la propia realidad personal de la profesora, ya que es consciente de que no todos sus estudiantes podrán sobresalir en el futuro en un entorno tan competitivo.

### *4.3. Enfado*

Las emociones de enojo y frustración se manifiestan en la prosodia, por lo que hemos observado en las entrevistas analizadas, a través de un tono alto (🎻) y una velocidad de habla rápida (🚴); esto refleja la gran carga emocional con la que el hablante busca enfatizar su punto de vista. Como se puede observar en los ejemplos 6, 7 y 8, estos patrones prosódicos son frecuentes situaciones de confrontación, donde el hablante desea comunicar su descontento o desacuerdo de manera vehemente. No se trata evidentemente de una discusión, dado que el hablante no confronta sus opiniones con el entrevistador, sino que manifiesta su desacuerdo con un posicionamiento habitualmente general o comúnmente compartido en la sociedad. Este distanciarse de otro punto de vista suele, por tanto, confrontar lo esperado por el hablante en un momento de su vida y la realidad encontrada (ejemplo 6) o lo que la manera de proceder de algún grupo en sociedad (la clase política, por ejemplo) y lo que el hablante considera que la sociedad realmente espera que, en este caso, suele alinearse con su pensamiento (ejemplo 7):

(6) 00:25:02 - 1990_alto_03_e: pero realmente por ejemplo volviendo al tema la biología de
00:25:05 - 🎻 1990_alto_03_e: yo pasé cinco años de biología y no tuve ni un solo 🎻
00:25:08 - 🎻 1990_alto_03_e: tema al respecto 🎻
00:25:09 - 🎻 1990_alto_03_e: ni uno solo 🎻
00:25:10 - 🎻 🔊 1990_alto_03_e: es más 🎻 🔊

00:25:11 - 🎻 1990_alto_03_e: si yo hubiera no estado convencido 🎻
00:25:13 - 1990_alto_03_e: que esta cuestión de era absolutamente vital
00:25:16 - 1990_alto_03_e: bueno es que hubiera sido imposible que lo hubiera desarrollado
00:25:19 - 1990_alto_03_e: yo y la desarrollé
00:25:20 - 🎻 1990_alto_03_e: como pude 🎻
00:25:21 - 1990_alto_03_e: a salto mata como podía y siempre
00:25:23 - 🎻 1990_alto_03_e: tenía esa perspectiva 🎻

En el ejemplo 6, el hablante usa un tono alto (🎻) y una intensidad marcada (🔊) para expresar su frustración por su experiencia académica; se enfatiza la importancia del tema tratado y su insatisfacción con la educación recibida.

(7) 00:20:09 - ⬆️ 🎻 1990_alto_04_e: veía justo que hubiese gente ⬆️ 🎻 ↗️
00:20:11 - ⬆️ 1990_alto_04_e: aunque estuviese fuera de la ley ⬆️
00:20:14 - ↘️ 1990_alto_04_e: que dijese ↘️
00:20:15 - 🚴 ↘️ 1990_alto_04_e: alto 🚴 ↘️
00:20:16 - 1990_alto_04_e: tú también puedes caer
00:20:18 - 🎻 1990_alto_04_e: y sobre todo porque exigen muchas cosas 🎻
00:20:22 - 🎻 1990_alto_04_e: exigen muchas cosas que ellos no cumplen 🎻
00:20:24 - 🎻 🔊 1990_alto_04_i: por qué tienen que exigir 🎻 🔊
00:20:27 - 🎻 🔊 1990_alto_04_i: que los presos vascos por ejemplo estén cerca del País Vasco 🎻 🔊
00:20:31 - ↘️ 1990_alto_04_e: no ↘️

En el ejemplo 7, el tono elevado (🎻), la inflexión tonal descendente marcada (↘️), una intensidad elevada (🔊) y el ritmo acelerado (🚴) refuerzan la crítica hacia las exigencias injustas, subrayando la inconformidad y la sensación de injusticia que percibe el hablante. Se trata de una disertación sobre los participantes en actos terroristas y su posterior comportamiento en actos jurídicos.

(8) 00:24:39 - 1990_alto_02_e: los problemas por ejemplo los problemas de política que hoy en día
00:24:43 - 1990_alto_02_e: se ven es la indecisión del gobierno

00:24:45 - ⬆️ 🎻 1990_alto_02_e: es decir ⬆️ 🎻
00:24:47 - 1990_alto_02_e: hoy digo una cosa mañana digo otra entonces
00:24:50 - 1990_alto_02_e: no es
00:24:51 - 1990_alto_02_e: la gente yo creo que está un poco
00:24:54 - 🚴 ↘️1990_alto_02_e: desconcertada 🚴 ↘️
00:24:56 - ⬆️ ⏳ 1990_alto_02_e: con los tipos de declaraciones que se están realizando desde el gobierno ⬆️ ⏳
00:25:01 - ⬆️ 1990_alto_02_e: es decir hoy ⬆️
00:25:02 - 1990_alto_02_e: es blanco mañana es negro
00:25:04 - 1990_alto_02_e: y pasado mañana es gris
00:25:06 - 1990_alto_02_e: entonces no sabe
00:25:08 - 1990_alto_02_e: en cierta medida por dónde va
00:25:10 - ⬆️ 🔇 1990_alto_02_e: es decir ⬆️ 🔇
00:25:11 - 1990_alto_02_e: vamos a ver
00:25:12 - ⬆️ 🎻 1990_alto_02_e: o esto o lo otro ⬆️ 🎻
00:25:14 - 1990_alto_02_e: quizás
00:25:15 - ⬆️ 1990_alto_02_e: hay algunos temas en que sí que se les ve ⬆️
00:25:18 - 1990_alto_02_e: con claridad la dirección que llevan
00:25:20 - 1990_alto_02_e: pero en la gran mayoría de los temas
00:25:24 - 🚴 1990_alto_02_e: no hay una concreción 🚴

El uso de las características prosódicas definidas en el ejemplo 7 también se manifiestan en el ejemplo 8, donde se discuten problemas de la clase política; se refuerza la idea de confusión y desconcierto y se transmite una fuerte insatisfacción con las decisiones de los gobiernos.

### *4.4. Miedo*

Debemos señalar que, en el marco de las entrevistas analizadas, la emoción del miedo ha sido la menos presente; no obstante, puede tratarse de una circunstancia eventual fruto de la fisiología de las entrevistas concretas analizadas. Existen otras entrevistas del corpus PRESEEA-Valencia, no las analizadas para este artículo, en las que hay preguntas sobre la muerte, sobre el miedo, etc. En los pocos casos de las entrevistas estudiadas en las que hemos percibido una transmisión de algo que puede llamarse temor o miedo, este se expresa mediante rangos tonales amplios (⬆️), inflexiones tonales descendentes (↘️) y tonos graves (🎷), lo que denota aprensión y expectativa ante la incertidumbre de eventos futuros. Este patrón prosódico, como se observa en los ejemplos 9 y 10, es característico en contextos donde se expresa temor o una sensación de inquietud.

(9) 00:37:24 - 🎻 1990_alto_01_i: ah mira 🎻
00:37:25 - ⬆️ 1990_alto_01_e: si tuviera mucho mucho mucho dinero yo me compraba un vagón de tren ⬆️
00:37:28 - 1990_alto_01_e: siempre me ha encantado
00:37:29 - ⬆️ ⏳ 1990_alto_01_e: yo un jet no me compraría porque me mareo y además me da miedo ⬆️ ⏳
00:37:33 - ⏳ 1990_alto_01_e: un avión ni se me ocurriría me da también miedo un vagón de tren es perfecto ⏳
00:37:36 - 1990_alto_01_e: como del oriente express el
00:37:38 - 1990_alto_01_i: ¿te gusta viajar sola y o acompañada?
00:37:39 - 1990_alto_01_e: no acompañada
00:37:40 - ↗️1990_alto_01_e: yo sola no viajo nunca ↗️
00:37:41 - ↘️ 1990_alto_01_e: no viajaría nunca ↘️
00:37:43 - 🎷 1990_alto_01_e: nunca 🎷
00:37:43 - 1990_alto_01_e: lo más que hago sola ir es al cine
00:37:45 - 1990_alto_01_e: pero yo sola no viajaría uy me una plorera entraría

En el ejemplo 9, el hablante muestra una mezcla de deseo y miedo al hablar de comprar un vagón de tren, utilizando una inflexión ascendente (↗️) que indica la incertidumbre y el temor subyacente a situaciones de riesgo, como volar en avión. Curiosamente, la hablante usa la palabra miedo en una secuencia en la que realmente no parece transmitir esa emoción; es un poco más adelante, cuando alude a que no viajaría nunca sola, en la que combina verdaderos marcas prosódicas que transmiten temor: la inflexión tonal descendente (↘️) y el tono bajo (🎷).

(10) 00:02:09 - 1990_alto_02_i: alguna anécdota que contar
00:02:12 - 1990_alto_02_e: pues
00:02:13 - ⏳ 1990_alto_02_e: del que más recuerdo algo es del que tuve en tercero de EGB ⏳
00:02:20 - 1990_alto_02_e: era un profesor de los de la antigua usanza
00:02:23 - 1990_alto_02_e: que llevaba todavía la vara
00:02:25 - 1990_alto_02_e: para preguntarte
00:02:26 - 1990_alto_02_e: los tiempos verbales
00:02:28 - ⬆️ 🎻 ↗️ 1990_alto_02_e: y recuerdo ⬆️ 🎻 ↗️
00:02:30 - 1990_alto_02_e: aprenderme todos los tiempos verbales de los verbos
00:02:33 - 1990_alto_02_e: irregulares regulares de y todos los verbos porque

00:02:37 - 1990_alto_02_e: era
00:02:39 - 1990_alto_02_e: especial
00:02:40 - ⬆️ 1990_alto_02_e: se llamaba don Pablo ⬆️
00:02:43 - 1990_alto_02_e: y en el momento que te veía un poco distraído
00:02:46 - 1990_alto_02_e: te llamaba la atención
00:02:48 - 1990_alto_02_e: y te hacía la mano para poner
00:02:50 - 1990_alto_02_e: pegarte con la vara

En el caso del ejemplo 10, hay una integración de marco narrativo que empieza por un grupo de entonación más largo de lo habitual: *del que más recuerdo algo es del que tuve en tercero de EGB*. Esta presentación de la figura de la que se va a hablar será seguida por rango tonal elevado (⬆️), junto con un tono agudo (🎻) y una inflexión ascendente (↗️); se trata de una narración en la que la hablante recuerda una experiencia escolar intimidante, en ella se resalta la mezcla de respeto y temor hacia la figura de autoridad representada por su profesor, que era causante de maltrato deliberado en el aula. La propia referencia al nombre del profesor (anonimizada en este artículo mediante un nombre falso) se introduce con un rango tonal elevado.

Así pues, en el caso del miedo, puede registrarse menor ocurrencia en entrevistas sociolingüistas, pero contrariamente una mayor variabilidad en el uso de recursos fónicos. En el futuro, podría añadirse al estudio paralingüístico del miedo la presencia de otros datos como la voz temblorosa o la mayor proliferación de pausas.

### *4.5. Otras emociones detectadas*

Cabe decir que más allá de las cuatro emociones que han sido analizadas, hemos detectado que muy frecuentemente, y sin estar asociado a ninguna de estas emociones en particular, hay un factor que aparece en secuencias discursivas ubicadas entre pausas (grupos de entonación, siguiendo la terminología de Quilis 1993). Se trata de casos en los que el hablante emite cadenas de habla extremadamente amplias, en las que pueden existir inflexiones melódicas internas mínimas, pero que destacan por su mayor duración (marcado en las transcripciones con el símbolo ⏳).

Generalmente, se trata de secuencias en las que se describe algún fenómeno o se narra algún hecho que resulta de interés para el hablante. De hecho, si tuviéramos que asociar alguna de estas secuencias con una emoción en particular, siguiendo el desglose de Plutchik (1980, 2001), podríamos hablar de tedio (o, incluso, de aburrimiento o neutralidad, términos no usados por el autor).

(11)

00:14:41 - 1990_alto_04_e: me gusta me gusta
00:14:43 - 1990_alto_04_e: me gusta pero lo que pasa es que
00:14:46 - 1990_alto_04_e: tienen que me convencer bastante
00:14:48 - 1990_alto_04_e: para salir de marcha
00:14:50 - 🎷 🔇 1990_alto_04_e: sí 🎷 🔇
00:14:51 - ⏳ 1990_alto_04_e: no soy de que arranca un los sábado enseguida llega a casa se pone a arreglarse y para ir con los amigos y tal ⏳
00:14:58 - 1990_alto_04_e: prefiero quedarme en casa o bien
00:15:00 - 1990_alto_04_e: ir con algunos amigos a algún bar y estar un rato

(12)

00:39:53 - 1990_alto_05_e: no lleva que
00:39:54 - 1990_alto_05_e: más o menos mi
00:39:55 - 1990_alto_05_e: o no vamos a poder acoplar nuestros ritmos
00:39:58 - 1990_alto_05_e: de
00:39:59 - 🚴 1990_alto_05_e: funcionamiento 🚴
00:40:01 - 1990_alto_05_e: o ella
00:40:02 - ⏳ 1990_alto_05_e: o esta otra persona va a preferirse a una discoteca por la noche en París en lugar ⏳
00:40:07 - 1990_alto_05_e: de levantarse pronto por la mañana ir e ver a
00:40:10 - 1990_alto_05_e: el palacio de Versalles
00:40:12 - 1990_alto_05_e: pues entonces me voy sola

Tanto los ejemplos 10 como 11 manifiestan situaciones de una cierta disconformidad o desacuerdo con algo expuesto previamente (salir de fiesta). La emoción utilizada no llega a los niveles de enfado ni tampoco de tristeza, sino que expone un posicionamiento personal y discursivo de los hablantes. Se hace de manera seguida, sin apenas vacilaciones y con ello se enuncia de un modo más taxativo la postura personal.

## 5. Conclusiones

Las conclusiones generales obtenidas a partir de este análisis destacan la importancia de la prosodia en la expresión emocional dentro del discurso y cómo esta relación es clave para entender la interacción sociolingüística. El estudio de la prosodia, como se refleja en trabajos como los de Mozziconacci (1995) y Bänziger y Scherer (2005), subraya que las variaciones en tono, intensidad y ritmo son cruciales para diferenciar emociones como la ira, la tristeza, y la alegría. Estos parámetros no solo identifican estados emocionales, sino que también reflejan cómo las emociones estructuran y dirigen las interacciones conversacionales.

Los hallazgos del presente trabajo, apoyados por el corpus PRESEEA, muestran que la prosodia juega un rol importante que modula la percepción emocional y la interpretación del mensaje. Investigaciones como las de Mozziconacci y Hermes (1999) o la de Padilla (2022) o Hidalgo (2020a, 2020b), entre muchas otras, señalan que la entonación y otros aspectos prosódicos no solo comunican emociones, sino que también están influenciados por normas culturales.

Además, el análisis que hemos realizado en este artículo sugiere que ciertas combinaciones de elementos prosódicos, como el tono alto, la intensidad alta o la velocidad alta, están frecuentemente asociadas con emociones de alta emotividad, como el enfado y la alegría, mientras que las combinaciones de tono bajo o intensidad baja se correlacionan con emociones de baja emotividad como la tristeza. Esto es consistente con las observaciones de Garrido Almiñana y Chica Sabariego (2018) y Cao et al. (2014), quienes identifican patrones prosódicos específicos que distinguen emociones en la línea que comentamos.

Aun habiendo obtenido interesantes resultados gracias al análisis efectuado, somos conscientes de que hay aspectos que pueden mejorarse de cara a un análisis en el futuro. En primer lugar, el corpus PRESEEA-Valencia debe ser analizado atendiendo a niveles de instrucción, sexo y edad o, al menos, asegurar una mayor representación sociolingüística. En segundo lugar, las emociones deben ser validadas por jueces para asegurar una mayor efectividad en la emoción analizada; si bien se ha operado con un análisis semiautomático, una prueba de consenso siempre será aconsejable para una fiabilidad más amplia. En tercer lugar, cabe preguntarse si puede refinarse más el análisis prosódico y hacer que de lo paralingüístico pase a lo lingüístico si, como se observa en los estudios de Molinacci (1995) o los de Cao *et al.* (2014), se atiende a los mismos valores estudiados (inflexión tonal, rango tonal, velocidad, intensidad y duración), pero en el marco de unidades más pequeñas: palabras y alófonos. Por supuesto, en cuarto lugar, quizá no asequible en fechas cercanas, se debería realizar un análisis contrastivo

con otros corpus PRESEEA, tanto en el ámbito del español peninsular como del español atlántico.

En conclusión, por lo que hemos visto en este artículo, la prosodia no solo sirve como un marcador de emoción en el discurso, sino que también desempeña un papel esencial en la mediación de las funciones comunicativas, dado que refleja las intenciones del hablante en el contexto interactivo en el que se desarrolla la conversación.

Este estudio inicial consideramos que puede servir de base para futuros análisis que incluyan una capa estadística más robusta y que integren, analicen y comenten un mayor número de entrevistas del corpus PRESEEA. De esta manera, se buscará mejorar la validez poblacional del estudio y aportar conclusiones más generalizables sobre la relación entre prosodia y emoción en diferentes contextos sociolingüísticos.

## Bibliografía

Ang, N., D. Bein, D. Dao, L. Sánchez, J. Tran, y N. Vurdien (2018), «Emotional prosody analysis on human voices», *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*: 737-741.

Bänziger, T., y K. R. Scherer (2005), «The role of intonation in emotional expressions», *Speech Communication*, 46 (3-4): 252-267.

Benus, S., A. Gravano, y J. B. Hirschberg (2007), «Prosody, emotions, and…'whatever'». *Interspeech, 8th Annual Conferente of the International Speech Comunication Association*.

Brezina, V. (2018), *Statistics in corpus linguistics: a practical guide* Cambridge, CUP. DOI: 10.1017/9781316410899.

Cabedo Nebot, A. (2022), *Oralstats* (Versión 1.3) [Software]. Disponible en: https://github.com/acabedo/oralstats.

Cabedo Nebot, A. (2023), *Oralstats Furious*. Disponible en: https://github.com/acabedo/furious.

Cao, H., Š. Beňuš, R. C. Gur, R. Verma, y A. Nenkova (2014), «Prosodic cues for emotion: analysis with discrete characterization of intonation», *Speech prosody (Urbana, Ill.)*: 130.

Chafe, W. (2002). «Prosody and emotion in a sample of real speech», *Relations and functions within and around language*: 277-315.

Cowie, R., y R. R. Cornelius (2003), «Describing the emotional states that are expressed in speech», *Speech Communication*, 40 (1-2): 5-32.

Estellés Arguedas, M. (2023), «Visualizando el conflicto discursivo a través de la expresión fónica: un estudio a partir de dos conversaciones», *Normas*, 13 (1): 224-247. DOI:10.7203/Normas.v13i1.27986

Garrido Almiñana, J. M., y J. A. C. Chica Sabariego (2018), «Pitch range and identification of emotions in Spanish speech: a perceptual study». *Estudios de Fonética Experimental*, 27: 13-36.

Gómez Molina, J. R. (2001), *El español hablado de Valencia, I. Materiales para su estudio: nivel sociocultural alto*, Valencia, Universitat de València.

Gómez Molina, J. R. (2005), *El español hablado de Valencia II. Materiales para su estudio: Nivel sociocultural medio*, Valencia, Universitat de València.

Gómez Molina, J. R. (2007), *El español hablado de Valencia III. Materiales para su estudio: Nivel sociocultural bajo*, Valencia, Universitat de València.

Hidalgo Navarro, A. (2020a), «Hacia una delimitación de parámetros acústicos aptos para el estudio de la entonación emocional». En O. Ivanova, C. V. Álvarez-Rosa, y M. Nevot Navarro (coords.), *Pragmática y discurso oral*, Salamanca, Ediciones de la Universidad de Salamanca: 81-99.

Hidalgo Navarro, A. (2020b), «Rasgos melódicos de la emoción: estudio de un corpus conversacional». *Phonica*, 16: 36-53.

Liu, B. (2012), *Sentiment analysis and opinion mining*, Springer International Publishing. DOI: 10.1007/978-3-031-02145-9.

Martínez, H., y D. R. Avendaño (2011), «Prosodia y emociones: datos acústicos, velocidad de habla y percepción de un corpus actuado», *Lengua y Habla*, 15 (1): 59-72.

Mohammad, S. M., y P. D. Turney (2013), «Crowdsourcing a word-emotion association lexicon». *Computational Intelligence*, 29 (3): 436-465.

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

Mozziconacci, S. (1995), «Pitch variations and emotions in speech», *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 1: 178-181.

Mozziconacci, S. J., y D. J. Hermes (1999), «Role of intonation patterns in conveying emotion in speech», *Proceedings of the 14th International congress of phonetic sciences, ICPhS-99*: 2001-2004.

Padilla García, Xose (2020) «Prosodia emocional y conversación espontánea: bases para el establecimiento de un protocolo de identificación perceptiva». *PHONICA*, 16: 4-35. DOI: 10.1344/phonica.2020.16.4-35.

Padilla García, X. (2022). «La voz como reacción emocional: de qué nos informa la prosodia», *Spanish in Context*, 19 (1): 72-98. DOI: 10.1075/sic.20029.pad.

Padilla García, X. (2023), «Cómo construimos las emociones en la entonación coloquial», *Estudios de Fonética Experimental*, 32: 155-168.

Quilis, A. (1993), *Tratado de fonética y fonología españolas*. Madrid, Gredos.

Plutchik, R. (1980), «A general psychoevolutionary theory of emotion», en R. Plutchik y H. Kellerman (eds.), *Theories of emotion*, Nueva York, Academic Press: 3-33. DOI: 10.1016/B978-0-12-558701-3.50007-7

Plutchik, R. (2001), «The Nature of emotions», *American Scientist*, 89 (4): 344. DOI: 10.1511/2001.28.344.

