LingCor2024: 3rd International Workshop on Spoken Corpus Linguistics

Institut für Romanistik,
Universität Wien

25.07. - 26.07.2024

Llibre de resums
Libro de resúmenes
Book of abstracts

Samia Aderdouch Derdouch
(Universitat de Barcelona)

El *Corpus Gradia de Mujeres Marroquíes de Cataluña* (COMMAR). Una aplicación para el estudio de las perífrasis verbales en situaciones de contacto de lenguas

El objetivo de esta comunicación consiste en presentar un corpus oral, el *Corpus Gradia de Mujeres Marroquíes de Cataluña* (COMMAR), y mostrar evidencias empíricas de su explotación para el estudio del contacto de lenguas. En concreto, se ha centrado el análisis en el estudio del contacto entre el árabe marroquí y el español en el empleo de las perífrasis verbales del español por parte de mujeres marroquíes de Barcelona.

El COMMAR incluye una muestra de la población femenina marroquí formada por un total de 100 informantes, repartidas en dos principales grupos según su lugar de escolarización. Por un lado, participantes escolarizadas en Marruecos y, por otro, informantes escolarizadas en España. Los métodos empleados para la recogida de datos han sido las conversaciones libres, de parejas de informantes, y las encuestas. De este modo, se han realizado un total de 50 grabaciones de audio, que representan 25 horas de grabación, y 100 encuestas individuales que han proporcionado datos sociolingüísticos sobre las informantes. También se ha manejado un corpus de control formado por informantes nativos de español, del que se han seleccionado únicamente las participantes de sexo femenino. Se trata del *Corpus Oral del Español de Barcelona* (GRADESBAR), compilado por el grupo GRADIA de la Universitat de Barcelona, y que ha servido para contrastar los resultados obtenidos.

Los resultados de esta investigación se fundamentan en los principios teóricos de la sociolingüística (Labov 1966, 1972; Silva-Corvalán 1989; Moreno Fernández 1990, 2009; López Morales 1993; Hernández Campoy y Almeida 2005) y del contacto de lenguas (Weinreich 1953, Thomason y Kaufman 1988, Thomason 2001, Heine y Kuteva 2005, Matras 2009, Palacios Alcaine 2017). Por último, después de realizar un vaciado de las perífrasis verbales del corpus mediante el programa AntConc y de descartar aquellos ejemplos no perifrásticos, los datos cuantitativos y cualitativos obtenidos han mostrado diferencias significativas en el empleo de las perífrasis verbales por las informantes marroquíes. Por ejemplo, se aprecia una destacada frecuencia de empleo del futuro analítico (*voy a ir*) en contraste con el futuro sintético (*iré*), en las informantes escolarizadas en Marruecos y en las participantes marroquíes escolarizadas en España, en comparación con las informantes nativas de español, que tienden a emplear con mayor frecuencia el futuro sintético, por el contacto con

el catalán (Blas Arroyo 2004, 2007, 2008, 2012; Enrique-Arias y Méndez Guerrero 2020; Garachana 2021). Este destacado uso de la forma perifrástica (*ir a* + INF) en las participantes marroquíes puede deberse al contacto con el árabe marroquí, dado que en él también existe una perífrasis similar formada por el verbo