Wittfoth, M., C. Schröder, D. M. Schardt, R. Dengler, H.-J. Heinze, y S. A. Kotz (2010), «On emotional conflict: interference resolution of happy and angry prosody reveals valence-specific effects», *Cerebral Cortex*, 20 (2): 383-392.

Carles Navarro-Carrascosa (ORCID: 0000-0002-0303-982X), «Descripción y análisis de un corpus de español oral de la comunidad de habla LGTBI». *Biblioteca de Babel: Revista de Filología Hispánica*, vol. extra. 2 (2024), pp. 239-261.
**DOI**: 10.15366/bibliotecababel2024.extra2.010 · **ISSNe**: 2695-6349 · Recibido: 31/07/2024 · Aceptado: 08/10/2024

# Descripción y análisis de un corpus de español oral de la comunidad de habla LGTBI

Carles Navarro-Carrascosa
*Universidad de Valladolid*
carles.navarro@uva.es

**Resumen**: Los corpus orales son herramientas esenciales para la investigación lingüística, ya que proporcionan datos auténticos del uso del lenguaje en contextos reales. En este trabajo se presenta el Corpus Oral de la Comunidad de Habla LGTBI (Navarro-Carrascosa 2023a), que recopila muestras de habla de esta comunidad, facilitando el estudio de sus códigos lingüísticos y prácticas comunicativas. Para mejorar este corpus se propone incluir muestras de habla de miembros del colectivo que no fueron acopiadas en los primeros registros, así como ampliar las muestras a más ciudades y áreas rurales, además de las variantes hispanohablantes de otros países para enriquecer el corpus con una mayor diversidad lingüística y cultural. Estas mejoras permitirán que el corpus beneficie investigaciones detalladas sobre variaciones lingüísticas, influencias socioculturales y la evolución del lenguaje en la comunidad LGTBI. Con una expansión y desarrollo continuos, el corpus seguirá proporcionando datos valiosos para una comprensión más completa y diversa del lenguaje dentro de esta comunidad.

**Palabras clave**: lingüística de corpus, lingüística *queer*, corpus orales, comunidad de habla LGTBI, sociolingüística.

## Description and analysis of an oral Spanish corpus of the LGTBI speaking community

**Abstract**: Oral corpora are essential tools for linguistic research, as they provide authentic data on language use in real contexts. This work presents the Oral Corpus of the LGTBI Speaking Community (Navarro-Carrascosa 2023a), which collects speech samples from this community, facilitating the study of its linguistic codes and communicative practices. To improve this corpus, it is proposed to include samples from community members that were not collected in the initial records, as well as to expand the recordings to more cities and rural areas, and to incorporate Spanish-speaking variants from other

countries to enrich the corpus with greater linguistic and cultural diversity. These improvements will allow the corpus to benefit detailed investigations into linguistic variations, sociocultural influences, and the evolution of language in the LGTBI community. With continuous expansion and development, the corpus will continue to provide valuable data for a more comprehensive and diverse understanding of language within this community.

**Keywords**: corpus linguistics, queer linguistics, oral corpora, LGTBI speaking community, sociolinguistics.

## 1. Introducción

Los corpus orales son herramientas indispensables para la investigación lingüística, ya que proporcionan datos auténticos del uso del lenguaje en contextos reales. Estas colecciones estructuradas de grabaciones de habla permiten a los investigadores analizar una amplia gama de fenómenos lingüísticos, desde variaciones dialectales hasta la pragmática y la sociolingüística. La importancia de los corpus orales radica en su capacidad para capturar la riqueza y diversidad del habla, ofreciendo una base empírica sólida para estudios detallados y precisos.

En este contexto, se presenta el Corpus Oral de la Comunidad de Habla LGTBI (Navarro-Carrascosa 2023a), un recurso pionero que recopila muestras de habla reales de esta comunidad. Este corpus ofrece numerosos beneficios, como la posibilidad de estudiar los códigos lingüísticos propios del colectivo LGTBI y las intenciones pragmalingüísticas en la comunicación interna entre sus miembros. Sin embargo, también enfrenta ciertos desafíos, especialmente en términos de la representación de subgrupos menos visibilizados y la obtención de grabaciones en modalidad secreta para captar el habla más espontánea posible.

Este trabajo no solo describe y explica el corpus actual, sino que también indica pautas para su ampliación y desarrollo futuros. Se explorarán estrategias para incluir a más personas transexuales binarias, intergénero, bisexuales y de otras categorías no registradas hasta ahora, como asexuales y demisexuales. Además, se discutirán métodos para registrar más grabaciones con personas de mayor edad, permitiendo así estudios diacrónicos sobre la evolución del argot a través de generaciones y la expansión geográfica del corpus a más ciudades y áreas rurales, así como a otras variantes hispanohablantes en países de habla española.

Finalmente, se presentarán las perspectivas de futuro del corpus y las posibles investigaciones que podría beneficiar. Desde estudios fonético-fonológicos y léxicos hasta análisis morfológicos y pragmáticos, este corpus tiene el potencial de enriquecer significativamente la investigación lingüística. Su desarrollo y expansión continuarán proporcionando datos valiosos para una comprensión más completa y diversa del lenguaje dentro de la comunidad de habla LGTBI.

## 2. Los corpus orales del español

Los corpus orales del español son colecciones estructuradas de grabaciones de habla en lengua española que se utilizan para el estudio y análisis lingüístico. Estos corpus contienen datos orales recolectados en diversas situaciones comunicativas y contextos, incluyendo conversaciones espontáneas, discursos formales, entrevistas, debates, y otros tipos de interacción verbal. El objetivo principal de estos corpus es proporcionar material auténtico y representativo del uso del idioma en situaciones reales.

### *2.1. Corpus orales generales*

Los corpus orales generales son aquellos que recogen muestras de habla de los miembros pertenecientes a una sociedad. Este tipo de corpus busca capturar una representación amplia y diversa del uso del lenguaje, incluyendo a personas de diferentes edades, géneros, niveles educativos, ocupaciones y orígenes geográficos. Un ejemplo destacado de corpus general es el de conversaciones coloquiales del grupo Val.Es.Co.[1], el cual, aunque tiene un objeto de estudio muy delimitado (el español coloquial), no discrimina a informantes, ya que cualquier miembro de una sociedad es susceptible de utilizar un registro familiar en determinados contextos.

Estos corpus son fundamentales para el estudio sociolingüístico, pues permiten analizar cómo se manifiestan las variaciones lingüísticas en distintas capas de la sociedad y cómo se utilizan diferentes registros del habla en contextos variados. Además, estos corpus son útiles para investigar fenómenos como la evolución del lenguaje, la influencia de factores sociales en el habla y la dinámica del cambio lingüístico.

Otros ejemplos son el corpus PRESEEA[2] (Proyecto para el Estudio Sociolingüístico del Español de España y de América), que recopila datos orales de diferentes ciudades hispanohablantes; y el CORLEC[3]

[1] https://www.valesco.es/#/pages/cod_hj3y7hwvuuajtlkq0ik/cod_m5em4hajg5ljtpxzuc1.
[2] https://preseea.uah.es/corpus-preseea.
[3] http://www.lllf.uam.es/ESP/Corlec.html.

(Corpus Oral de Referencia de la Lengua Española Contemporánea), que se centra en la recopilación de muestras de habla de diversas regiones y contextos.

### *2.2. Corpus orales parciales o sectoriales*

Son aquellos que se limitan a un sector social específico, como determinados grupos sociales, profesionales o regionales. Estos corpus están diseñados para estudiar y describir la variedad del lenguaje en contextos más acotados y concretos, proporcionando información detallada sobre el uso del lenguaje en comunidades particulares.

Al tener los corpus parciales objetivos mucho más definidos que los corpus generales, los investigadores pueden realizar análisis más detallados y específicos, facilitando el estudio de fenómenos lingüísticos particulares dentro de esos grupos.

Algunos ejemplos de corpus parciales son el COVJA (Corpus Oral de la Variedad Juvenil Universitaria del Español de Alicante) (Azorín Fernández 2005); otro ejemplo más reciente es el corpus del proyecto ESPRINT[4], que se enfoca en las interacciones de parejas; el COSER[5] (Corpus Oral y Sonoro del Español Rural), que proporciona muestras de las variantes dialectales rurales del español de España, en concreto, de personas mayores; el CHCS (Corpus del Habla Culta de Salamanca), de Fernández Juncal (2005); el COJEM[6] (Corpus Oral Juvenil del Español de Mallorca); o el COLEM (Corpus Oral de la Lengua Española en Montreal), de Pato (2020).

## 3. Fases para la elaboración de un corpus oral parcial de una comunidad lingüística específica

Moreno Fernández (2005) enumera los fundamentos metodológicos que él y su equipo llevaron a cabo para la elaboración de los distintos corpus orales del proyecto PRESEEA. Nosotros los hemos sintetizado en cuatro procedimientos metodológicos que deben seguirse para la construcción de un corpus oral: la elección de informantes, el tamaño de la muestra, el método de recolección y la transcripción. Todos estos criterios responden a decisiones que los investigadores tendrán que afrontar para desarrollar su tarea. Sin embargo, todos los puntos son mucho más delicados cuando el objetivo es la elaboración de un realia oral de una comunidad de habla específica.

[4] Estrategias pragmático-retóricas en la interacción conversacional conflictiva entre íntimos y conocidos: intensificación, atenuación y gestión interaccional. Ministerio de Ciencia e Innovación (PID2020-114805GB-100). Directoras: Marta Albelda / Maria Estellés (Universitat de València).
[5] http://www.corpusrural.es.
[6] http://www.linred.es/numero13_corpus-1.html.

### *3.1. La elección de informantes*

Sobre este aspecto, Briz (2012) ya señalaba la falta de representatividad de todos los corpus orales, una carencia que se acentúa aún más en aquellos de carácter parcial, pues en ellos se pone el foco en un sector de la sociedad más acotado. Para definir el tipo de informante necesario para un corpus oral sectorial es importante plantear algunas cuestiones específicas. Generalmente, se trata de criterios demandados por los objetivos de las investigaciones para las que se acopian los textos orales: «núcleos urbanos, monolingües o bilingües, con una población hispanohablante bien asentada, con conciencia de comunidad y suficientemente diversa desde un punto de vista sociológico» (Moreno Fernández 2005: 127). Las decisiones que deben tomarse, a este respecto, tienen que ver con criterios demográficos, como la edad, los años de residencia, el género, el nivel de instrucción, etc. (Carcelén y Uclés 2019).

Estos son los procedimientos más habituales, pero, dependiendo del tipo de investigación, podrían surgir otros más específicos o particulares que acotarán mucho más el perfil de los informantes. Por ejemplo, Sanmartín (1998, 1999a y 1999b), para su trabajo sobre el español de la delincuencia elaboró un corpus de habla cuyos informantes eran reclusos de la prisión de Valencia. Manejó variables generales y otras más específicas, como el tipo de delito que había cometido el informante o los años que llevaba recluido en el momento de la entrevista.

### *3.2. El tamaño de la muestra*

El tamaño de la muestra es una de las cuestiones más polémicas entre los distintos investigadores en sociolingüística (Briz y Val. Es.Co. 2002a). Lo habitual es aplicar principios estadísticos para obtener muestras representativas de la población lingüística (Torruela y Llisterri 1999). Esta cuestión va aparejada al número de participantes.

> El plan de muestreo diseñado para una muestra aleatoria simple supone un nivel de confianza del 90 % en los resultados obtenidos, con un margen de error en esa confianza del 6 %. Este nivel de precisión exige una muestra de 189 informantes, valor que nos permitirá un alto grado de seguridad en las inferencias obtenidas (Briz y Grupo Val.Es.Co. 2002a: 14).

Para el corpus PRESEEA se estableció que el tamaño de la muestra debe estar relacionado con la población total del municipio de la recopilación:

> El estudio de las comunidades con poblaciones de entre medio millón y un millón de habitantes se aborda mediante muestras formadas por 54 informantes. Cuando la ciudad es de un tamaño mayor y de una mayor complejidad sociológica, se recomienda el empleo de muestras de 72 a 108 hablantes (Moreno Fernández 2005: 128).

No obstante, este criterio puede variar atendiendo a varios factores, entre ellos el tipo de recolección (que se verá en el siguiente apartado). Los corpus vivos (aquellos cuya elaboración nunca finaliza, pues siempre van ampliándose, como es el caso del corpus de conversaciones coloquiales del grupo Val.Es.Co), al ser ilimitados y estar en constante elaboración, no siempre establecen un número de informantes.

En los corpus que buscan plasmar los modos de habla de comunidades específicas, se pueden presentar diversos problemas a la hora de establecer un criterio cuantitativo en el número de participantes. En primer lugar, si las comunidades son minoritarias, la dificultad puede girar en torno a encontrar informantes o a que estos deseen participar. Por ejemplo, el corpus de Montecino Soto (2008), limitado a personas sin hogar de Santiago de Chile, necesitó de más de dos años para conseguir once entrevistas.

### *3.3. El método de recolección*

La codificación y presentación de datos orales implica decisiones sobre el formato de registro y las convenciones de transcripción (Pons y Ruiz Gurillo 2005; Recalde y Vázquez Rozas 2009). Los corpus orales suelen recurrir a dos técnicas de acopio de muestras: la entrevista sociolingüística y la grabación secreta.

#### *3.3.1. La entrevista sociolingüística*

La entrevista sociolingüística es una técnica utilizada por lingüistas investigadores para obtener muestras de habla reales en determinados contextos sociales. Es una metodología original de William Labov, diseñada en los años 60 y con el objetivo de estudiar la variación dialectal y el cambio lingüístico a través del registro y el análisis de las muestras acopiadas.

Existen varias técnicas metodológicas empleadas para desarrollar una entrevista sociolingüística; algunas de ellas están diseñadas para minimizar la paradoja del observador[7] (Labov 1972). Estas técnicas se clasifican de la siguiente manera:

[7] La paradoja del observador se refiere al fenómeno donde los sujetos modifican su comportamiento cuando saben que están siendo observados.

— **La entrevista estructurada**: está basada en un conjunto de preguntas preparadas con anterioridad (en ocasiones, acordadas), lo que facilita la comparación de respuestas entre los distintos entrevistados. No obstante, varios autores destacan que las respuestas de estas entrevistas suelen resultar menos espontáneas y, por tanto, su validez es menor (Calderón Noguera y Alvarado Castellanos 2011).

— **La entrevista semiestructurada o semidirigida**: el entrevistador sigue un guion básico, pero puede adaptar las preguntas según las respuestas del informante, fomentando así una conversación más natural y detallada. «Es una estrategia de recolección de materiales de punto medio, en la que las preguntas planificadas con anterioridad tienen como propósito buscar registros de actuación lingüística […] o datos lingüísticos de manera directa» (Calderón Noguera y Alvarado Castellanos 2011). Según Moreno Fernández (2005: 128), las entrevistas sociolingüísticas «suelen recogerse mediante conversaciones semidirigidas y grabadas con magnetófono [o cualquier dispositivo digital habilitado] a la vista en situación de entrevista».

— **La entrevista no estructurada**: también llamada entrevista libre (Fernández Sanmartín 2022). Su principal característica es la total flexibilidad y espontaneidad, lo que permite una conversación más natural. Es útil para registrar el habla vernácula, aunque dificulta un análisis sistemático.

Tanto en la entrevista semidirigida como en la no estructurada, se pueden utilizar diferentes estrategias con el objetivo de mejorar la naturalidad de las respuestas, como la inclusión de múltiples entrevistadores (Calderón Noguera y Alvarado Castellanos 2011) o la realización de entrevistas grupales con más de un informante a la vez donde los participantes se sientan más cómodos y menos observados (Moreno Fernández 2011; Navarro-Carrascosa 2023a).

### *3.3.2. La grabación secreta*

La grabación secreta es una técnica de recopilación de muestras de habla reales que consiste en registrar conversaciones sin que los participantes sepan que están siendo grabados. El objetivo es que las manifestaciones lingüísticas sean lo más espontáneas posible, evitando probables moderaciones expresivas de los hablantes al ser conscientes del registro.

La grabación secreta puede ser con observación participante (el investigador que consigue la muestra participa en ella) o sin esta. Ambas modalidades son cada vez más habituales en la investigación sociolingüística por considerarse «la forma más eficaz de obtención de datos

del español coloquial y permite soslayar inconvenientes teóricos como la llamada *paradoja del entrevistador*» (Briz y Val.Es.Co. 2002a: 17).

Con la legislación anterior se permitía que el investigador informara debidamente, una vez finalizado el registro, a los participantes de que la grabación había tenido lugar y les solicitara, *a posteriori*, un consentimiento por escrito firmado para que el documento resultante fuera utilizado. En la actualidad, el consentimiento debe otorgarse siempre antes del registro, lo cual dificulta más la espontaneidad de las muestras, especialmente con miembros de comunidades específicas, cuyos informantes tienen características más concretas y menos habituales. Además, cualquier dato que permita la identificación de cualquiera de los informantes deberá ser falseado.

### *3.4. La transcripción*

La transcripción de textos orales con fines lingüísticos implica la representación escrita de lo dicho, pero también la preservación de elementos paralingüísticos y contextuales esenciales que permitirán un análisis exhaustivo. Cabe añadir que este proceso no es una mera traducción entre modos de expresión; es, además, un acto forzosamente interpretativo, en el que quien transcribe incorpora evaluaciones y toma decisiones sobre cómo representar cada detalle del habla. Algunos autores consideran que es un método fundamental para el análisis lingüístico actual, ya que facilita el manejo de datos mediante *software* especializado para la visualización, etiquetado morfosintáctico, entre otros objetivos (Ridao Rodrigo 2021).

Podemos enumerar tres tipos de transcripción: la ortográfica, la fonética y fonológica y la anotada:

— La transcripción ortográfica plasma la grabación oral usando los recursos ortográficos estándar. Se trata de trasladar al papel el texto oral (Bejarano, Llanos, Rubio y Bonilla 2018).

— La transcripción fonética y fonológica procura capturar con mayor precisión los sonidos del habla, utilizando una serie de símbolos para representar detalles prosódicos, entre otros.

— La transcripción anotada incluye anotaciones adicionales sobre aspectos paralingüísticos y contextuales, como elementos no verbales que acompañan al discurso (Ridao Rodrigo 2021).

Para facilitar la transcripción y garantizar la uniformidad se han desarrollado diversos protocolos. Por ejemplo, el Instituto Caro y Cuervo ha creado el protocolo de transcripción ortográfica CLICC (Bejarano, Llanos, Rubio y Bonilla 2018), que establece pautas claras

para transcribir corpus orales. Sin embargo, los métodos más conocidos y utilizados en sociolingüística son el sistema PRESEEA y Val. Es.Co., que aplican sistemas de transcripción de tipo anotada.

El sistema de Briz y el grupo Val.Es.Co (2002b), uno de los más populares en la actualidad, se basa en tres principios fundamentales: transcripción ortográfica adaptada, es decir, respecta la normativa ortográfica e incluye símbolos convencionales, como signos de exclamación o interrogación; incorporación de elementos paralingüísticos, como risas, vacilaciones, reformulaciones, cambios de tono, etc. Estos se marcan como etiquetas específicas dentro del texto para que el lector pueda interpretarlas de manera clara. Por último, el uso de etiquetas para señalar aspectos relevantes de la interacción, como las interrupciones, los solapamientos y las palabras ininteligibles, entre otros.

## 4. Necesidad de un corpus oral de la comunidad de habla LGTBI

Desde hace algunas décadas, la academia estadounidense ha abierto un espacio para los estudios *queer*, es decir, aquellos estudios pertenecientes a las distintas áreas de conocimiento que desarrollan sus investigaciones desde una perspectiva de las identidades de género y sexoafectivas disidentes. En los últimos años, esta tendencia ha empezado a surgir en el hispanismo y podemos encontrar varios trabajos académicos de diferentes disciplinas (arquitectura, literatura, educación, derechos, etc.) desde un prisma LGTBI. La lingüística *queer* no ha sido una excepción y está ganando un espacio significativo en los estudios hispanistas.

En este sentido, para poder ofrecer materiales de estudio para una lingüística *queer* se hace necesario el Corpus oral de la comunidad de habla LGTBI (Navarro-Carrascosa 2023a), cuyo objetivo es

> ofrecer materiales para estudiar y analizar los códigos lingüísticos propios de esta subcultura y las intenciones pragmalingüísticas que se dan en la comunicación interna entre sus miembros y, de este modo, servir de herramienta en las investigaciones de la Lingüística *queer* hispánica (Navarro-Carrascosa 2023a: 11).

Cada colectivo alberga una serie de características actitudinales que las distinguen del resto. Entre ellas se pueden destacar comportamientos lingüísticos determinados. En el caso de la comunidad de habla LGTBI, estos se hacen especialmente particulares, ya que provienen de necesidades históricas de mantenerse en la clandestinidad e interactuar con otros miembros del colectivo. Este contexto represivo para sus miembros motivó una serie de mecanismos lingüísticos secretos

que hoy en día ya no son necesarios. Sin embargo, algunos de ellos se han mantenido y han evolucionado, ya descodificados para el resto de la sociedad gracias a publicaciones literarias y a los medios de comunicación (Rodríguez 2008).

> Este argot propio del colectivo LGTBI no solo refleja los intereses y necesidades de esta comunidad, sino que también expresa roles, comportamientos sexuales específicos dentro de esta, reforzando la cohesión interna y creando una alternativa a la heterosexualidad hegemónica, tal y como apunta Bengoechea (2015).

Para poder clasificar, organizar y analizar todas estas estrategias lingüísticas se hizo necesaria la compilación de todas las muestras de habla reales que conforman el corpus y se mantiene hoy día la necesidad de seguir ampliándolo con más conversaciones y entrevistas que mantengan el carácter sincrónico del *realia* y amplíen la variedad de informantes en todas las variantes que un corpus de estas características debe manejar (la identidad de género y la orientación sexual, fundamentalmente).

## 5. El corpus oral de la comunidad de habla LGTBI

Como se señalaba en el apartado anterior, uno de los corpus orales parciales o sectoriales del español es el que recogen muestras de habla reales de la comunidad LGTBI[8]. El corpus consta de un total de cincuenta y nueve informantes distribuidos en dos tipos de muestras: conversaciones coloquiales y entrevistas semiestructuradas.

Las conversaciones coloquiales incluyen diecisiete informantes, organizados en cinco grabaciones: dos en Valencia (C.VA.1 con cuatro participantes y C.VA.2 con cinco) y tres en Madrid (C.MA.1 con dos, C.MA.2 con tres y C.MA.3 con tres). Estas conversaciones fueron obtenidas mediante grabación secreta para preservar la naturalidad de las interacciones, con una duración total de 282 minutos y 38 segundos.

Las entrevistas cuentan con cuarenta y dos informantes, realizadas en Barcelona, Madrid y Valencia. Se han llevado a cabo tres entrevistas en Barcelona con diez participantes; nueve en Madrid con diecinueve participantes; y ocho en Valencia con tres participantes. La duración acumulada de las entrevistas es de 1175 minutos y 43 segundos.

[8] Se entiende por comunidad LGTBI a aquel grupo social compuesto «por personas que representan realidades sexoafectivas y/o de género que no están integradas dentro de lo que socialmente es considerado como normal» (Navarro-Carrascosa 2023b: 45). Por tanto, se considerará que los miembros de esta comunidad serán hombres y mujeres cisgénero homosexuales, bisexuales, pansexuales, mujeres trans, hombres trans y personas no binarias.

Este corpus, iniciado por Navarro-Carrascosa (2023a), sigue una serie de pautas y criterios de elaboración atendiendo a las particularidades que el contexto de acopio requiere, los cuales se desarrollan a continuación para la posible ampliación.

## *5.1. Informantes*

Para seleccionar con precisión los participantes en la recolección de datos de una investigación lingüística enfocada en un grupo social específico se debe identificar las variables sociolingüísticas pertinentes que definan a los miembros de la comunidad de estudio y que los diferencien del resto. En el caso del colectivo LGTBI estas variables se centraban en la orientación sexual, la identidad de género y el círculo social del informante, determinando si las personas que lo conforman son miembros del colectivo o no. A continuación, se presentan las variables más importantes que resultan determinantes para la selección de informantes del corpus oral LGTBI[9]:

— La orientación sexual: las variables que se manejan en esta categoría son tres: la heterosexualidad (aquellos que se sienten atraídos hacia el género opuesto); la homosexualidad (la atracción sexual hacia el mismo género); y la bisexualidad[10] (correspondiente a personas que se sienten atraídas a otras personas, independientemente del género de estas).

— La identidad de género: no se corresponde la identidad de género con la biología de los genitales de cada individuo. La primera, se refiere a los roles de género que cada uno siente que le son propios. Nuevamente, destacamos tres variables: personas cisgénero (la identidad de género está alineada con lo socialmente atribuido a cada uno de los géneros; es decir, una persona con genitales masculinos se identifica con los roles de género socialmente asociados al hombre); personas transgénero, esto es, la identidad del informante no está determinada por sus genitales de nacimiento[11]. Por último, en cuanto a la identidad de género destacamos una última categoría para identificar a los informantes del corpus: se trata de las personas no binarias (NB) o intergénero, las cuales no se identifican identitariamente ni como hombres ni como mujeres. Llegados a este punto, nos planteamos la posibilidad de un cuarto grupo en cuanto al género: la intersexualidad[12],

[9] Los términos referidos a identidades de género y orientaciones sexuales se definen en profundidad en obras lexicográficas parciales, como las de Mira (1999), Pereda (2004) y Rodríguez (2008).
[10] Los estudios *queer* consideran la ruptura entre la dicotomía de género masculino-femenino, lo que cuestionaría el término bisexualidad (si se asume que existen más de dos géneros) y haría necesario hablar de pansexualidad, la atracción hacia las personas independientemente de su género, entendiendo que hay más de dos.
[11] Dentro de este grupo de pueden distinguir mujeres transgénero (mujeres que nacieron con genitales masculinos) y hombres transgénero (hombres que nacieron con genitales masculinos).
[12] La intersexualidad es una condición biológica en la que una persona nace con características

es decir, personas que han nacido con genitales ambiguos o variaciones cromosómicas.

— El círculo social del informante: como ya se ha apuntado, es muy relevante para la elección de participantes de este corpus, pues es relevante que un hombre cisgénero heterosexual forme parte del inventario si su círculo social está conformado, en parte, por personas LGTBI, ya que probablemente esto afecte a su expresión en contextos de familiaridad; del mismo modo que un miembro del colectivo LGTBI que no se relacione con otros no es considerado un hablante representativo del argot de esta comunidad. Por tanto, las posibilidades son dos: informantes con contacto social con personas LGTBI y los informantes que no lo tienen.

— La edad: es una variable muy relevante en la mayoría de los estudios sociolingüísticos. Para las investigaciones en lingüística *queer* también lo es, pues resulta de interés observar si hay cambios en las formas de expresarse de los miembros de esta comunidad lingüística atendiendo a este factor. La división en las franjas de edad que se propone para catalogar a los informantes difiere de la clasificación clásica en sociolingüística (aplicada a estudios como los de PRESEEA 2021; entre otros), pues se parte de la base de que el contexto social de los miembros del colectivo LGTBI tiene particularidades socio-históricas relevantes y, por tanto, la clasificación de los grupos de edad que se ha establecido es la siguiente:

a) De 18 a 25 años: este grupo ha crecido en una sociedad en la que el colectivo LGTBI está totalmente visibilizado, han alcanzado varios derechos sociales y, por tanto, su contexto no ha experimentado tantas censuras como los grupos de mayor edad.

b) De 26 a 35 años: los participantes de esta franja de edad también han crecido en un contexto donde las personas del colectivo están más o menos visibilizadas, pero la imagen que se proyecta de ellas se relaciona con la burla. Además, algunos aspectos relacionados con este colectivo todavía se consideran un tabú (Mira 2004).

c) De 36 a 45 años: el contexto en el que han crecido las personas de esta franja de edad tiene un grado de aceptación hacia el colectivo que, aunque empieza a abrirse, todavía es reducido (Mira 2004).

d) Más de 45 años: se han criado durante la dictadura o durante la transición, es decir, en la época en la que la homosexualidad estaba penalizada. Por tanto, la imagen que se proyecta de los gais y las

sexuales físicas que no encajan en las típicas definiciones binarias de masculino o femenino. Estas características pueden incluir variaciones en los genitales, las gónadas (ovarios o testículos), los cromosomas sexuales (XX, XY o combinaciones atípicas), o los niveles hormonales.

lesbianas es muy negativa. Por otro lado, el resto del colectivo está totalmente invisibilizado (Mira 2004).

### *5.2. Recopilación de datos*

La recopilación de datos en los estudios sociolingüísticos y dialectológicos sigue un conjunto de características generales que aseguran la calidad, ética y confidencialidad de la información obtenida. A continuación, se describen las características fundamentales para la recopilación de datos en este contexto:

— La **transcripción** de los datos: se realiza utilizando el sistema Val.Es.Co, desarrollado por Briz y el Grupo Val.Es.Co. (2002a). Este sistema de transcripción es detallado y permite capturar con precisión las características del habla, incluyendo pausas, entonación y otros fenómenos prosódicos.

— Para proteger la identidad de los participantes y cumplir con la legalidad vigente se lleva a cabo la **anonimización** de los datos. Esto implica el cambio de nombres y cualquier otra información que pudiera permitir el reconocimiento de los informantes.

— Los investigadores redactan un **compromiso de confidencialidad**, en el cual se comprometen a mantener la anonimización y a tratar las grabaciones de manera correcta y ética. Este documento debe ser archivado y conservado como parte de la documentación del estudio, asegurando así el cumplimiento de las normas éticas.

— Tanto para las entrevistas como para las grabaciones secretas de conversaciones coloquiales, se utiliza una **ficha identificatoria** basada en la metodología de Briz y el Grupo Val.Es.Co. (2002a). Esta ficha incluye información relevante del registro, como la fecha y lugar de la grabación, el investigador principal, el número de participantes, sus edades, profesiones y la duración del registro, entre otros datos. Esta información es esencial para contextualizar los datos y facilitar su análisis posterior. Dadas las particularidades de los participantes en las grabaciones del corpus oral de la comunidad de habla LGTBI, la ficha ha sido completada para que incluya rasgos como la identidad de género y la orientación sexual.

#### *5.2.1. Entrevistas*

En el caso de las entrevistas, para garantizar su clasificación e identificación es importante el sistema de etiquetado, que sigue las siguientes pautas:

— Cada entrevista se marca con una *E* al inicio del etiquetado para diferenciarlas de las grabaciones secretas.

— Se indica la ciudad en la que se ha llevado a cabo el registro mediante un código específico (*VA*, para Valencia; *BA* indica que el registro se ha hecho en Barcelona; y *MA* señala que la entrevista ha tenido lugar en Madrid). Para ciudades nuevas en las que todavía no se han recopilado muestras, se asigna una marca específica que procure no coincidir con otras ya asignadas. Por ejemplo: *SE*, para Sevilla; *VAll*, para Valladolid; o *MAl*, para Málaga.

— El etiquetado finaliza con el número asignado al documento, generalmente siguiendo el orden en el que fueron recopiladas: *E.VA.3*, por tanto, es la tercera entrevista obtenida en Valencia; mientras que *E.BA.2* se corresponde con la segunda grabada en Barcelona.

El sistema de etiquetado permite una rápida localización de los documentos (tanto de los audios como de las transcripciones) y un sencillo manejo de los datos, facilitando el análisis comparativo entre diferentes registros.

Para maximizar la naturalidad del discurso y mitigar la paradoja del observador (Labov 1972), es recomendable propiciar un ambiente cómodo para los entrevistados. Para ello la metodología de registro de entrevistas del Corpus Oral de la Comunidad de Habla LGTBI establece las siguientes directrices:

— El **lugar de la grabación** debe ser un espacio familiar para los entrevistados. Idealmente, el hogar de uno de los participantes o un comercio habitual para ellos, siempre de su elección. Esto permite que los entrevistados se sientan en un entorno seguro y conocido, lo que facilita una mayor naturalidad en sus intervenciones.

— El **número de participantes** por entrevista se recomienda que no sea inferior a dos. Además, la **relación** entre ellos debe ser afectiva (familiares, amigos, pareja). Esta vinculación emocional promueve un ambiente relajado y permite que los participantes se sientan más cómodos al expresarse.

— El entrevistador lleva un esquema con temas específicos para la entrevista, enfocándose en asuntos relativos a la comunidad LGTBI. Sin embargo, este esquema no condiciona el desarrollo de la sesión y se permite que los participantes introduzcan otros temas si así lo desean. Este enfoque flexible facilita una conversación más espontánea y fluida, ya que los entrevistados pueden hablar sobre asuntos que les resulten más naturales e interesantes durante el registro.

En definitiva, para lograr un entorno propicio y tener más posibilidades de obtener muestras de habla espontáneas y, sobre todo, coloquiales, el entrevistador adopta una actitud relajada y amistosa, interviniendo como un participante más, manifestando sus opiniones en determinados asuntos y evitando imponer una estructura rígida en la conversación.

### *5.2.2. Grabaciones secretas*

Las grabaciones secretas son registros realizados de manera oculta para captar el habla más natural y espontánea posible. Por lo general, suelen realizarse los registros sin el conocimiento expreso del informante.

El papel del investigador en este tipo de acopio de muestras de habla es la de infiltrarse en un grupo social (con al menos dos interlocutores, además del recopilador) y grabar sus interacciones lingüísticas[13]. A los informantes se les comunica que han sido grabados y se les solicita un permiso para utilizar el registro, siempre con fines científicos. Dicha autorización es firmada por los participantes, que aceptan a cambio de que cualquier dato personal que pudiera identificarlos en la grabación sea suprimido o anonimizado.

Al igual que en el caso de las entrevistas, los documentos generados por las grabaciones secretas (tanto los archivos de audio como las transcripciones) se etiquetan. El sistema utilizado es el mismo que el mencionado en las entrevistas, pero se marca con una *C* para diferenciarlo, indicando que se trata de una conversación coloquial espontánea.

## *5.3. Investigaciones en las que se puede integrar el corpus*

El corpus oral de la comunidad de habla LGTBI es una herramienta valiosa, no solo porque supone el primer acopio de registros reales de este colectivo, sino porque, además, abre un camino en las investigaciones sobre el argot de esta comunidad.

A continuación, se describen algunas de las posibles investigaciones en las que este corpus (tanto en su estado actual como cuando sea ampliado) puede resultar particularmente útil.

[13] La problemática ética y legal que este tipo de técnicas puede presentar en la actualidad se plantea y se desarrolla en el apartado 5.4.3.

### *5.3.1. Estudios fonético-fonológicos*

Este corpus puede facilitar el análisis de las características fonéticas y fonológicas del habla de la comunidad LGTBI. Investigaciones sobre patrones entonativos distintivos, uso marcado de ciertos fonemas y modulaciones prosódicas que reflejan la construcción de identidad en contextos de interacción social, especialmente en ámbitos como el uso de la entonación afectiva o variaciones en la pronunciación de vocales, consonantes o el seseo para proyectar afiliación a subgrupos, serían de gran relevancia.

### *5.3.2. Estudios léxicos y fraseológicos*

El corpus permite estudiar el léxico específico y las expresiones idiomáticas que son propias de la comunidad LGTBI, enfocándose en el uso y evolución de términos específicos como *maricón* (Navarro-Carrascosa, 2019), *pluma* o *drag*, que han sido resignificados y reapropriados. La investigación podría abordar la creatividad léxica, incluyendo el uso de prefijos y sufijos con connotaciones humorísticas o irónicas (la formación de palabras, el uso de prefijos y sufijos, y las expresiones características ofrecen un campo amplio para investigar cómo se crean y utilizan estos términos en diferentes contextos), así como expresiones que forman parte del argot LGTBI en distintos contextos sociales y culturales, tales como el *ball culture* o *voguing*.

### *5.3.3. Estudios morfológicos*

La morfología en el habla LGTBI representa una herramienta para reflejar identidades de género diversas. Este corpus permite investigar la adaptación de estructuras morfológicas para visibilizar el género no binario y analizar cómo se emplean innovaciones como el uso de terminaciones neutras, referidas al género gramatical no marcado (por ejemplo, *-e*), o formas flexionadas alternativas en un contexto donde se resignifican los sistemas tradicionales de género.

### *5.3.4. Estudios pragmáticos y sociolingüísticos*

El corpus puede resultar de utilidad para estudios pragmáticos que analicen las intenciones comunicativas y el uso del lenguaje en interacciones sociales dentro de la comunidad LGTBI, como el uso de la ironía, el doble sentido o la subversión de normas sociales mediante el lenguaje. Además, permite analizar cómo las identidades de género y las orientaciones sexuales se expresan a través de actos de habla específicos, tales como saludos, cumplidos o estrategias de afrontamiento en contextos de discriminación o solidaridad.

### *5.3.5. Estudios sobre identidades de género y uso del lenguaje*

Otras investigaciones que pueden sacar provecho de este corpus son aquellas que busquen entender cómo las personas no binarias y transgénero utilizan el lenguaje para expresar su identidad. Este compendio de muestras de habla proporciona datos sobre el uso de pronombres y formas no marcadas de género, así como sobre la adopción de términos específicos para describir experiencias únicas.

### *5.3.6. Estudios comparativos*

El corpus también puede utilizarse en estudios comparativos entre el lenguaje de la comunidad LGTBI y el lenguaje de otros grupos sociales. Esto puede incluir comparaciones entre diferentes ciudades, edades y niveles socioeconómicos, proporcionando una visión más amplia de la variación lingüística.

También permite estudios comparativos internos dentro del colectivo, tales como el análisis de diferencias entre el habla de personas gais, lesbianas, bisexuales, trans y no binarias, o entre grupos de diferentes contextos culturales. Comparar el lenguaje empleado en ciudades específicas o en función de factores como la clase social o la edad puede proporcionar una perspectiva más matizada de la variación y el cambio lingüístico en esta comunidad.

## *5.4. Problemas del corpus*

A pesar de los esfuerzos por construir un corpus oral amplio y representativo de la comunidad de habla LGTBI, existen varios problemas importantes que deben ser enfrentados para mejorar la calidad y la integridad de los datos recopilados. Los obstáculos más destacados incluyen la obtención de muestras de subgrupos menos visibilizados y la necesidad de incrementar los registros en modalidad de grabaciones secretas.

### *5.4.1. Obtención de muestras en subgrupos menos visibilizados*

Uno de los retos más significativos en la recopilación de datos es la inclusión de subgrupos menos visibilizados, como las personas no binarias o las intersexuales. Estos grupos suelen estar subrepresentados en los estudios lingüísticos, lo que limita nuestra comprensión de la diversidad y riqueza del uso del lenguaje en diferentes identidades de género. Su representación en las primeras muestras del corpus publicadas por Navarro-Carrascosa (2023a) es mínima.

Para abordar este problema es esencial desarrollar estrategias inclusivas y sensibles que permitan contactar y reclutar participantes de estos subgrupos, asegurando así su representación adecuada en el corpus. A continuación, se enumeran algunas posibilidades:

— Buscar la colaboración de organizaciones LGTBI para facilitar el acceso a participantes de estos subgrupos.

— Desarrollar campañas de sensibilización que informen y animen a la participación de personas no binarias e intersexuales, así como otras minorías menos representadas.

— Adaptar las metodologías de recolección de datos para ser más inclusivas y respetuosas con todas las identidades de género, asegurando que todos los participantes se sientan cómodos y representados.

### *5.4.2. Tamaño de la muestra*

Como se ha indicado anteriormente, este es uno de los principales obstáculos en la elaboración de los corpus orales parciales. En el caso del corpus oral de la comunidad de habla LGTBI para calcular cuál es una buena muestra de partida es pertinente conocer el número de personas en España que pertenecen al colectivo (asumiendo que el corpus se limita a este país, aunque posteriormente pueda crecer con otros países hispanohablantes). Sin embargo, «no todas las personas con una orientación sexual o una identidad de género no normativas lo declaran en público y, por lo tanto, los censos pueden ser engañosos» (Navarro-Carrascosa 2023b: 92). No obstante, las asociaciones LGTBI ofrecen algunos datos a partir de los cuales se pueden realizar cálculos aproximativos[14].

Por otro lado, además, es importante que las muestras sean representativas de todas las realidades LGTBI en proporción. Así, se establece un número aproximado de participantes que sean hombres cisgénero homosexuales, mujeres cisgénero homosexuales, mujeres transgénero, hombres transgénero, personas bisexuales (o pansexuales), personas intergénero, etc. Conseguir muestras representativas de todos los subgrupos resulta uno de los problemas que se presentan a la hora de elaborar el corpus. En la publicación de Navarro-Carrascosa (2023a), de los 60 informantes que participaron tanto en las entrevistas semidirigidas como en las grabaciones secretas, solo cinco de ellas eran bisexuales; y solo una, intergénero.

[14] En 2023 se proporcionan datos de la asociación FELGTBI+ que señalan que entre el 7 y el 8 % de la población española pertenece al colectivo LGTBI: https://felgtbi.org/wp-content/uploads/2023/11/I-Informe-Estado-socioeconomico_felgtbi.pdf.

### *5.4.3. Incrementar los registros en modalidad de grabaciones secretas*

Otro problema que se debe abordar es aumentar el número de registros obtenidos mediante grabaciones secretas. Este método resulta valioso porque puede proporcionar una visión más auténtica y espontánea del habla, minimizando la paradoja del observador, donde la presencia del investigador puede influir en el comportamiento lingüístico de los participantes. Sin embargo, este tipo de registro presenta varios desafíos legales y éticos. La legislación vigente no permite grabar sin el conocimiento y consentimiento previo de los participantes, por tanto, es importante preparar un documento en el que se informe a los informantes de que van a ser grabados en algún momento futuro sin previo aviso y estos lo firman a modo de autorización. Tal y como se especifica en el proyecto Ameresco[15], el consentimiento se firma en múltiples ocasiones para asegurar la claridad y el acuerdo continuo:

> Esta primera firma manifiesta el consentimiento del hablante a ser grabado y recoge la fecha en que lo autorizó. En el segundo apartado, una vez realizada la grabación y habiendo sido informado el hablante de que acaba de ser grabado, se le da la posibilidad de escuchar la conversación y, si consiente en cederla para su análisis lingüístico, deberá firmar y fechar el segundo apartado de la autorización. Con esta firma, por tanto, el hablante manifiesta que ha escuchado la conversación y que está de acuerdo con que se haga pública para fines de investigación, previa anonimización de nombres y lugares. En último lugar, los hablantes deben firmar la sección para el tratamiento de datos personales, de acuerdo con la normativa vigente, y aceptar los términos. En el caso de no obtener dicha autorización o de que esta esté incompleta, el archivo no podrá utilizarse y deberá ser destruido (Carcelén y Uclés 2019: 22).

Otro obstáculo en la implementación de la grabación secreta, relacionado con el cumplimiento de la ley de protección de datos, es el desconocimiento de algunos participantes por parte del investigador o la persona encargada de realizar la grabación. En este caso, encontrar el momento adecuado para la grabación sin que los participantes lo esperen puede ser más complicado. Se requiere, por tanto, de una planificación cuidadosa y un enfoque sensible para no comprometer la ética del estudio.

Tratar estos retos es complejo, no obstante, existen algunas acciones que se pueden proponer al respecto:

— Delegar la responsabilidad de las grabaciones secretas en otros participantes que, previamente, hayan firmado el consentimiento, junto con el resto de los informantes. A modo de infiltrado,

[15] Proyecto que estudia y analiza el español coloquial de los países americanos, dirigido por Marta Albelda y Maria Estellés (Universitat de València).

grabará las conversaciones sin que el resto de los participantes sepa que va a hacerlo y su presencia no condicionará la actuación lingüística y conversacional.

— Proponer varios encuentros del investigador con el grupo que va a ser grabado sin que los participantes sepan en cuál se desarrollará el registro.

— Uso de tecnologías discretas, siempre que haya una financiación por parte de un proyecto y/o institución. En caso de que existan fondos, se pueden implementar micrófonos ambientales, colocados en lugares comunes donde los participantes suelen estar, o dispositivos portátiles pequeños, como grabadoras que puedan ser llevadas en bolsillos o mochilas.

## 6. El futuro del corpus oral de la comunidad de habla LGTBI

El Corpus Oral de la Comunidad de Habla LGTBI ha sentado una base sólida para el estudio de la lingüística *queer* hispánica, pero su desarrollo y expansión son esenciales para una comprensión más completa y diversa del lenguaje dentro de esta comunidad. A continuación, y a modo de conclusión, se detallan los principales objetivos y planes futuros para el corpus:

— Inclusión de grupos más invisibilizados: como ya se ha especificado en apartados previos, resulta fundamental obtener más muestras de habla reales de personas pertenecientes a grupos más invisibilizados dentro del colectivo LGTBI. Esto incluye a personas transexuales binarias, intergénero, bisexuales, y categorías que hasta ahora no han sido registradas en el corpus, como asexuales, demisexuales, entre otros. La inclusión de estos subgrupos permitirá una visión más completa y representativa de la diversidad lingüística y cultural del colectivo LGTBI.

— Registro de habla de generaciones mayores: registrar más grabaciones con personas de la franja de edad más mayor es un objetivo clave para poder comparar la evolución del argot LGTBI a través de distintas generaciones. Esto abrirá la puerta a estudios diacrónicos en lingüística *queer*, permitiendo analizar cómo ha cambiado y se ha adaptado el lenguaje de la comunidad a lo largo del tiempo, reflejando cambios sociales, culturales y estructurales.

— Expansión geográfica del corpus: la primera versión del corpus se limitó a las tres principales capitales de España: Madrid, Barcelona y Valencia. Resulta fundamental ampliar los registros a más ciudades

y municipios para capturar una mayor diversidad de dialectos y usos lingüísticos. Además, buscar registros en núcleos más rurales será interesante para identificar y estudiar diferencias lingüísticas significativas que puedan existir entre las áreas urbanas y rurales.

— Inclusión de variantes hispanoamericanas: siguiendo la misma línea que el punto anterior, el corpus debe, en su futura expansión, buscar registros de las distintas variantes hispanohablantes en otros países de habla española. La inclusión de datos de comunidades LGTBI en América Latina y otras regiones hispanohablantes permitirá comparaciones interculturales y enriquecerá el corpus con una variedad más amplia de influencias lingüísticas y culturales.

## Bibliografía

Azorín Fernández, Dolores (2005), «Corpus oral para el estudio del lenguaje juvenil y del español hablado en Alicante. El corpus ALCORE y COVJA», *Oralia*, 8: 265-287.

Bejarano, Daniel, Andrea Llanos, Ruth Rubio, y Jonhatan Bonilla (2018), *Protocolo de transcripción ortográfica CLICC*, Bogotá, Instituto Caro y Cuervo.

Bengoechea Bartolomé, Mercedes (2015), *Lengua y género*, Madrid, Síntesis.

Briz Gómez, Antonio, y Grupo Val.Es.Co. (2002a), *Corpus de conversaciones coloquiales*, Madrid, Arco/Libros.

Briz Gómez, Antonio (2012), «Los déficits de los corpus orales del español», en Tomás Eduardo Jiménez Juliá, Belén López Meirama, Victoria Vázquez Rozas y Alexandre Veiga Rodríguez (coords.), *Cum corde et in nova grammatica: estudios ofrecidos a Guillermo Rojo*, Santiago de Compostela, Universidad de Santiago de Compostela: 115-137.

Calderón Noguera, Donald Freddy, y Julia Alvarado Castellanos (2011), «El papel de la entrevista en la investigación sociolingüística», *Cuadernos de Lingüística Hispánica*, 17: 11-24.

Carcelén Guerrero, Andrea, y Gloria Uclés Ramada (2019), «Diseño y construcción de un corpus oral multidialectal. El corpus AMERESCO», *Normas*, 9: 17-36.

Fernández Juncal, Carmen (2005), *Corpus de habla culta de Salamanca (CHCS)*, Burgos, Fundación Instituto Castellano y Leonés de la Lengua.

Fernández Sanmartín, Alba (2022), *Teoría y métodos para la elaboración de corpus orales: la entrevista sociolingüística*, Santiago de Compostela, Universidad de Santiago de Compostela.

Labov, William (1972), *Sociolinguistic patterns*, Philadelphia, University of Pennsylvania.

Mira, Alberto (1999), *Para entendernos: diccionario de cultura homosexual, gay y lésbica*, Barcelona, Ediciones de la Tempestad.

Mira, Alberto (2004), *De Sodoma a Chueca: una historia cultural de la homosexualidad en España en el siglo* xx, Madrid, Egales.

Moreno Fernández, Francisco (2005), «Corpus para el estudio del español en su variación geográfica y social. El corpus PRESEEA», *Oralia*, 8: 123-139.

Moreno Fernández, Francisco (2011), «La entrevista sociolingüística: esquema de perspectivas», *Linred. Lingüística en la red*, 9.

Navarro-Carrascosa, Carles (2019), «Resignificación y reapropiación en el español coloquial: el caso de *maricón*», en Adrián Cabedo Nebot y Antonio Hidalgo Navarro (eds.), *Pragmática del español hablado: hacia nuevos horizontes*, Valencia, Universitat de València: 169-183.

Navarro-Carrascosa, Carles (2023a), *Corpus oral de la comunidad de habla LGTBI: materiales para la investigación en lingüística* queer *hispánica*, Alcalá de Henares, Universidad de Alcalá.

Navarro-Carrascosa, Carles (2023b), *Lingüística* queer *hispánica: las formas nominales de tratamiento de la comunidad de habla LGTBI*, Berna, Peter Lang.

Pato, Enrique (2020), «El español en contacto con el francés en Quebec y su estudio gracias al 'Corpus Oral de la Lengua Española en Montreal' (COLEM)», *Boletín Hispánico Helvético*, 35-36: 263-287.

Pereda, Ferran (2004), *El cancaneo: diccionario petardo de argot gay, lesbi y trans*, Madrid, Laertes.

Pons Bordería, Salvador, y Leonor Ruiz Gurillo (2005), «Corpus para el estudio de la conversación coloquial: el corpus Val.Es.C.o. (Valencia. Español Coloquial)», *Oralia*, 8: 243-264.

PRESEEA (2021), *Metodología del «Proyecto para el estudio sociolingüístico del español de España y de América» (PRESEEA)*, Disponible en: http://preseea.linguas.net [Fecha de consulta: 03/03/2024]

Recalde, Montserrat, y Victoria Vázquez Rozas (2009), «Problemas metodológicos en la formación de corpus orales», en Pascual Cantos Gómez y Aquilino Sánchez Pérez (eds.), *A Survey on corpus-based research /Panorama de investigaciones basadas en corpus*, Murcia, Asociación Española de Lingüística de Corpus: 37-49.

Ridao Rodrigo, Susana (2021), «Aproximación a la transcripción de corpus orales: los símbolos de transcripción en corpus judiciales», *Revista de Llengua i Dret*, 77: 93-110.

Rodríguez González, Félix (2008), *Diccionario gay-lésbico*, Madrid, Gredos.

Sanmartín Sáez, Julia (1998), *Lenguaje y cultura marginal: el argot de la delincuencia*, Valencia, Universitat de València.

Sanmartín Sáez, Julia (1999a), *Palabras desde el talego: el argot en la prisión de Valencia*, Valencia, Edicions Alfons El Magnànim.

Sanmartín Sáez, Julia (1999b), *Diccionario de argot*, Madrid, Espasa.

Torruella, Joan, y Joaquim Llisterri (1999), «Diseño de corpus textuales y orales», en José Manuel Blecua, Gloria Clavería, Carlos Sánchez y Joan Torruella (eds.), *Filología e informática: nuevas tecnologías en los estudios filológicos*, Barcelona, Editorial Milenio/Universitat Autònoma de Barcelona: 45-77.

Carlota de Benito Moreno (ORCID: 0000-0001-9112-5471), Antonio Corredor Aveledo (ORCID: 0000-0002-0092-9667) y Elena Padrón Castilla (ORCID: 0000-0002-1103-8341) «Heterogeneidad e innovación en la isla de La Palma: aproximación sociolingüística y dialectal». *Biblioteca de Babel: Revista de Filología Hispánica*, vol. extra. 2 (2024), pp. 263-289.
**DOI**: 10.15366/bibliotecababel2024.extra2.011 · **ISSNe**: 2695-6349 · Recibido: 03/10/2024 · Aceptado: 15/11/2024

# Heterogeneidad e innovación en la isla de La Palma: aproximación sociolingüística y dialectal[1]

Carlota de Benito Moreno
*Universidad Autónoma de Madrid*
carlota.debenito@uam.es

Antonio Corredor Aveledo
*Université de Neuchâtel*
antonio.corredor@unine.ch

Elena Padrón Castilla
*Université de Neuchâtel*
elena.padron@unine.ch

**Resumen**: A partir de veinticuatro fenómenos lingüísticos de distintos niveles gramaticales, estudiamos la variación lingüística en cuatro enclaves de La Palma, con el objetivo de describir el grado de heterogeneidad y de innovación de las hablas de la isla de acuerdo con parámetros sociolingüísticos y geográficos. Usamos el corpus Ruricán, cuyas entrevistas se han anotado por medio de un procedimiento de doble escucha, permitiendo un análisis cuantitativo del comportamiento de los hablantes respecto de cada una de las variables estudiadas.

**Palabras clave**: español de Canarias, sociolingüística rural, dialectología, innovación, homogeneidad lingüística.

## Heterogeneity and innovation in La Palma: a sociolinguistic and dialectal approach

**Abstract**: Using twenty-four linguistic features of different grammatical levels, we analyse the linguistic variation in four towns of La Palma, in order to describe the degree of heterogeneity and innovation of the island's Spanish according to sociolinguistic and geographical parameters. We use the Ruricán corpus, whose interviews have been annotated through a double-listening procedure, allowing a quantitative analysis

[1] Este trabajo ha sido posible gracias al proyecto con referencia 197401 (Rural Sociolinguistics in the Canary Islands: Linguistic Innovation and Diffusion, RurICan) del Fondo Nacional Suizo (Swiss National Science Fundation).

of the behaviour of the speakers with respect to each of the variables under study.

**Keywords**: Canary Spanish, rural sociolinguistics, dialectology, innovation, linguistic homogeneity

## 1. Introducción

El propósito de este trabajo es contribuir al estudio del grado de heterogeneidad y de innovación de los espacios lingüísticos periféricos. Para ello, investigamos los patrones de variación del español hablado en cuatro enclaves de La Palma, a partir de veinticuatro variables de todos los niveles lingüísticos, estudiadas en un corpus de entrevistas. En § 2 se reflexiona sobre la variación lingüística en sus vertientes geográfica y social, mientras que en § 3 se caracteriza la variedad lingüística palmera. En § 4 se describe el corpus empleado y en § 5 se detallan los veinticuatro fenómenos. En § 6 se explica la metodología de anotación de los datos, cuyo análisis se ofrece en § 7. En § 8 se recogen las conclusiones.

## 2. Consideraciones teóricas

La sociolingüística desplazó el foco de interés a las ciudades desde el espacio rural, privilegiado por la dialectología. Esta, además, se centraba en los hablantes más «genuinos», creando la apariencia de que las comunidades rurales son lingüísticamente estáticas (cf. Vandekerckhove 2010, Villena Ponsoda 2010). La concepción dicotómica de campo y ciudad presenta a estas como espacios de innovación lingüística, frente al conservadurismo de los hablantes rurales, y las apuntala como centros de difusión de las innovaciones, debido a su mayor prestigio y población (Vandekerckhove 2010; Chambers y Trudgill 1980; Britain 2004).

Sin embargo, esta dicotomía ha sido cuestionada por Britain (2012, 2017), que observa que no hay por qué esperar procesos lingüísticos cualitativamente diferentes en ambos contextos. De hecho, la concepción (simplificada) expuesta anteriormente choca con observaciones bien establecidas, como que las comunidades rurales son lingüísticamente heterogéneas (Gauchat 1905; Borrego Nieto 1981; Parodi y Santa Ana 1997) o que los centros urbanos pueden difundir normas conservadoras, como ocurre en los procesos de estandarización, donde al prestigio urbano se suma la intervención planificada por una

autoridad, difundida por medios de comunicación y centros educativos, entre otros (López Serena / Méndez García de Paredes 2019), especialmente en el contexto hispanohablante, donde las Academias desempeñan un importante papel normativo (López Serena 2011). Por otro lado, López Izquierdo (2014) critica la oposición entre variedades innovadoras o conservadoras como una heurística para definir y clasificar lenguas o variedades de forma global, por ser excesivamente simplificadora.

Así, la sociolingüística rural debe iluminar las dinámicas variacionales fuera de los espacios urbanos. Este trabajo quiere contribuir a ello a partir de tres objetivos: 1) describir el comportamiento lingüístico de veinticuatro variables lingüísticas en La Palma; 2) comparar el grado de heterogeneidad lingüística de las comunidades rurales con las semiurbanas y urbanas, y 3) enmarcar los resultados en la discusión entre innovación y conservación.

## 3. El español de La Palma en el contexto canario e hispánico

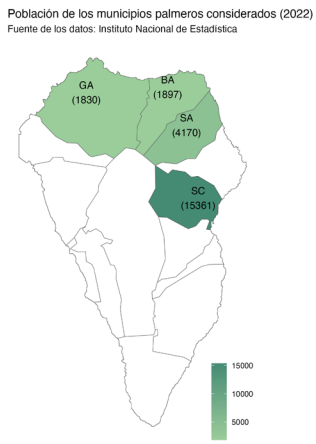
Es frecuente aludir al papel central de Canarias en la expansión atlántica del español. Dada su ubicación geográfica y su relevancia histórica en el dominio del Atlántico, se ha considerado puente entre el español americano y el peninsular y, por tanto, el archipiélago desempeña un papel central en la conformación del español en el continente americano (Catalán 1958; Álvarez Nazario 1972; Lüdtke 2014). La Palma se incorporó a la Corona de Castilla como isla realenga desde su conquista, que finalizó en 1493 (Aznar Vallejo 1983). Desde la segunda mitad del siglo xvi, Santa Cruz se convirtió en uno de los principales puertos atlánticos (Santana Pérez 2020) y la isla atrajo a un gran número de comerciantes —especialmente flamencos, genoveses y portugueses—. Este esplendor, que dotó de gran dinamismo a su sociedad en el siglo xvi, decayó en la primera mitad del siglo xvii por la pérdida de su papel central como puerto atlántico a favor de Santa Cruz de Tenerife, sobre todo en el comercio con América, del que dependía enormemente (Santana Pérez 2020). Desde entonces, el sustento principal de la población en la isla de La Palma ha sido la agricultura y la ganadería y su historia está marcada por la fuerte emigración de palmeros en el siglo xix y en la primera mitad del xx —sobre todo a Cuba y Venezuela— y el progresivo despoblamiento del norte de la isla. En la actualidad, la población de su capital, Santa Cruz, está disminuyendo: desde principios de siglo el núcleo urbano de mayor tamaño y crecimiento poblacional es Los Llanos de Aridane (INE).

En el marco de la discusión sobre el pluricentrismo de la lengua española, parece haber acuerdo en que Canarias no llega a constituir un centro fuerte de irradiación de la norma (Amorós-Negre y Prieto de los Mozos 2017), lo que permite considerar el español canario como una variedad periférica[2]. Sin embargo, el archipiélago contiene sus propios centros y periferias (Catalán 2003: 58-62): Las Palmas de Gran Canaria y Santa Cruz de Tenerife se consideran centros irradiadores de la innovación lingüística y, por tanto, del cambio (véase, entre otros, Samper Padilla 1996). Estas dinámicas pueden ilustrarse con la historia de las realizaciones de la /s/ implosiva en Canarias, con Las Palmas como centro difusión de las distintas fases del cambio (primero la aspiración; después la pérdida y alargamiento de la consonante siguiente), que se difunde de forma gradual por dos rutas: a las islas más cercanas y al otro centro urbano más importante (Santa Cruz de Tenerife), desde donde se extiende al resto de islas occidentales (Morera 2007). En las islas periféricas, entre las que se incluye La Palma, confluyen fenómenos conservadores y efectos de la influencia de su centro capitalino de referencia (cfr. Catalán 1989 [1964]), que, en el caso palmero, es Santa Cruz de Tenerife. Por otro lado, la cuestión de si existe una variedad estándar del español propia de Canarias está lejos de estar resuelta. Como señala Medina López, la dificultad de establecer el grado de estandarización de los rasgos lingüísticos isleños se debe, en gran parte, a la escasez de corpus y estudios «que cuantifiquen en términos estadísticos y cualitativos el avance o retroceso de ciertos fenómenos» (Medina López 1992-1993: 178, nota 6), no solo de las capitales, sino también de las áreas rurales.

En este trabajo nos centramos en rasgos concretos, en vez de en la descripción general de la variedad. Incluimos municipios con distinto grado de desarrollo urbano, para comparar sus dinámicas de estandarización: 1) uno de los principales centros urbanos de la isla (la capital, Santa Cruz de La Palma); 2) dos localidades fundamentalmente rurales (Garafía y Barlovento), de reducido tamaño, población dispersa y alto aislamiento geográfico, y 3) San Andrés y Sauces, que se sitúa entre estos dos extremos: buenas comunicaciones con Santa Cruz (especialmente en Los Sauces), con una población que dobla la de Garafía y Barlovento[3]. El mapa 1 ofrece la ubicación geográfica de estos municipios, junto con su población (en 2022).

[2] Usamos el término *periférico* para señalar, únicamente, la condición de alejado del centro, en este caso, de los centros de difusión de la norma, es decir, sin —evidentemente— ninguna connotación peyorativa.

[3] Las entrevistas se han realizado en distintos barrios o localidades del mismo municipio. En San Andrés y Sauces hemos entrevistado principalmente en San Andrés y Los Sauces, mientras que en Garafía y Barlovento hemos entrevistado también en núcleos como Gallegos, Franceses, Llano Negro, etc. Las entrevistas de Santa Cruz proceden generalmente del núcleo urbano, aunque también se entrevistó a algunas personas de zonas periféricas de carácter agrícola.




*Mapa 1. Población de los municipios palmeros considerados.*

## 4. El corpus

Nuestros datos proceden de entrevistas orales semidirigidas[4]. El corpus, resultado del proyecto Ruricán, se inspira en el Corpus Oral y Sonoro del Español Rural (COSER) (Fernández-Ordóñez 2005–), pero añade la dimensión social, entrevistando a múltiples hablantes por localidad. El corpus está formado por 113 entrevistas semidirigidas a adultos, recogidas en dos campañas (septiembre-diciembre de 2021 y julio de 2022). En la selección de los informantes se ha buscado la diversidad de género, de edad (con tres grupos: el 1, con hablantes menores de 35 años; el 2, con hablantes de entre 35 y 64 años, y el 3, con hablantes de más de 64 años) y de nivel educativo, donde distinguimos dos niveles (básico y superior), aunque el significado de estos no es el mismo en todas las generaciones. Mientras que en la primera y la segunda se corresponden con los niveles de educación reglada (en el nivel superior están aquellos que han realizado estudios universitarios o ciclos formativos superiores), en la tercera generación se ha clasificado dentro del nivel básico a aquellos que solo recibieron la educación obligatoria (generalmente hasta los 14 años), mientras que aquellos que continuaron su educación en distintos niveles (ciclos de formación profesional de cualquier nivel, estudios universitarios) se han clasificado dentro del nivel superior. En líneas generales, esta diferencia se asocia a dos perfiles profesionales distintos: respectivamente, uno agropecuario y otro liberal. Por último, además de buscar diversidad demográfica en la muestra, hemos tratado de entrevistar a personas

[4] Los estudios previos sobre el español palmero se basan en otro tipo de datos. Desde los años cuarenta del siglo pasado, La Palma cuenta con una nómina de filólogos interesados en describir el acervo léxico de la isla (p. ej., Pérez Vidal 1946, 1949, 1987; Régulo Pérez 1968-1969; Díaz Alayón 1983, 1990, 2020; Rodríguez Concepción 1991), así como, aunque en menor medida, algunos de sus rasgos gramaticales (Régulo Pérez 1968-1969; Leal Cruz 2003). La Palma tiene presencia también en el Atlas Lingüístico y Etnográfico de las Islas Canarias (ALEICan).

con relaciones familiares o de amistad entre sí, lo que explica que la muestra, cuya distribución demográfica se presenta en la tabla 1, no esté equilibrada.

| | | **GEt: 1** | **GEt: 2** | **GEt: 3** |
|---|---|---|---|---|
| Barlovento | NEd: B | 1H - 1M | 3H - 2M | 1H - 3M |
| (n = 22) | NEd: S | 1H - 2M | 1H - 4M | 2H - 1M |
| San Andrés y Sauces | NEd: B | 2H - 2M | 2H - 1M | 2H - 2M |
| (n = 26) | NEd: S | 2H - 3M | 2H - 4M | 2H - 2M |
| Santa Cruz de la Palma | NEd: B | – - 1M | 1H - 2M | 1H - 4M |
| (n = 31) | NEd: S | 2H - 3M | 8H - 5M | 1H - 3M |
| Garafía | NEd: B | 2H - 1M | 4H - 4M | 5H - 6M |
| (n = 34) | NEd: S | – - 4M | 5H - 3M | – - – |

*Tabla 1. Distribución demográfica y geográfica de los informantes*

Las entrevistas versan sobre el modo de vida pasada y presente en las diferentes localidades, lo que favorece que los informantes se sientan cómodos. Contábamos con algunas preguntas diseñadas para provocar algunos contextos de interés lingüístico, pero no se preguntó explícitamente por rasgos lingüísticos hasta el final de la entrevista, donde se orientó la conversación a explorar la conciencia y las actitudes lingüísticas de los informantes, indagando después sobre fenómenos lingüísticos concretos, especialmente las formas de tratamiento.

## 5. Los fenómenos estudiados

Para la selección de las veinticuatro variables lingüísticas consideradas, se partió de las encuestas palmeras del COSER, de las que se seleccionaron fenómenos con un cierto grado de variación y una frecuencia de aparición elevada[5]. Generalmente, de las dos (o, en ocasiones, tres) variantes que compiten, una es considerada estándar, frente a otra que se suele describir como subestándar, rural o vulgar. Es el caso, por ejemplo, de *nadie/nadien*, *yo he comido/yo ha comido* o *después/dispués*, que encajan, por tanto, en el español «estigmatizado» de Parodi y Santa Ana (1997). Otros rasgos no se ubican tan fácilmente dentro del diasistema o espacio variacional (Del Barrio 2018), pues no han sido apartados explícitamente de la norma, y se caracterizan mejor como regionalismos (o variantes del español regional, Parodi y Santa Ana 1997), que se oponen al español general —y no al estándar— (por ejemplo, el ascenso del cuantificador en las superlativas relativas, Peña Rueda 2022a; Medina López 2023)[6]. Para facilitar la visualización de los

[5] A pesar de esta preocupación previa, algunos de los fenómenos escogidos por presentar variación en las entrevistas del COSER apenas la muestran en las de Ruricán, como se verá más adelante, lo que puede quizá explicarse por el ámbito geográfico más reducido de este último.

[6] No es evidente, sin embargo, que puedan caracterizarse como formas propias de un estándar regional (cf. la discusión en López Serena 2011), ya que la propia idea de un estándar regional

datos en § 7, hemos asignado a cada una de las variantes una etiqueta neutral, a saber, V1, V2 y, cuando sea necesario, V3. En la mayoría de los casos, las variantes estándar o propias del español general llevan la etiqueta V1. Con todo, algunos casos escapan a esta clasificación, por gozar de consideración normativa escurridiza, como ocurre con las formas de 2ª pl. *ustedes* y *vosotros*, ambas perfectamente normativas.

Por último, para poner a prueba el carácter conservador de las hablas periféricas, clasificamos los fenómenos analizados según su carácter históricamente innovador o conservador, atendiendo a dos criterios: 1) en el caso de variantes etimológicamente relacionadas entre sí, la variante innovadora es la que muestra cambios producidos a partir de la otra variante, y 2) en el caso de variantes sin relación etimológica, son innovadoras las más tardías. Por tanto, tampoco todas las variantes pueden caracterizarse en estos términos.

## *5.1. Fenómenos de orden fonético*

Nuestro estudio no incluye fenómenos como el seseo, la aspiración de *-s* implosiva o la pronunciación aspirada de la /x/, rasgos que no presentan variación dentro del espacio palmero (o la que presentan debe estudiarse con técnicas más sofisticadas de las aquí empleadas).

### *5.1.1 -e paragógica tras /r/ final prepausal*

A los casos sin paragoge del tipo *trabajar*, *comer*, *bar* (V1) se oponen las formas con paragoge en el infinitivo: *trabajare*, *comere* (V2, véase (1a)) o en sustantivos, del tipo *bare* (V3, véase (1b)), documentadas en las «zonas rurales conservadoras» (Almeida y Díaz Alayón 1988: 122, Régulo Pérez 1968-1969: 42). La paragoge se da en las hablas del occidente peninsular (Castillo *et al.* 2022), probable origen de su presencia en las islas. El fenómeno existe, al menos, en las islas occidentales y Lanzarote (Alvar 1959, Castillo *et al.* 2022). Así pues, la paragoge es tanto un fenómeno subestándar como, históricamente, una innovación fonética.

(1) a. Nos escaldaba gofio para comere (Ruricán, BA-004)
b. La batata nos la escachaba ella con un tenedore (Ruricán, BA-004)

canario, especialmente en lo que respecta al nivel morfológico o sintáctico, es todavía discutida y discutible.

### *5.1.2. Cierre de vocales átonas (/e/~/i/)*

Evaluado en una forma léxica concreta, a saber, la palabra *después* (V1), con la variante (fonéticamente más innovadora) *dispués* (V2, véase (2)), que se documenta en las hablas peninsulares occidentales (Régulo Pérez 1968-1969: 134) y orientales (Zamora Vicente 1979 [1967]: 276) y se ha considerado vulgar (NGLE 2011: §3.7i).

(2) Para la… botarlos acá aquí dispués (Ruricán, BA-003)

### *5.1.3 Cierre de vocales átonas (/a/~/e/)*

También evaluado en un lexema concreto, a saber, la palabra *entonces* (V1). La variante innovadora, *antonces* (V2, véase (3)), se considera, como la vacilación de vocales en general, vulgar (NGLE 2011: §3.7i). Régulo Pérez (1968-1969: 35) la documenta en La Palma y Zamora Vicente (1979 [1967]: 199) recoge *antoncies* en leonés.

(3) Antonces se usaban las falditas grandes (SD-034)

### *5.1.4 Diptongación decreciente en el sufijo* -ero

Este fenómeno, también de raigambre occidental, se estudia en la palabra *tunera* (V1) / *tuneira* (V2). Régulo Pérez (1968-1969: 40) lo considera arcaísmo o portuguesismo: históricamente, el diptongo decreciente es más antiguo que su reducción (que sería, por tanto, la variante innovadora), pero como préstamo del portugués —interpretación más adecuada en este lexema— debe considerarse una innovación.

## *5.2. Fenómenos de orden morfológico*

Incluimos en este nivel fenómenos que, como señala Rodríguez Molina (2015: 1050, n. 2), presentan una variación que podría ser de orden fonético o morfológico (se trata de los casos de § 5.2.1 a § 5.2.3). Los consideramos morfológicos porque no son procesos generales, sino que su ámbito de actuación está restringido léxicamente, siendo especialmente notables en palabras gramaticales. Solo un análisis detallado de su comportamiento permitiría saber si las variantes tienen usos semánticos diferenciales, que no se han descrito en la bibliografía.

### *5.2.1 Paragoge consonántica de nasal final en el indefinido negativo:* nadie / nadien

Históricamente, la variante estándar *nadie* (V1) se opone a la más innovadora *nadien* (V2, véase (4)), que se ha considerado rural, popular o vulgar (Rosenblat 1946: 150, Almeida y Díaz Alayón 1988: 118, Régulo Pérez 1968-1969: 51, *NGLE* 2009 § 48.1c).

(4) Pero aquí antes nadien compraba verduras (Ruricán, SA-008)

### *5.2.2 Paragoge consonántica de nasal final en los adverbios de lugar:* aquí, ahí / aquín, ahín

A pesar de no haber sido descritas para el español de La Palma, en el COSER se documenta una variante innovadora, con nasal final, de los adverbios *aquí*, *ahí* (V1): *aquín*, *ahín* (V2, véase (5)), probablemente a semejanza de otros adverbios como *asín*.

(5) Ya te digo que la placa estaba allín (Ruricán, SC-002)

### *5.2.3* todavía / entodavía

La variante estándar, *todavía* (V1), es históricamente más conservadora que *entodavía* (V2, véase (6)), que se ha considerado vulgar, rural o popular (Álvarez Martínez 1987: 13, *NGLE* 2009 § 30.8f) y cuyo origen podría estar en la combinación de *en* y *todavía* (Morera 1999) o de *aún* y *todavía* (Fernández-Ordóñez 2011). Se ha registrado en todas las islas (Alvar 1959: 78; Régulo Pérez 1968-1969: 65, Almeida y Díaz Alayón 1988: 129).

(6) No quedarán muchas, pero sí quedan entoavía unas miles. (Ruricán, SD-001)

### *5.2.4 Apócope de* primera *en posición prenominal*

Ante sustantivo femenino, *primera* (V1) puede apocoparse en *primer* (V2, véase (7)), forma caracterizada como arcaísmo (*NGLE* 2009 § 21.4e) o poco culta (Almeida y Díaz Alayón 1988: 106). Es, históricamente, más innovadora, documentándose desde el siglo xvi (*NGLE* 2009 § 21.4d).

(7) Subes la cuestita esa, y coges la general, la primer casa a la derecha. (SD-010)

### *5.2.5 Morfema* -mos/-nos *en las formas verbales esdrújulas de 1ª pl.*

La forma estándar (*estábamos*, V1) se opone a la analógica y más innovadora (*estábanos*, V2, ver (8)), documentada en las variedades del español americano y europeo, sobre todo en hablantes mayores y de escasa instrucción (Catalán (1989 [1964]: 150; *NGLE* 2009: §16.1l; Pato 2015).

(8) Y llegábanos a casa y mamá nos guisaba un boniato. (Ruricán, BA-004).

### *5.2.6 Morfema* -amos/-emos *en las formas de 1ª pl. del pretérito perfecto simple.*

La forma estándar (*ayer cantamos*, V1) es más antigua que la innovadora *ayer cantemos* (V2, ver (9)), cuyo origen parece estar en la analogía con la 1ª sg. (Menéndez Pidal 1941: §118.4; García de Diego 1946: 317). V2 se ha descrito como un vulgarismo general o extendido (García de Diego 1946: 317, Menéndez Pidal 1941: §118.4), muy documentado en Canarias (ALEICan III, mapa 1141; Almeida y Díaz Alayón 1988: 121; Catalán 1989 [1964]: 196; Ortega Ojeda 1987-1988; Leal Cruz 2003: 126).

(9) Ya después compremos un coche y salíamos por ahí (Ruricán, SD-008)

### *5.2.7 Morfema* -ron/-ran *en las formas de 3ª pl. del pretérito perfecto simple*

A la variante estándar (*ayer cantaron*, V1) se le opone *ayer cantaran* (V2, ver (10)), que Leal Cruz (2003: 225) documenta ampliamente en La Palma y que considera un portuguesismo, por lo que la consideramos una innovación por contacto.

(10) Ahí se cargó las últimas varas que se cargaran ahí (Ruricán, SD-001)

### *5.2.8 Morfología del auxiliar* haber *en el pretérito perfecto compuesto*

La forma estándar *yo he comido* (V1) alterna con la forma analógica (e innovadora) *yo ha comido* (V2, ver (11)), que se ha considerado vulgar

(García de Diego 1946: 318; Catalán 1989 [1964]: 195-196; Álvarez Martínez 1987: 14; Almeida y Díaz Alayón 1988: 127).

(11) Digo que eso lo ha vivido yo. (Ruricán, BA-004)

### *5.2.9 Incremento velar en los subjuntivos de* haber *e* ir

Las formas subestándar *haiga*, *vaiga* (V2, ver (12)) son una innovación respecto de *haya*, *vaya* (V1), pues el incremento velar es analógico —y tardío (de Benito Moreno 2020)—. Se consideran un vulgarismo extendido (Menéndez Pidal 1941: §113.2b; García de Diego 1946: 318) y están bien documentadas en Canarias (Régulo Pérez 1968-1969: 61, 64; ALEICan (III, mapa 1162); Almeida y Díaz Alayón 1988: 126-127).

(12) a. Yo ahora no recuerdo que me haigan echado algún cuento (Ruricán, SD-026)
b. No quieres que vaiga (Ruricán, BA-004)

### *5.2.10 Reducción del lexema del verbo* ver

En el imperfecto del verbo *ver* conviven formas del tipo *veía* (V1), propias del estándar, y las que pierden la vocal lexemática, como *vía* (V2, ver (13)), registradas en La Palma (Leal Cruz 2003: 130), zonas rurales de Canarias (Almeida y Díaz Alayón 1988: 126) y de Murcia (Zamora Vicente 1979 [1967]: 343), y consideradas arcaicas (Zamora Vicente 1979 [1967]: 343) y vulgares (García de Diego 1946: 318). La naturaleza conservadora o innovadora de estas variantes es difícil de establecer en términos históricos, pues ambas están testimoniadas desde la Edad Media y probablemente se deban a un doble paradigma *veer* / *ver*, como señalan Alvar y Pottier (1983: §120.1, n. 27).

(13) Yo vía a la gente antes más unidas para las cosas que hoy (Ruricán, SD-034)

### *5.2.11 Reducción del lexema del verbo* haber

La reducción de la raíz, por aféresis de la /a/ inicial, en el verbo *haber* (*había*, *haber*, *habemos* (V1) frente a *bía*, *ber*, *vemos*, V2, ver (14)) es una innovación considerada popular o arcaica y documentada en Andalucía y Canarias (Alvar y Pottier 1983: §158; Almeida y Díaz Alayón 1988: 127, Alvar 1959: 55)[7].

[7] Hay que señalar que los fenómenos 2.5.6 y 2.5.7 plantean un problema metodológico, pues pueden llegar a confundirse, ya que las variantes V2 son homófonas, lo que puede dificultar la

(14) El volcán le bía jodido una parte de la casa (Ruricán, BA-003)

### *5.2.12 Las formas de tratamiento de segunda persona plural*

En Canarias se usa mayoritariamente ustedes como única forma de 2ª pl., pero se documentan usos tradicionales de *vosotros* y su paradigma en las islas occidentales (Régulo Pérez 1968-1969: 47; Catalán 1989 [1964]: 147-148; Lorenzo Ramos 2003). El paradigma de *vosotros* puede también documentarse en todas las islas por efecto del estándar nacional (Ortega Ojeda 1981; Medina López 1992-1993; 2013; Ortega Ojeda y García Rivero 2020). Además, en La Palma documentamos formas híbridas de paradigmas de 2ª pl. y 3ª pl[8]. Para limitar la complejidad del fenómeno, nos limitamos a dos contextos: a) las formas verbales de 2ª pl. y b) la forma del sujeto explícito en combinación con formas verbales de 2ª pl. En el caso de a), se han anotado dos variantes: presencia de desinencias de tercera persona gramatical (V1) y presencia de desinencias de segunda persona gramatical (V2), ilustradas en (15a) en el mismo hablante. En el caso de b), tenemos tres variantes: *ustedes* + formas de 3ª pl. (V1, véase (15b)); *vosotros* + formas de 2ª pl. (V2, véase (15c)), y *ustedes* + formas de 2ª pl. (V3, véase (15d))[9].

(15) a. Digo: «no, ¡eh! Ustedes no me muevan eso!» Porque... ya somos viejos pero no quiero que... Mira, yo tengo leche condensada si queréis eh... (Ruricán, BA-003)
b. Lo que pasa es que yo tengo el coche averiado y no lo tengo ahora. Si no les llevaba a Mirca. Si ustedes quieren ir a Mirca, ves ets-| ta y tal… (Ruricán, SC-002)
c. pero vosotros l-| hoy en día, miro yo que sí, y y si ya tenéis pues dos o tres vacas, no fuera a decir que tengáis diez o que tengáis pero, si teniendo… (Ruricán, BA-003b)
d. En el camino y dice: «Y le dices a tu mujer que camine. Y si ella no pega aquí estáis ustedes caminando. Para que ese salga». (Ruricán, SD-005)

anotación: *bía mucha gente* [<había mucha gente] vs. *ví(a) mucha gente* [< veía a mucha gente]. Los casos ambiguos se han descartado.

[8] Álvarez Martínez (1987: 14) califica estos usos híbridos como «excepciones aisladas, pertenecientes a hablas muy vulgares».

[9] En aquellos hablantes que disponen de más de una de estas variantes, no es descartable que estas se pongan al servicio de consideraciones pragmáticas, asunto que se tratará con más detenimiento en el marco del proyecto Ruricán (Padrón Castilla en preparación).

### *5.3. Fenómenos del orden de la sintaxis*

Por motivos operativos, hemos restringido muchos de los fenómenos sintácticos —como los fonéticos— a contextos léxicos específicos, lo que facilita su anotación (véase § 6).

#### *5.3.1* a veces/veces

Hay variación entre la presencia o ausencia de la preposición en la locución adverbial *a veces* (V1, variante general) y *veces* (V2, véase (16)), que se documenta profusamente en Canarias (Álvarez Martínez, 1987: 13; Álvarez Martínez 1996: 73, nota 16; Leal Cruz 2003: 139), en todos los niveles del habla (Almeida y Díaz Alayón 1988: 134).

(16) Y los niños veces no se crían con muchos valores (Ruricán, SC-028)

#### *5.3.2 Preposiciones en adjuntos temporales de edad*

Tres preposiciones conviven en los adjuntos temporales de edad: *a* (V1, véase (17a)), *de* (V2, (17b)) y *con* (V3, (17c)). V2 es la más antigua y está restringida al ámbito rural en el español peninsular (De Benito Moreno 2020), mientras que V1 y V3 son generales en el español europeo.

(17) a. Pero antes, realmente el colegio, ¿qué se empezaba, a los 6? (Ruricán, SA-007)
b. Después que los tenía criados, se murió. A mí me dejó […] de ocho años (Ruricán, SA-004)
c. Con 10 años, así, empezaban a estudiar música para meterse en la banda (Ruricán, SA-007)

#### *5.3.3* Haber *en expresiones temporales*

*Hacer* impersonal en expresiones temporales (*hace muchos años*, V1) convive con *haber* (*hay muchos años*, V2), documentado en Canarias (Almeida y Díaz Alayón 1988: 127; Leal Cruz 2003: 135) y en andaluz occidental (Fernández-Ordóñez 2016). V2 es ya medieval, mientras que *hacer* aparece en el Siglo de Oro (Díez Itza 1992; Pérez Toral 1992). En (18) se registran ambas formas en el mismo enunciado.

(18) Había un convento no hace… ¿qué habrá, mamá? Ciento veinte años (Ruricán, SA-007)

### *5.3.4* sobre de / sobre

La preposición *sobre* presenta, además del uso estándar en solitario (*sobre la mesa*, V1), la posibilidad del régimen indirecto, con *de* (*sobre de la mesa*, V2, véase (19a)), y la combinación con *por* (*por sobre (de) la mesa*, V3, véase (19b)), variantes documentadas en Canarias (Alvar 1959: 79; Álvarez Martínez 1987: 13; Leal Cruz 2003:150). Tanto V2 como V3 se documentan desde la Edad Media (Octavio de Toledo 2016: 54, 210).

(19) a. Y yo me ponía allí sentada sobre de una piedra allí (Ruricán, SD-004)
b. Le raspé la pint-| le saqué un montón de pintura por sobre el capó (Ruricán, SD-004)

### *5.3.5* somos muchos / habemos muchos

El uso existencial de *haber* en 1ª pl. (V2, véase (20)) es innovador respecto del uso de *ser* o *estar* (V1)[10]. V2 se registra en todas las variedades del español, incluida Canarias (Almeida y Díaz Alayón 1988: 127; Álvarez Martínez 1987: 14; Leal Cruz 2003: 136), con marcación diastrática y diafásica variable, aunque esté desterrada por la norma académica (Castillo Lluch y Octavio de Toledo 2016). En Canarias no parece tener la marcación baja que tiene en la península (Álvarez Martínez 1987: 14).

(20) Si estos son cuatro pelagatos los que habemos aquí (Ruricán, SA-003)

### *5.3.6* recuerdo / me recuerdo

La variante no pronominal (*recuerdo*, V1 (21a)) y la pronominal (*me recuerdo*, V2 (21b)) se documentan desde el Medioevo, aunque esta es menos prestigiosa (*NGLE* 2009: § 36.3d)[11]. Aunque entre ambas puede haber una diferencia sintáctica (ilustrada en (21), véase *NGLE* 2009: § 43.6q), parecen ser sinónimas.

(21) a. Yo no recuerdo ver a mis abuelos haciendo chorizos ni morcillas (Ruricán, BA-023)
b. Porque yo me recuerdo muchísimo de ingeniarnos cosas (Ruricán, BA-023)

[10] Pero quizá no respecto de *haber* no concordado en 3ª pl., véase Castillo Lluch y Octavio de Toledo (2016).
[11] Anotamos únicamente la 1ª singular, por ser la forma que con más frecuencia presenta la variante pronominal en las hablas rurales peninsulares (De Benito Moreno 2015).

### *5.3.7 Ascenso del cuantificador en superlativas relativas*

En español general, el cuantificador de las superlativas relativas sigue al pronombre relativo (*lo que más me gusta es el queso*, V1), pero en Canarias y Puerto Rico puede precederlo (*lo más que me gusta es el queso*, V2 (22)) (Álvarez Martínez 1987: 20-21; Álvarez Martínez 1996: 74-79; Almeida y Díaz Alayón 1988: 135; Peña Rueda 2022a). V2 no goza de consideración baja, aunque en Canarias su uso presenta cierta estratificación (Peña Rueda 2022a). Los primeros ejemplos de V2 podrían datar del s. xiii (Peña Rueda 2022a), siendo así coetáneos a los de V1.

(22) Sobre todo el turismo alemán es el más que viene a La Palma (Ruricán, SC-015)

### *5.3.8 Posición de* ya *respecto del sujeto (1ª sg.)*

En la alternancia entre *yo ya* (V1, véase (23a)) y *ya yo* (V2, véase (23b)) no es fácil discriminar diferencias sociales (Frago Gracia 2003: 73) o de antigüedad: ambas se documentan ya en textos medievales (Peña Rueda 2022b)[12].

(23) a. Yo ya no tengo familia en Gran Canaria (Ruricán, SC-027)
b. Tengo un jardín muy grande y entonces ya yo soy feliz (Ruricán, SC-027)

## 6. Metodología de anotación y análisis

Las variantes no se han anotado a partir de las transcripciones de las entrevistas, todavía no disponibles, sino a partir de dos escuchas atentas por entrevista, durante las cuales se anotó la aparición de cada variante en una plantilla. Las anotaciones se compararon posteriormente, solucionando las discrepancias en una tercera escucha[13]. Para cada fenómeno y hablante existen distintas posibilidades, a saber: que se haya documentado una única variante; que se hayan documentado dos o más variantes, o que no dispongamos de información al respecto, por no haber ocurrido el contexto relevante. Presentamos un análisis cuantitativo y exploratorio, a partir de la visualización del comportamiento de los hablantes respecto de cada una de las variables,

[12] También aquí nos limitamos al pronombre nominativo de 1ª sg., por ser la combinación más frecuente (cfr. Peña Rueda 2022b).
[13] Salvo que las discrepancias estuvieran en la marcación de uno de los contextos propios de las variedades de los anotadores, que, por serlo, son más difíciles de detectar solo por medio de la escucha. En estos escasos casos se decidió confiar en la anotación explícita del contexto por parte de uno de los investigadores.

atendiendo a estas distintas posibilidades. Como se explicó en § 5, las variantes se denominan con etiquetas genéricas (V1, V2 o V3).

## 7. Análisis

Abordamos primero el análisis dialectal. La figura 1 muestra el número de hablantes que, en cada localidad, presenta los distintos comportamientos posibles para cada fenómeno. Aunque estos presentan variación entre sí, se pueden extraer algunas generalizaciones. En primer lugar, la mayoría de las variantes se documentan en todas las localidades, lo que sugiere que la variación lingüística en la isla de La Palma, al menos en su mitad nororiental, no está principalmente determinada por el eje geográfico. La excepción son algunas variantes muy poco frecuentes, documentadas en una o dos localidades: los adverbios con nasal final (*aquín*, *ahín*); el cierre vocálico en *dispués*; el morfema de pasado de 3ª pl. *–ran*, y la forma *entoavía*, cuya baja frecuencia impide certificar que se trate de diferencias dialectales.

En segundo lugar, la frecuencia de las variantes no es la misma en todas las localidades. En general, se observa un patrón según el cual las formas menos marcadas (ya sea como variantes rurales o subestándar o como formas no generales) aparecen más en Santa Cruz y en San Andrés y Sauces. Es el caso de la *-e* paragógica (cuya variante más extrema, en sustantivos, no se documenta en estas localidades); del uso de *hay* con valor temporal; del incremento velar en el subjuntivo de *haber* e *ir*, del auxiliar de 1ª sg. *ha*, del morfema de imperfecto de 1ª pl. *-mos*, de las formas de *vosotros* y de la forma reducida del imperfecto de *ver*, que son más frecuentes en Barlovento y Garafía; las localidades de menor tamaño, de economía más rural y de ubicación más aislada. Es decir, la distribución observada no parece estar determinada por el factor geográfico en sí mismo, sino por el perfil socioeconómico de las localidades.

Por último, no siempre encontramos esta jerarquización geográfica. En algunos fenómenos no se observan diferencias entre las localidades: es el caso de la diptongación de *tuneras* (no documentada), de la concordancia de 1ª pl. de *haber* existencial, del uso de *veces* ‘a veces’ y, prácticamente, del incremento nasal con el indefinido *nadie* (menos frecuente en Barlovento), la apócope de *primera* y el ascenso de *más* en las superlativas (menos comunes en Santa Cruz).

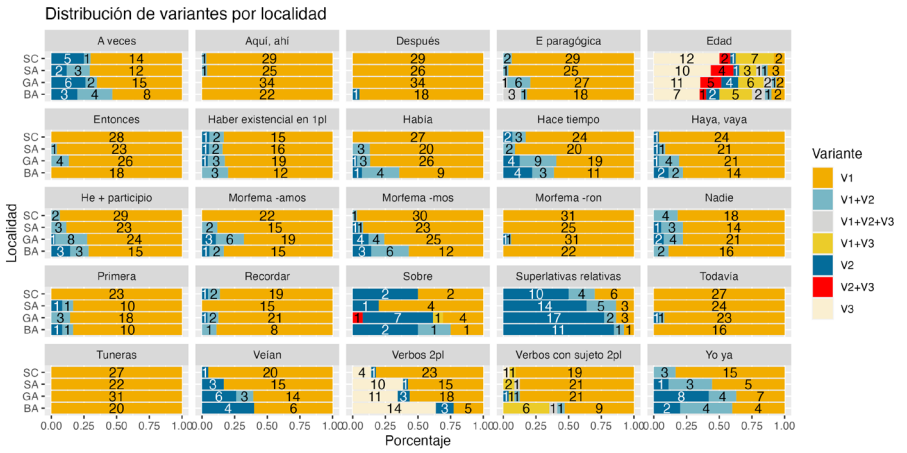



*Figura 1. Distribución de variantes por localidad.*

A continuación, exploramos los mismos datos según el nivel de estudios y la edad de los hablantes, agrupados así en seis categorías[14]. La figura 2 recoge los resultados de los cuatro fenómenos fonéticos investigados, que muestran un patrón bastante homogéneo: en tres de ellos la variante estándar (V1) es completamente general (no documentamos *tuneira*, forma que sí documenta el COSER) o abrumadoramente mayoritaria (solo un hablante, del grupo 3_B, produjo la forma *dispués*, mientras que cinco hablantes, de los grupos 1_B y 3_B, produjeron la forma *antonces*). La excepción a este alto nivel de estandarización es la paragoge de *-e* (V2 y V3), bastante más frecuente (14 hablantes proporcionan ejemplos). La diferencia podría estribar en una cuestión metodológica, pues es el único caso no limitado a un ítem léxico. Sea como fuere, se observa una clara estratificación social tanto de V2 —que solo no aparece en el grupo 1_S; despliega una frecuencia directamente proporcional al grupo etario, y es siempre menos frecuente en los hablantes de nivel educativo alto— como de V3 —atestiguada solo en hablantes de los grupos 2_B y 3_B, siendo más frecuente en este último.

[14] Las etiquetas numéricas corresponden a los grupos etarios, mientras que las alfabéticas indican el nivel educativo (básico o superior), siguiendo las convenciones usadas en § 4.

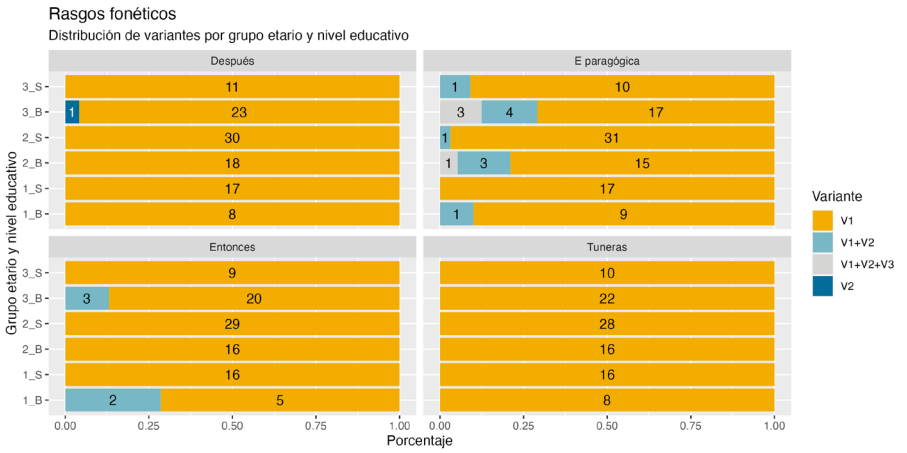



*Figura 2. Distribución de variantes fonéticas por grupo etario y nivel educativo*

La variación morfológica se ilustra en las figuras 3 y 4. En tres de los fenómenos estudiados (la desinencia *-ron* de la 3ª pl. del pasado, la paragoge de *-n* en *aquí*, *allí* y *ahí*, y la forma de *todavía*) encontramos que las respectivas formas no estándar (V2) apenas se documentan, con la excepción de dos hablantes (grupo 3_B) en el caso de la desinencia de 3ª pl.; de dos hablantes (grupos 3_B y 3_S) en el de la paragoge, y dos hablantes (grupos 1_B y 1_S) en el de *entodavía*. En el resto de los casos de morfología verbal (figura 3), la variación aparece claramente estratificada en términos sociales. En primer lugar, las formas no estándar no se documentan en los hablantes más jóvenes, con una única excepción: un hablante del grupo 1_S emplea *haiga* (ejemplo de (12a)). En el resto de los grupos etarios, todas las variantes no estándar analizadas son más frecuentes entre los hablantes de la generación 3 que entre los de la 2 y en hablantes con estudios básicos que en aquellos con estudios superiores.

En el resto de los casos de morfología no verbal la estratificación social no es tan clara, en cambio. Sí se observa con la forma *nadien*, documentada con más frecuencia en la generación 3, independientemente del nivel de estudios, así como en el grupo 2_B, aunque también se documenta en dos hablantes de los grupos 1_S y 2_S. Sin embargo, la forma femenina *primer* se documenta en las generaciones 2 y 3 y en todos los niveles de estudios, sin una jerarquía claramente determinada por estos parámetros (pero nótese que el contexto fue difícil de documentar). En cuanto a las formas verbales con valor de 2ª pl., todos los grupos documentan formas de 2ª pl., generalmente en convivencia con las de 3ª, aunque, sorprendentemente, con menos frecuencia en los grupos de mayor edad y menor nivel de estudios. Cuando el sujeto está explícito, la forma más empleada es *ustedes* + 3ª pl. (V1), que evidentemente forma parte del repertorio de todos los hablantes, pero

no se observan diferencias de importancia entre generaciones o niveles educativos. Todos los grupos documentan casos de *ustedes* + 2ª pl. (V2), más raramente, y/o, más frecuentemente, usos híbridos (V3).

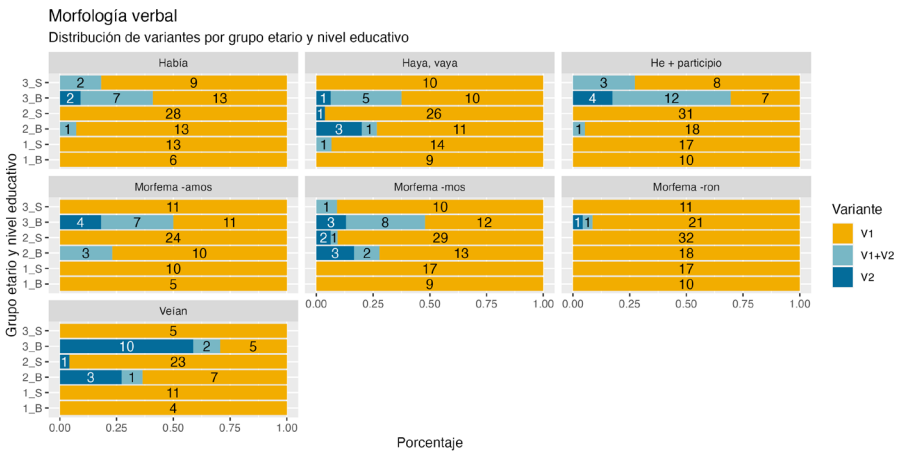



*Figura 3. Distribución de variantes morfológicas (ámbito verbal) por grupo etario y nivel educativo*

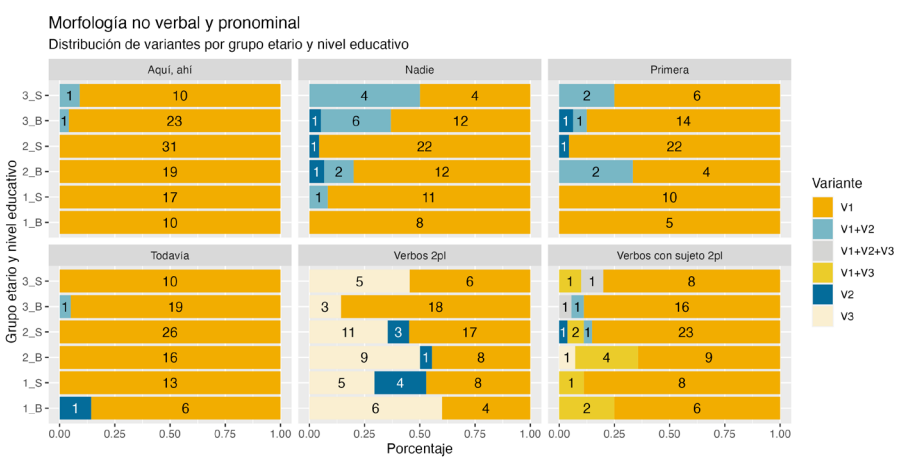



*Figura 4. Distribución de variantes morfológicas (ámbito no verbal) por grupo etario y nivel educativo*

En las variantes sintácticas (figura 5) se observa una mayor frecuencia de las variantes no estándar o no generales (V2 y/o V3). En algunos casos hay también indicios de estratificación. Así, la forma *veces* (frente a *a veces*) parece estar estratificada por edad, pues no se documenta en la generación 1 y es más frecuente en la 3. El uso pronominal de *recordar*, de baja frecuencia global, solo se documenta en hablantes de las generaciones 2 y 3. En el caso de las superlativas relativas, la forma regional (V2) es la más frecuente en términos globales, pero los grupos

que menos la usan son 1_S y 2_S. Tanto *haber* temporal (V2) como la preposición de *en* adjuntos etarios (V2) son sustancialmente más frecuentes en el grupo 3_B, aunque en el resto de los grupos no se observa ningún orden relativo evidente. El caso de la preposición *sobre* es más peliagudo, por haber sido el contexto especialmente difícil de documentar. En cualquier caso, la variante *sobre de* (V3) alcanza frecuencias no despreciables en los grupos 2_B y 3_B. La anteposición de *ya* no está socialmente estratificada, lo que es esperable si el orden responde a factores contextuales (Peña Rueda 2022b). Tampoco lo están las formas *habemos/habíamos* existenciales, a pesar de estar claramente censuradas por la norma. Hay indicios, empero, de que los hablantes no son conscientes de esta censura y de que, antes al contrario, las consideran arcaicas, formales, elaboradas o incluso cultas (Peña Rueda 2024).

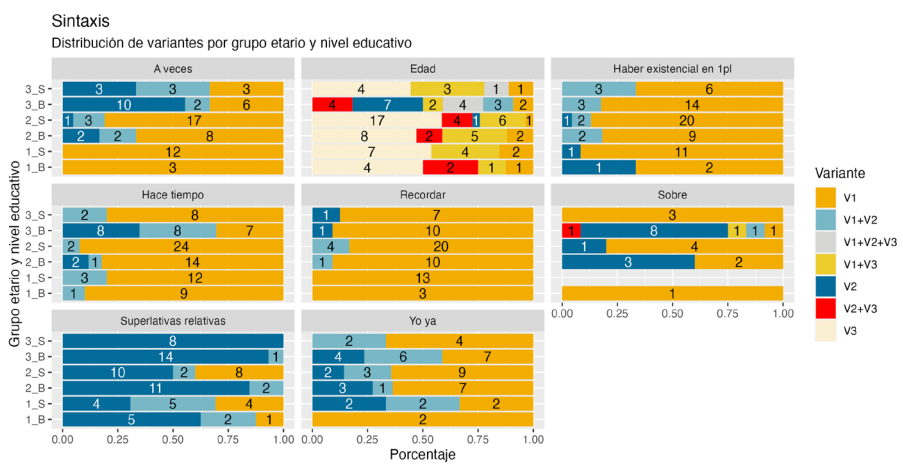



*Figura 5. Distribución de variantes sintácticas por grupo etario y nivel educativo*

## 8. Discusión y conclusiones

Nuestros resultados descriptivos pueden resumirse en dos generalizaciones y una advertencia. Primero, la variación lingüística observada no se organiza en torno a parámetros geográficos, sino sociales, relacionados con la edad y el nivel de estudios, como corresponde a un proceso de estandarización, en el que los hablantes más jóvenes y de mayor nivel educativo abandonan las formas marcadas. Segundo, la estandarización se observa más en los ámbitos fonético y morfológico, mientras que los rasgos sintácticos marcados despliegan mayor permeabilidad social. Esto puede deberse tanto a una menor conciencia de la norma sintáctica como a una mayor dificultad del control consciente de estos rasgos. Por último, no nos hallamos ante un proceso lineal y uniforme de abandono de rasgos marcados, sino que cada fenómeno

tiene su propia historia, escapando algunos a estas generalizaciones, quizá por estar envueltos en dinámicas sociales locales o porque las variantes estudiadas se vean afectadas por factores contextuales: se hace necesario un análisis detallado de cada fenómeno para el futuro.

En cuanto al grado de homogeneidad lingüística y al carácter innovador, nuestros datos no confirman la idea de que las ciudades son más heterogéneas. El análisis muestra que este parámetro es poco explicativo, pero que, cuando hay diferencias entre los enclaves más rurales (Barlovento y Garafía) y los más urbanos o semiurbanos (Santa Cruz y San Andrés y Sauces), son los primeros los que presentan mayor variación y mayor frecuencia de variantes innovadoras. Esto es razonable en un proceso de estandarización: las formas no marcadas suelen ser más conservadoras históricamente y triunfar más rápidamente en las ciudades.

## Bibliografía

Arad, Maya (1998), *VP-Structure and the syntax-lexicon interface*, tesis doctoral, University College London.

[ALEICan] = Alvar, Manuel (1975-1978), *Atlas Lingüístico y Etnográfico de las Islas Canarias (ALElCan)*, Las Palmas de Gran Canaria, Ediciones del Excmo. Cabildo Insular de Gran Canaria.

Almeida, Manuel y Carmen Díaz Alayón (1988), *El español de Canarias*, Santa Cruz de Tenerife, S/E.

Alvar, Manuel (1959), *El español hablado en Tenerife*, Madrid, Consejo Superior de Investigaciones Científicas.

Alvar, Manuel y Bernard Pottier (1983), *Morfología histórica del español*, Madrid, Gredos.

Álvarez Martínez, María Ángeles (1987), *Rasgos gramaticales del español de Canarias*, La Laguna, Instituto de Estudios Canarios.

Álvarez Nazario, Manuel (1972), *La herencia lingüística de Canarias en Puerto Rico: estudio histórico-dialectal*, San Juan de Puerto Rico, Instituto de Cultura Puertorriqueña.

Amorós-Negre, Carla y Emilio Prieto de los Mozos (2017), «El grado de pluricentrismo de la lengua española», *Language Problems and Language Planning*, 41 (3): 245–64. DOI: 10.1075/lplp.00004.amo.

Aznar Vallejo, Eduardo (1983), *La integración de las Islas Canarias en la Corona de Castilla (1478-1520)*, Madrid, Secretariado de

publicaciones de la Universidad de La Laguna y Universidad de Sevilla.

Borrego Nieto, Julio (1981), *Sociolingüística rural: investigación en Villadepera de Sayago*, Salamanca, Ed. Universidad de Salamanca.

Britain, David (2004), «Geolinguistics: diffusion of language», en Ulrich Ammon, Norbert Dittmar, Klaus Mattheier y Peter Trudgill (eds.), *Sociolinguistics: international handbook of the science of language and society*, Berlin, Mouton De Gruyter: 34-48. DOI: 10.1515/9783110141894.1.1.34.

Britain, David (2012), «Countering the urbanist agenda in variationist sociolinguistics: dialect contact, demographic change and the rural-urban dichotomy», en Sandra Hansen, Christian Schwarz, Philipp Stoeckle y Tobias Streck (eds.), *Dialectological and folk dialectological concepts of space*, Berlín, de Gruyter: 12-30. DOI: 10.1515/9783110229127.12.

Britain, David (2017), «Which way to look?: Perspectives on "Urban" and "Rural" in dialectology», en Emma Moore y Chris Montgomery (eds.), *A sense of place: studies in language and region*, Cambridge, Cambridge University Press: 171-188. DOI: 10.1017/9781316162477.010.

Castillo Lluch, Mónica y Álvaro Octavio de Toledo y Huerta (2016), «*Habemos muchos que hablamos español*: distribución e historia de la concordancia existencial en primera persona de plural», en Carlota de Benito Moreno y Álvaro Octavio de Toledo y Huerta (eds.), *En torno a 'haber': construcciones, usos y variación desde el latín hasta la actualidad*, Nueva York / Frankfurt, Peter Lang: 111-168.

Castillo Lluch, Mónica, Cristina Peña Rueda, y Michiel De Vaan (2022), «¿Pronunciar o pronunciare? Esa es la cuestione», en Ana Estrada, Beatriz Martín y Carlota de Benito (eds.), *Como dicen en mi pueblo: el habla de los pueblos españoles*, Madrid, Pie de Página: 63-75.

Catalán, Diego (1958), «Génesis del español atlántico: ondas varias a través del océano», *Revista de Historia Canaria*, 123-124:. 233-242.

Catalán, Diego (1989 [1964]), «El español en Canarias», en El español. Orígenes de su diversidad, Madrid, Paraninfo: 145-201.

Catalán, Diego (2003), «Centralidad teórica de las hablas fronterizas», en Carmen Díaz Alayón, Marcial Morera y Gonzalo Ortega (eds.), *Estudios sobre el español de Canarias. Actas del I Congreso Internacional sobre el español de Canarias*, Islas Canarias, Academia Canaria de La Lengua: 43-62.

Chambers, J. K. y Peter Trudgill (1980), *Dialectology*, Cambridge, Cambridge University Press.

De Benito Moreno, Carlota (2015), *Las construcciones con "se" desde una perspectiva variacionista y dialectal*, tesis doctoral, Universidad Autónoma de Madrid.

De Benito Moreno, Carlota (2020), «Reflexiones sobre la "lengua vulgar dialectal" y el vulgarismo», en Inés Fernández-Ordóñez (ed.), *El legado de Ramón Menéndez Pidal (1869-1968) a principios del siglo XXI*, Madrid, Consejo Superior de Investigaciones Científicas: Vol. 2, 19-56.

Del Barrio De La Rosa, Florencio (2018), *Espacio variacional y cambio lingüístico en español*, Madrid, Visor.

Díaz Alayón, Carmen (1983), «Nuevas aportaciones al léxico de la lluvia en La Palma», R*evista de Filología (Universidad de La Laguna)*, 2: 71-82.

Díaz Alayón, Carmen (1990): «Notas de dialectología canaria: el léxico palmero», *Revista de Filología (Universidad de La Laguna)*, 8-9: 127-144.

Díaz Alayón, Carmen (2020), «El español de La Palma: una mirada a sus peculiaridades», en Manuel Poggio Capote, Víctor J. Hernández Correa y Antonio Lorenzo Tena (eds.), *Cinco mitos para cinco siglos: 525º aniversario de la fundación de Santa Cruz de La Palma, Santa Cruz de La Palma*, Cabildo Insular de La Palma, II: 217-241.

Díez Itza, Eliseo (1992), «*Ha*, *hay*, *hace* temporales en el Siglo de Oro», en Manuel Ariza Viguera *et al.*, (eds.), *Actas del II Congreso Internacional de Historia de la Lengua Española*, Madrid, Pabellón de España, Vol. 1: 373-380.

Fernández-Ordóñez, Inés (dir.) (2005- ), *Corpus Oral y Sonoro del Español Rural*. Disponible en http://www.corpusrural.es/.

Fernández-Ordóñez, Inés. (2011), «Nuevos horizontes en el estudio de la variación gramatical del español: el Corpus Oral y Sonoro del Español Rural», en Germán Colón i Domènech y L. Gimeno Betí (eds.), *Noves tendències en la dialectología contemporània*, Castellón de la Plana, Universitat Jaume I: 173-203.

Fernández Ordóñez, Inés (2016), «Dialectos del español peninsular», en Javier Gutiérrez-Rexach (ed.), *Enciclopedia de Lingüística Hispánica*, Londres/Nueva York, Routledge: Vol. 2, 387-404. DOI: 10.4324/9781315713441-108.

Frago Gracia, Juan Antonio (2003), «Origen peninsular e influjos americanos del español de Canarias», en Carmen Díaz Alayón, Marcial Morera y Gonzalo Ortega (eds.), *Estudios sobre el español de Canarias. Actas del I Congreso Internacional sobre el español de Canarias*, Islas Canarias, Academia Canaria de La Lengua: 63-84.

García de Diego, Vicente (1946), *Manual de dialectología española*, Madrid, Instituto de Cultura Hispánica.

Gauchat, Louis (1905), *L'unité phonétique dans le patois d'une commune*, Halle, Max Niemeyer.

[INE]: Instituto Nacional de Estadística. Disponible en: https://www.ine.es/up/347MEYje. [Fecha de consulta: 31 de mayo de 2024].

Leal Cruz, Pedro N. (2003), *El español tradicional de La Palm*a, La Laguna, Gobierno de Canarias, Cabildo de La Palma/CajaCanarias/Centro de la Cultura Popular Canaria.

López Izquierdo, Marta (2014), «Sobre la distinción innovador / conservador y los modelos secuenciales en la lingüística histórica», *RILCE*, 30 (3): 776-806. DOI: 10.15581/008.30.386.

López Serena, Araceli (2011), «El andaluz y el español de América en la distancia comunicativa. ¿Hacia una norma panhispánica?», *Itinerarios: Revista de Estudios Lingüísticos, Literarios, Históricos y Antropológicos*, 14: 47-73.

López Serena, Araceli y Elena Méndez García de Paredes (2019), «¿Puede hablarse y desde cuándo de una norma para Andalucía occidental?», en Viorica Codita (ed.), Eugenio Bustos Gisbert, Juan Pedro Sánchez Méndez (coords.), *La configuración histórica de las normas del castellano*, Valencia, Tirant Humanidades: 79-108.

Lorenzo Ramos, Antonio (2003), «El uso de los pronombres en el español de Canarias. Analogías y diferencias con el de otras variedades del español», en Carmen Díaz Alayón, Marcial Morera y Gonzalo Ortega (eds.), *Estudios sobre el español de Canarias. Actas del I Congreso Internacional sobre el español de Canarias. Gran Canaria*, Islas Canarias, Academia Canaria de La Lengua: 129-151.

Medina López, Javier (1992-1993), «Estandarización lingüística en las hablas canarias», *Universitas Tarraconensis: Revista de Filologia*, 14: 175-188.

Medina López, Javier (2013), «La formación lingüística de Canarias: sustratos, contactos e historia: un balance de cinco siglos», *Zeitschrift für romanische Philologie*, 129 (2): 413-445. DOI: 10.1515/zrp-2013-0039.

Medina López, Javier (2023): «Percepciones y actitudes ante la lengua en las redes sociales. Un ejemplo a propósito de la consulta @ RAEINFORMA: "El que más me gusta" vs. "El más que me gusta"», en Alberto Hernando García-Cervigón (ed. y coord.), *Ciencia del lenguaje y discurso*, Madrid, Visor: 203-226.

Medina López, Javier y Dolores Corbella (eds.) (1996), *El español de Canarias hoy: análisis y perspectivas*, Frankfurt am Main/Madrid, Vervuert/Iberoamericana. DOI: 10.31819/9783865278340-002.

Menéndez Pidal, Ramón (1941), *Manual de gramática histórica española*, Madrid, Espasa Calpe.

Morera, Marcial (1999), «Origen y evolución del adverbio temporal español *todavía*», *Revista de Filología de la Universidad de La Laguna*, 17: 511-518.

Morera, Marcial (2007): «Unidad y variedad del español de Canarias», *Revista de Filología*, 25: 443-455.

[NGLE 2009]: Real Academia Española y Asociación de Academias de la Lengua Española (2009), *Nueva gramática de la lengua española*, Madrid, Espasa.

[NGLE 2011]: Real Academia Española y Asociación de Academias de la Lengua Española (2011), *Nueva gramática de la lengua española: fonética y fonología*, Madrid, Espasa.

Octavio de Toledo y Huerta, Álvaro (2016), *Los relacionantes locativos en la historia del español*, Berlín/Boston, De Gruyter Mouton. DOI: 10.1515/9783110458510.

Ortega Ojeda, Gonzalo (1981), «El español hablado en Canarias: visión sociolingüística», *Revista de Filología de la Universidad de La Laguna*, 0: 111-116.

Ortega Ojeda, Gonzalo (1987-1988), «Las formas *cantemos* y *cántemos* en Canarias: ¿algo más que un simple vulgarismo?», *Revista de Filología de la Universidad de La Laguna*, 6-7: 347-356.

Ortega Ojeda, Gonzalo y Narés García Rivero (2020), «Medios de comunicación y normalización lingüística en Canarias», *ACL. Revista de la Academia Canaria de la Lengua*, 1.

Padrón Castilla, Elena (en preparación), «Formas de tratamiento en el corpus Ruricán: datos de La Palma».

Parodi, Claudia y Otto Santa Ana (1997), «Tipología de comunidades de habla: del español rural al estándar», *Nueva Revista de Filología Hispánica*, 45 (2): 305–320. DOI: 10.24201/nrfh.v45i2.1999.

Pato, Enrique (2015), «*Estábanos* por *estábamos*, o la desgramaticalización de un vernáculo», *Hápax*, 8: 113-132.

Pérez Vidal, José (1946), «Los estudios lingüísticos y La Palma», *Diario de avisos* (Santa Cruz de La Palma, 2 de agosto).

Pérez Vidal, José (1949), «Nombres de la lluvia menuda en la isla de La Palma (Canarias)» *Revista de Dialectología y Tradiciones Populares*, 5: 177-197.

Pérez Vidal, José (1987), E*l romancero en la isla de La Palma*, Santa Cruz de La Palma, Cabildo Insular de La Palma.

Peña Rueda, Cristina (2022a), *Fenómenos de orden de palabras en el español rural de Canarias,* tesis doctoral, Université de Lausanne.

Peña Rueda, Cristina (2022b), «La secuencia «*ya* + pronombre personal sujeto + verbo»: distribución geográfica actual y trayectoria histórica», en María de los Ángeles Sidrach de Cardona López et al. (eds.), *Una lengua diversa y mudable: nuevas perspectivas en historiografía e historia de la lengua española*, Berlín, Peter Lang: 71-86.

Peña Rueda, Cristina (2024), «Desorientación normativa y variación gramatical en el español de Canarias», *Energeia*, 9: 57-90.

Régulo Pérez, Juan (1968-1969), «Notas acerca del habla de la isla de La Palma», *Revista de Historia Canaria*, 32, 157-164, 12-174.

Rodríguez Concepción, Anelio (1991): «En torno al léxico de los tabaqueros en la isla de La Palma», en César Hernández, et al. (eds.): *El español de América: actas del III Congreso Internacional de «El español de América»*, Valladolid: Consejería de Cultura y Turismo, Junta de Castilla y León, vol. II: 863-869.

Rodríguez Molina, Javier (2015), «El adverbio *así* en español medieval: variantes morfofonéticas», en José María García Martín (dir.), Teresa Bastardín Candón y Manuel Rivas Zancarrón (coords.), *Actas del IX Congreso Internacional de Historia de la Lengua Española (Cádiz, 2012)*, Madrid/Frankfurt am Main, Iberoamericana/ Vervuert: tomo I: 1049-1064. DOI: 10.31819/9783964566492-059.

Rosenblat, Angel (1946), «Notas de morfología dialectal», en *Biblioteca de Dialectología Hispanoamericana*, II, Buenos Aires, Instituto de Filología, Facultad de Filosofía y Letras de la Universidad de Buenos Aires: 103-316.

Samper Padilla, José Antonio (1996), «El estudio de la norma lingüística culta del español en Las Palmas de Gran Canaria», en Javier Medina López y Dolores Corbella Díaz (eds.),

*El español de Canarias hoy: análisis y perspectivas*, Madrid/Frankfurt am Main, Iberoamericana/Vervuert: 255-284. DOI: 10.31819/9783865278340-010.

Santana Pérez, Germán (2020), «Santa Cruz de La Palma: ¿tercer puerto del imperio?», en Manuel Poggio Capote, Víctor J. Hernández Correa y Antonio Lorenzo Tena (eds.), *Cinco mitos para cinco siglos: 525 aniversario de la fundación de Santa Cruz de La Palma*, vol. 1, La Palma, Cabildo Insular de La Palma: 21-38.

Vandekerckhove, Reinhild (2010), «Urban and rural language», En Peter Auer y Jürgen F. Schmidt (eds.), *Language and space. An international handbook of linguistic variation. Vol. 1: theory and methods*, Berlín/Nueva York, Mouton de Gruyter: 315-331.

Villena Ponsoda, Juan Andrés (2010), «Community-based investigations: from traditional dialect grammar to sociolinguistic studies», en Peter Auer y Jürgen F. Schmidt (eds.). *Language and space. An international handbook of linguistic variation. Vol. 1: theory and methods*, Berlín/Nueva York, Mouton de Gruyter: 613-631.

Zamora Vicente, Alonso (1979 [1967]), *Dialectología española*, Madrid, Gredos, 2ª ed. aumentada.

Adriana Raya Palmer(ORCID: 0009-0000-4524-8200), «Youth speech in translated fiction: a corpus-based comparison of selected pragmatic markers in Catalan and Spanish». *Biblioteca de Babel: Revista de Filología Hispánica*, vol. extra. 2 (2024), pp. 291-305.
**DOI**: 10.15366/bibliotecababel2024.extra2.012 · **ISSNe**: 2695-6349 · Recibido: 10/09/2024 · Aceptado: 06/11/2024

# Youth speech in translated fiction: a corpus-based comparison of selected pragmatic markers in Catalan and Spanish

Adriana Raya Palmer
*University College Dublin*
adriana.rayapalmer@ucd.ie

**Abstract**: The aim of this study is to explore fictional youth speech in translated novels in contrast to real youth speech in spoken conversation. Specifically, this analysis focuses on a selection of pragmatic markers as one of the many features of orality that can be found in literary dialogue. These markers function as expressive turn-management units between characters and are particularly prevalent in youth speech. The study examines the frequency and distribution of the pragmatic markers of interest in two languages, Catalan and Spanish, using a parallel corpus of translated dialogues from contemporary novels and two spoken corpora. The results show that, in line with previous descriptions of fictional orality, pragmatic markers are less common in fictional youth speech than they are in real conversation. However, there are some exceptions that highlight the characteristics of translated language and the literary traditions of Catalan and Spanish.

**Keywords**: translation, fictional orality, pragmatic markers, corpora, youth speech

## El lenguaje juvenil en la ficción traducida: una comparación basada en corpus de una selección de marcadores pragmáticos en catalán y español

**Resumen**: El objetivo de este estudio es explorar el habla juvenil ficticia en novelas traducidas, en comparación con el habla juvenil real en la conversación oral. En concreto, el análisis se centra en una selección de marcadores pragmáticos, uno de los muchos rasgos de la oralidad presentes en los diálogos de la ficción. Estos marcadores, que funcionan como elementos expresivos y de manejo de los turnos de habla, son particularmente frecuentes en el habla juvenil. El presente estudio analiza la frecuencia y distribución de los marcadores pragmáticos en

dos lenguas, el catalán y el español, mediante un corpus paralelo de diálogos traducidos de novelas contemporáneas y dos corpus orales. Los resultados indican que, acorde con las descripciones previas de la oralidad ficticia, los marcadores pragmáticos son menos frecuentes en el habla juvenil ficticia que en las conversaciones reales. Sin embargo, ciertas excepciones destacan las particularidades del lenguaje traducido y las tradiciones literarias del catalán y el español.

**Palabras clave**: traducción, oralidad ficticia, marcadores pragmáticos, corpus, habla juvenilIntroduction

## 1. Introduction

Pragmatic markers (PMs) that are typical of informal speech, such as *well*, *like*, *so*, *yeah*, are commonly used in the creation of fictional dialogue in works of fiction to evoke orality. In everyday conversation, PMs are crucial to guide discourse, negotiate turns, or to signal the stance of the speaker, given that spoken interaction is commonly unplanned and spontaneous. In contrast, fictional texts are carefully planned, yet still authors introduce PMs in their work, especially in the direct speech of characters to make dialogue more life-like.

While this is a stylistic technique used commonly for all types of characters, it is especially productive when representing adolescence. In many societies, teenagers are openly stigmatized by older generations by how they speak, an attitude that is present in popular media and art in the form of stereotypes or gags. An example is young Millat in Zadie Smith's *White Teeth* (2001), who overuses the tag *yeah?*, as can be seen in the example below. Thus, another function of PMs emerges: they are used for discourse and stance purposes, to evoke speech in fiction, but they can also be used to recreate effective stereotypes of speakers, especially teenagers.

(1) I just say, **yeah**? One for Bradford, **yeah**? You got some problem, **yeah**? Speaka da English? This is King's Cross, **yeah**? One for Bradford, innit? [YouLiL_EN, WT, 74]

Youth speech can be considered a social variety of a language, strongly linked to diatopic, or geographical, varieties. As such, its use in a work of fiction is effective in indexing social and emotional cues for the intended reader in the context where the text is published. Consequently, and as has been reported extensively in the literature, relaying the particularities of youth speech in another language

proves a considerable challenge for the translator, who will have to consider an array of constraints and priorities linked to the target language and the target context (see Van Coillie 2012). In this paper, the focus is on Catalan and Spanish as target languages. These languages are comparable as they are formally, geographically, and socially close. However, there is an unequal influence of one over the other, especially as regards youth language: it has been shown that Catalan youth speech makes use of Spanish loanwords or code-switching into Spanish (Pujolar 1997), while Spanish youth speech, in general, is not influenced by Catalan in the same way.

This paper aims to explore the degree of orality in translated fictional dialogue, with a focus on PMs in youth speech. The method consists of carrying out a quantitative comparison of a selection of PMs in a parallel corpus of translated dialogues in Catalan and Spanish, in comparison to a corpus of real youth speech in Catalan and another in Spanish.

## 2. Pragmatic markers in fiction and their translation

The term "pragmatic marker" is an umbrella term that covers all those words with little propositional meaning that exist outside of the syntactic structure, that guide discourse, be it in the capacity of organizing or of expressing the speaker's stance on what is being said, and that are usually short (Aijmer & Simon-Vandenbergen 2011). Also known as discourse markers, inserts, or small words, their classification is still a matter of debate among linguists. To contextualize the analysis in this paper, let us examine the role of PMs in youth speech and its representation in fiction, and how they have been studied in the context of translation.

### *2.1. PMs in youth speech*

The study of youth speech has been of interest to linguists in the past decades chiefly due to its innovative nature and the social situations that make it arise (e.g., Eckert 1989). Teenagers interact with their peers in structured groups, where they build their identity in relation to each other as they leave childhood. This is the ideal context for new linguistic forms to emerge, although few spread to the speech of other generations.

In their extensive account of the speech of British teenagers, Stenström et al. (2002) establish a set of features that define youth language and that they name "Slanguage" or "slangy language". It is composed by slang words, dirty words, vague words, among others,

and, lastly, "small words", i.e., PMs. The authors focus especially on invariant tags (*eh*, *okay*, *yeah*, *right*, *innit*), due to their high frequency of use in teenage conversation, a feature that they even suggest might be a "universal of teenage talk" (Stenström *et al.* 2002: 166). Another in-depth account of youth speech is carried out by Tagliamonte (2016), who delves into "sentence starters" (*like*, *well*, *so*) and "sentence enders" (*whatever*, *you know*).

Research on youth language in Spanish and Catalan often goes hand in hand with the study of informal or colloquial language and it has focused extensively on lexical aspects, such as word formation or swearing (see Regueiro Rodríguez 2023). In Spanish, there are studies on PMs in youth speech, such as the description of vocatives (Stenström 2020), or innovative features like *en plan* ('like') (De Smet & Enghels 2020), to name a few. In Catalan, however, the study of colloquial and youth language takes the form of comprehensive descriptive studies of communicative situations, where PMs are not usually the main concern (e.g., Bernal & Sinner 2009).

## *2.2. Translating PMs*

In fictional orality, PMs convey metalinguistic and turn-management strategies, making them fundamental items in the creation of dialogue. Bublitz (2017) defines orality in fiction as "reduced orality", since the features that index speech—coordination rather than subordination, generalised vocabulary, hesitation and self-correction, among others—are purposefully placed in the dialogue for stylistic effect. Therefore, while PMs are essential, they do not appear at the same frequency as they do in real speech (see also López Serena 2007 on the features of colloquial Spanish in literary fiction). Bednarek (2010) suggests that PMs are less frequent in fiction because, as non-lexical items, they are not productive towards advancing narration. In turn, this also makes them more noticeable when they are used and are thus a productive strategy for characterization.

As with most features of orality, the context-boundness and multiple functions of PMs render them challenging to translate. Practices that are observed in translated target texts at large, such as explicitation, standardization, lexicalization, or omission, are also observed in the translation of PMs in particular (González 2012). On the other hand, the most prevalent markers are usually translated as markers in the target language, but not necessarily in a one-to-one fashion (see González Villar & Arias Badia 2017). Other studies have found that the more stable the function of a marker, the more homogeneous its translation (Brumme & Schmid 2021).

The target language and target context of a translation conditions the translation choices or solutions that are taken by the translator. Translating spoken, colloquial, or youth language leads to a negotiation between what is expected in the tradition of literary translation in a given context and what is closest to communication between real speakers. In the case of Catalan, this distance between fiction and reality is heightened by the presence of Spanish and English code-switching and loanwords in youth speech. Regardless of the presence of foreign languages, translators, writers, and scholars advocate that colloquial Catalan in fiction can be more transgressive with the strategies that spoken Catalan lends itself to, without having to resort to Spanish or English (see Ainaud *et al.* 2020; Cabal Guarro 2024; Gurt 2024).

## 3. Methods

This study relies on the comparison of a selection of PMs across several corpora. Youth speech in fiction is observed in the YouLiL corpus1, a parallel corpus of dialogues translated from English into Spanish and Catalan that have been extracted from five contemporary novels: *White Teeth* (Smith 2001), *Middlesex* (Eugenides 2002), *I am the Messenger* (Zusak 2002), *Paper Towns* (Green 2008), and *The Casual Vacancy* (Rowling 2012)[2]. The novels depict young characters in realistic, urban settings and across Western English-speaking regions. The token count in the Catalan subcorpus is 22 307 tokens, while in the Spanish subcorpus it is 20 920 tokens.

The dialogue is compared to reference corpora of real spontaneous conversation between young speakers. For Catalan, the COC corpus is used (initials in Catalan for Spoken Corpus of Colloquial Conversation), a corpus compiled by researchers at Universitat de Barcelona (accessible in Payrató & Alturo 2002). COC consists of conversations in informal settings, namely gatherings between friends or family, that were recorded between 1993 and 1997. For this study, only turns by young people aged 14-29 were considered: a total of five conversations and 18 123 tokens.

The Spanish subcorpus of dialogues is compared to sections in Val.Es.Co 3.0 (Pons Bordería 2024), a corpus of colloquial conversation in Spanish created by Universitat de València. This is a larger corpus, with conversations that span from 1989 to 2022. For this analysis, however, only seven conversations from 1994 and 1996 were chosen, given that they explicitly portrayed young speakers (mostly university students,

[1] See Raya Palmer (2023: 3.2 and 3.3) for a detailed description of the YouLiL corpus.
[2] These novels will henceforth be abbreviated as WT, MS, IATM, PT, and TCV respectively.

in the age group 18-34), and to ensure comparability with COC. The total token count is 16 673.

The transcriptions of the selected conversations were extracted from COC and Val.Es.Co 3.0 and were introduced into #Lancsbox (Brezina et al. 2020), a language data analysis software that enables the quantification of words and n-grams (i.e., co-occurrences of words). The PMs were chosen, initially, according to the most relevant markers for the creation of youth speech found in the Catalan subcorpus of translated dialogues in Raya Palmer (2023: 4.2.2.3). The markers were contrasted in fictional and real speech, and in Spanish and Catalan, with the aim of observing their frequency and functions across corpora. In sum, this is a form-to-function methodology, which, as Aijmer points out, "has the advantage that linguistic elements (…) can be studied with great precision" (2020: 30).

Comparing the frequencies of lexical items across corpora requires the normalization of frequencies (in this case, per 10k tokens). It is also crucial to consider the distribution of items within the sections of the different corpora; to do so, the percentage coefficient of variation (CV %) is observed, with > 50 % as the threshold for uneven distribution (Brezina 2018: 51). One of the benefits of dealing with small corpora is that the automatic analysis can be complemented with a manual analysis, which is especially valuable to ensure the correct selection of PMs, as they often share forms with lexical words (e.g., *mira* 'look', can be used as a marker or as a verb).

## 4. Analysis

As an introduction to the results, Figure 1 below presents the relative frequencies per 10k tokens across fictional dialogue (FD) and spoken conversation (SC) and in the Spanish and Catalan corpora. The focus is on *doncs*, *pues* and *bueno* ('well'); *és que* and *es que* ('it is that'); and the question tags *oi?*, *no?*, *val?*, *eh?* and *¿verdad?*, *¿no?*, *¿vale?*, and *¿eh?*. At first glance, it appears that most items are more frequent in spoken conversation than in fictional dialogue, as is expected. However, there are some units—*doncs*, *oi?*, and *val?*, in Catalan, and *¿verdad?*, and *¿vale?* in Spanish—that are more frequent in fictional dialogue.

The subsections that follow provide a more detailed description of each marker, alongside the values of their frequency and distribution across the corpora.

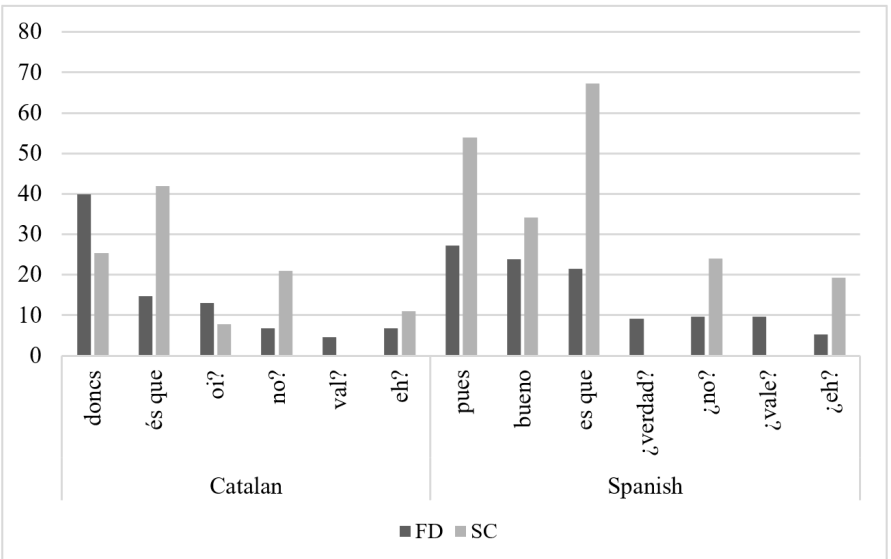



*Figure 1. Relative frequencies per 10k tokens of PMs in FD and SC in Catalan and Spanish*

## *4.1.* Doncs, pues *and* bueno

In the Catalan subcorpus of fictional dialogues, *doncs* is one of the most frequent markers, as are *pues* and *bueno* in the Spanish parallel subcorpus. These are often translated from *well* in the source texts (ST) in English, which is used widely as "a discourse boundary marker, a response utterance initiator, a generalized starter, and an attention-getter" (Tagliamonte 2016: 111). Similarly, *doncs* can be used as a connector that signals continuity with what has been said previously (IEC 2023: 26.4.1) and *pues* introduces new information (Briz et al. 2008). *Bueno* can be used both to signal continuity with what has been said previously and to introduce complete or partial agreement (Briz et al. 2008). In the example below, the PMs introduce an explanation with a slight hesitation, both in the ST and the target texts (TT).

| | | |
|---|---|---|
| (2) | ST | What'd you do to her?<br>Well, when she told me about Jase, I sort of shot the messenger. |
| | TT CA | I què li has fet?<br>Doncs quan m'ha explicat això d'en Jase, es podria dir que he matat el missatger. |
| | TT SP | ¿Qué le hiciste?<br>Bueno, cuando me contó lo de Jase, de alguna manera maté al mensajero. [YouLiL, PT, 225-26] |

These markers are present across subsections in the corpora, as shown by the CV% values in Table 1. While *pues* and *bueno* follow the expected pattern, whereby they are more frequent in spoken conversation than in fictional dialogue, *doncs* does not. This is most probably due to the fact that in Catalan spoken conversation the functions of introducing turns and partial agreement are also carried out by Spanish or non-standard *pues* (and variant *pos*) and *bueno*, with absolute token counts of 28 (rel. freq. 15.5) and 75 (rel. freq. 41.4) respectively. Though common in everyday speech, they are not conventionally accepted in literary dialogue. Conversely, the preferred option in literary writing bé, equivalent in many contexts to Spanish *bueno*, is used in Catalan fictional dialogue (albeit rarely, with only six occurrences), but does not appear once in spoken conversation.

| Language | Marker | Corpus | Abs. freq. | Rel. freq. | CV % |
|---|---|---|---|---|---|
| Catalan | *doncs* | FD | 78 | 40.0 | 17.0 |
| | | SC | 46 | 25.4 | 11.0 |
| Spanish | *pues* | FD | 57 | 27.3 | 18.0 |
| | | SC | 90 | 54.0 | 26.9 |
| Spanish | *bueno* | FD | 50 | 23.9 | 27.6 |
| | | SC | 57 | 34.2 | 21.6 |

*Table 1 Frequency and distribution values of* doncs, pues, *and* bueno.

Aside from frequency, patterns in the co-occurrence of markers (see Cuenca & Marín 2009) consistently present differences between fictional dialogue and spoken conversation in both languages. In Catalan spoken conversation, *doncs* co-occurs with *ah* (*ah doncs*, 'oh, well'), *bueno* (*bueno doncs*, 'well, well'), and also non-standard *vale* (*vale doncs*, 'okay, well'). In fictional dialogue it co-occurs with *així* (*així doncs*, 'so, well'), *molt bé* (*molt bé doncs*, 'very good, well'), *mira* (*doncs mira*, 'well, look'), and *sí* (*doncs sí*, 'well yes').

In Spanish, *bueno* and *pues* often co-occur together, both in fictional dialogue and in spoken conversation. In spoken conversation, *pues* also co-occurs with *sí* (*sí pues*, 'yes, well') and *nada* (*pues nada*, 'well, nothing'), while *bueno* co-occurs most commonly with *y* (*bueno y*, 'well, and') and *es que* (*bueno es que*, 'well, the thing is'). In Spanish fictional dialogue, both *pues* and *bueno* co-occur with *sí* ('yes') and *vale* ('okay'). In sum, although these markers are present across registers, patterns of co-occurrence differ considerably. A unifying feature is that the markers are significantly more frequent as single units across all corpora than in pairs of co-occurrence.

## *4.2.* És que *and* es que

*És que/es que* ('it is that' or 'the thing is') is used to introduce a justification of what has been said previously, an excuse, an apology or a mitigated objection. It can also introduce a reaction to an implied rejection or a polite excuse (Briz *et al.* 2008) and is sometimes used for emphasis (Marín & Cuenca 2012). This analysis considers the grammaticalized marker alone, but also *és que/es que* pseudo-cleft constructions with a generic noun or clause, such as *el cas és que* or *el caso es que* ('the case is that') (see Marín & Cuenca 2012).

In the corpora of fictional dialogue, characters mostly use this marker to clarify what they are saying, thus avoiding conflict, and as a result of hesitation, as in example 3. *És que/es que* appears in the corpus as a translation from 'I mean', 'just', and 'the thing is', but also as an addition where there is no marker in the ST, as in example 4. In contrast to the markers in the previous section, pairs in the TTs where this marker is used in both Catalan and Spanish are scarce.

(3) ST **I just** thought… like when she says, here: Then will I swear, beauty herself is black… (...)
TT CA **És que** em pensava… com això que diu aquí: Llavors juraré que la bellesa és negra… (...)
TT SP ∅ Yo pensé… Como aquí dice: «Entonces juraré que es negra la hermosura misma…». (...) [YouLiL, WT, 162]

(4) ST Okay. Sometimes I think I have bad breath. (...)
TT CA D'acord. ∅ A vegades em penso que em put l'alè. (...)
TT SP Vale. Es que a veces me parece que me huele el aliento. (...) [YouLiL, MS, 217]

The token counts reveal that this marker is more frequent in Spanish than it is in Catalan, both in fictional dialogue and in spoken conversation; in parallel, it is used much more in spoken conversation than in fictional dialogue in both languages. As Table 2 shows, the distribution is even across corpora.

| Language | Marker | Corpus | Abs. freq. | Rel. freq. | CV % |
|---|---|---|---|---|---|
| Catalan | *és que* | FD | 33 | 14.8 | 9.0 |
| | | SC | 76 | 41.9 | 18.5 |
| Spanish | *es que* | FD | 45 | 21.5 | 14.0 |
| | | SC | 112 | 67.2 | 13.5 |

*Table 2. Frequency and distribution values of* és que *and* es que.

Not only do the overall frequencies reveal differing tendencies in fictional dialogue and spoken conversation, but the type of construction varies, too: in fictional dialogue, pseudo-cleft structures with general nouns make up 55 % of cases in Catalan and 47 % in Spanish; in spoken conversation, they make up only 11 % of cases in Catalan and 13% in Spanish. In other words, fictional dialogue in the corpora of this study prefers the marker *és que/es que* in pseudo-cleft structures, instead of the grammaticalized PM. The most common pseudo-cleft constructions in fictional dialogue, aside from the pair in the example below *el cas és que*/l*o que pasa es que* ('the thing is that') , are *la verdad es que* ('the truth is') and l*o que quiero decir es que* ('what I want to say is that') in Spanish, and its equivalent *el que vull dir és que* in Catalan. Interestingly, while this final construction does appear in the Catalan corpus of spoken conversation, its equivalent does not appear at all in the Spanish subcorpus, suggesting that it might be a construction reserved for fictional dialogue.

| | | |
|---|---|---|
| (5) | ST | It's just… I wish it was easier, for me, you know? |
| | TT CA | **El cas és que**… Voldria que fos més fàcil, que em resultés més fàcil, ¿m'entens? |
| | TT SP | **Lo que pasa es que** me gustaría que fuera más fácil para mí, ¿sabes? [YouLiL, IATM, 283] |

### *4.3.* Oi?, ¿no?, *and other question tags*

The final set of markers to consider are question tags. In the TTs in YouLiL, they arise when the STs use canonical and non-canonical question tags, such as *isn't he?*, *right?*, *innit?*, etc. The functions of question tags usually involve seeking confirmation from the listener about what the speaker is saying, as in the example below, inviting further information, or as a contact-check to make sure the listener is following the conversation (see Andersen 2001; Cuenca & Castellà 1995).

| | | |
|---|---|---|
| (6) | ST | Cubby upset, **was he?** |
| | TT CA | En Cubby està molt afectat, **oi?** |
| | TT SP | Cuby está muy afectado, **¿no?** [YouLiL, TCV, 109] |

In contrast to the previous sets of markers, these are far less frequent overall, and are less evenly distributed, as can be appreciated in Table 3. They are worth contemplating, however, as notable contributors to the portrayal of youth stereotypes; as noted in Raya Palmer (2023) and

in the example in the start of the study, they are highly productive for character delineation.

| Language | Marker | Corpus | Abs. freq. | Rel. freq. | CV% |
|---|---|---|---|---|---|
| Catalan | oi? | FD | 29 | 13.0 | 31.5 |
| | | SC | 14 | 7.7 | 58.4* |
| | no? | FD | 15 | 6.7 | 55.7* |
| | | SC | 38 | 21.0 | 39.8 |
| | val? | FD | 10 | 4.5 | 100.0* |
| | | SC | 0 | 0.0 | 0.0 |
| | eh? | FD | 15 | 6.7 | 37.5 |
| | | SC | 10 | 11.0 | 41.0 |
| Spanish | ¿verdad? | FD | 19 | 9.1 | 30.0 |
| | | SC | 3 | 0.0 | 100.0* |
| | ¿no? | FD | 20 | 9.6 | 50.7* |
| | | SC | 40 | 24.0 | 18.6 |
| | ¿vale? | FD | 20 | 9.6 | 46.8 |
| | | SC | 1 | 0.0 | 100.0* |
| | ¿eh? | FD | 11 | 5.3 | 33.3 |
| | | SC | 32 | 19.2 | 25.8 |

*Table 3. Frequency and distribution values of response elicitors. Note: The asterisks (*) mark the CV % values that present an uneven distribution across subsections.*

The preferred marker in Catalan fictional dialogue is *oi?*, which, like *doncs* above, is more common in fictional dialogue than in spoken conversation. It is also the most evenly distributed marker in the corpus, followed by *eh?*, *no?* and *val?*, which are more restricted, with *val?* (non-standard 'okay?') appearing in only one novel. This last marker is an exception, as the translator chooses a non-standard form to add a layer of colloquiality to the dialogue; however, it is not supported by the subcorpus of real spoken conversation in Catalan[3].

In the Spanish subcorpus of dialogues, the tags present a wider variety of use, with *¿verdad?*, *¿no?* and *¿vale?* being used at similar relative frequencies across novels. *¿No?*, on the other hand, presents an uneven distribution, like in Catalan. As is the case with Catalan, *¿no?* and *¿eh?* are more frequent in spoken conversation than in fictional dialogue. In fact, *¿verdad?* and *¿vale?* present extremely low frequencies in spoken conversation, with only three and one occurrence respectively.

## 5. Discussion and final thoughts

This paper set out to establish potential differences between the frequency of PMs in fictional and real youth speech, in Catalan and

[3] On the other hand, there are instances of code-switching to Spanish *vale*, although there are only three occurrences.

Spanish. Given that this is a broad topic, the analysis focused on a selection of PMs: the Catalan sentence starters *doncs* and *és que*; the Spanish sentence starters *pues*, *bueno*, and *es que*; the Catalan question tags *oi?*, *no?*, *val?*, and *eh?*; and the Spanish question tags *¿verdad?*, *¿no?*, *¿vale?*, *¿eh?*. These PMs were chosen due to their frequency in the corpora and their significance in portraying youth speech. The contrast between the corpora reveals that, in general, these PMs are more prevalent in the corpora of real youth speech than in the fictional corpora, thus agreeing with the insights of Bublitz (2017) and Bednarek (2010) on the reduction of PMs in fiction.

In contrast, there are two cases in Catalan that present an inverse pattern, that of *doncs* and *oi?*; markers that, though belonging to the structures of spoken interaction, appear to be more typical of fictional dialogue. Comparing dialogues to spoken conversation among different age groups could assert whether this is specific to young speakers or to Catalan speakers in general. If the latter is true, a preference for certain markers of orality in fictional dialogue might imply the existence of an established style in Catalan fictional dialogue, which can be differentiated from both spoken Catalan and from literary Catalan.

The repertoire of question tags in the fiction corpora suggests varied and unfixed translation solutions. There is a deliberate intent of reproducing the wide range of tags in the STs, despite a clear preference in spoken conversation in both languages for *no* and *eh*. Another result that stems from the translation process is the preference for explicitation in the *és que*/*es que* pseudo-cleft sentences, instead of the more informal grammaticalized marker. These findings lead one to consider other consequences of the translation process: for instance, there seems to be a higher frequency of PMs in Spanish spoken corpora than in Catalan corpora, yet the translated dialogues do not follow these distributions to the same extent. Presumably, this is due to the influence of the STs in English. Further research could study the distribution of these markers in STs in Catalan and Spanish, to assert whether they are closer to patterns in real conversation than the ones in this paper.

This study has contributed to the definition of fictional dialogue, especially in relation to youth speech. However, it is important to note that these results are particular to the data analyzed. While the data in the corpus of fictional dialogues is supra-local and follows the conventions of translated literature across Catalan and Spanish-speaking regions, the COC and Val.Es.Co 3.0 corpora are specific to the regions where they were created. In this sense, the patterns in this study would benefit from being contrasted to larger corpora with conversations by speakers of different language varieties of Catalan and Spanish.

## 6. Funding

This article is part of the Youth Engagement in European Language Preservation Project, which has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 802695).


## References

Aijmer, Karin (2020), "Contrastive Pragmatics and Corpora", *Contrastive Pragmatics* 1 (1): 28-57. DOI: 10.1163/26660393-12340004.

Aijmer, Karin & Anne-Marie Simon-Vandenbergen (2011), "Pragmatic markers", in Jan Zienkowski, Jan-Ola Östman & Jef Versucheren (eds.), *Discursive Pragmatics*, Amsterdam, John Benjamins Publishing Company: 223–247.

Ainaud, Jordi, Anna Espunya & Dídac Pujol (2020), *Manual de traducció anglès-català, Barcelona*, Universitat Pompeu Fabra.

Andersen, Gisle (2001), *Pragmatic Markers and Sociolinguistic Variation: A Relevance-Theoretic Approach to the Language of Adolescents*, Amsterdam & New York, John Benjamins Publishing Company.

Bednarek, Monika (2010), *The Language of Fictional Television: Drama and Identity*, London and New York, Continuum.

Bernal, Elisenda & Carsten Sinner (2009), "Al seu rotllo: Aproximació al llenguatge juvenil català", *Zeitschrift für Katalanistik*, 22: 7-36. DOI: 10.46586/ZfK.2009.7-36.

Brezina, Vaclav (2018), *Statistics in Corpus Linguistics*, Cambridge, Cambridge University Press.

Brezina, Vaclav, Pierre Weill-Tessier & Tony McEnery (2020), #LancsBox 5.x and 6.x. [software package].

Briz, Antonio, Salvador Pons & José Portolés (coords.) (2008), *Diccionario de partículas discursivas del español*, Available at: www.dpde.es. [Accessed: August 3rd, 2024].

Brumme, Jenny & Beatrice Schmid, (2021), "Convergències i divergències en l'ús de les partícules discursives", *Actes Del XVIIIè Col·loqui de l'AILLC*, 96–108. DOI: 10.2436/15.8090.02.6.

Bublitz, Wolfgang, (2017), "Oral features in fiction", in Mariam Locher & Andreas. H. Jucker (eds.), *Pragmatics of Fiction*, Berlin and Boston, De Gruyter: 235-263.

Cabal Guarro, Miquel (2024), "*Tenim de traduir-ho, això*: la (re)creació del col·loquial en la traducció literària", in Elisenda Bernal (ed.), Col·loquial(s): *Estudis de l'ús del català actual*, Leipzig, Leipziger Universitätsverlag: 101-118.

Cuenca, Maria Josep & Josep Maria Castellà (1995), "Una caracterització cognitiva de les preguntes confirmatòries (*question tags*)", *Caplletra*, 18(Spring): 65–84.

Cuenca, Maria Josep & Maria Josep Marín (2009), "Co-occurrence of discourse markers in Catalan and Spanish oral narrative", J*ournal of pragmatics*, 41 (5): 899-914. DOI: 10.1016/j.pragma.2008.08.010.

De Smet, Emma & Renata Enghels (2020), "Los datos en Twitter como fuente del discurso oral coloquial: Estudio de caso del marcador discursivo *en plan*", *Oralia*, 23 (2): 199-218. DOI: 10.25115/oralia.v23i2.6379.

Eckert, Penelope (1989), *Jocks and Burnouts: Social categories and identity in the high school*, New York, Teachers College Press.

González, Montserrat (2012), "Pragmatic markers in translation", in Jenny Brumme & Anna Espunya (eds.), *The Translation of Fictive Dialogue*, Amsterdam and New York, Rodopi: 218-232.

González Villar, Alejandro & Blanca Arias Badia, (2017), "Marcadors conversacionals en la traducció literària alemany-català: *also i na a Jeder stirbt für sich allein*", *Zeitschrift Für Katalanistik*, 30, 245–267. DOI: https://doi.org/10.46586/ZfK.2017.245-267.

Gurt, Carlota (2024), "El català col·loquial des de la trinxera de l'escriptor: la teoria de l'equalitzador", in Elisenda Bernal (ed.), Col·loquial(s): *Estudis de l'ús del català actual*, Leipzig, Leipziger Universitätsverlag: 87-100.

Institut d'Estudis Catalans (IEC) (2023), *Gramàtica essencial de la llengua catalana (GEIEC)*, [online ed.], Accessible at: https://geiec.iec.cat/.

López Serena (2007), *Oralidad y escrituralidad en la recreación literaria del español coloquial*. Madrid, Gredos.

Marín, Maria Josep & Maria Josep Cuenca (2012), "De l'atribució a la modalitat: Construccions amb és que en català oral", *Caplletra* (Spring): 65–94. DOI: https://doi.org/10.7203/caplletra.52.4699.

Raya Palmer, Adriana (2023), *The Representation of Youth Language in Fictional Dialogue in English and its Translation into Catalan*, tesis doctoral, Universitat Pompeu Fabra.

Payrató, Lluís & Núria Alturo (2002), *Corpus oral de conversa col·loquial. Materials de treball,* Barcelona, Publicacions de la Universitat de Barcelona.

Pons Bordería, Salvador (2024), *Corpus Val.Es.Co 3.0.*, Available at: http://www.valesco.es. [Accessed: July 21st, 2024].

Pujolar, Joan (1997), *De què vas, tio?*, Barcelona, Empúries.

Regueiro Rodríguez, María Luisa (2023), *Diccionario del léxico juvenil en España: del lenguaje juvenil al español coloquial*. Navarra, EUNSA. Ediciones Universidad de Navarra, SA.

Stenström, Anna-Brita (2020), "English- and Spanish-speaking teenagers' use of rude vocatives", in Nico Nassenstein & Anne Storch (eds.), *Swearing and Cursing: Contexts and practices in a critical linguistic perspective*, Berlin and Boston, De Gruyter: 281-302.

Stenström, Anna-Brita, Gisle Andersen, & Ingrid Kristine Hasund (2002), *Trends in Teenage Talk*, Amsterdam and Philadelphia, John Benjamins Publishing Company.

Tagliamonte, Sali A. (2016), *Teen Talk: The language of adolescents*, Cambridge, Cambridge University Press.

Van Coillie, Jan (2012), "*Cool*, *geil*, *gaaf*, *chouette* or *super*. The challenges of translating teenage speech", in Martin Fischer & Maria Wirf Naro (eds.), *Translating Fictional Dialogue for Children and Young People*, Berlin, Frank & Timme: 217-232.


UÁM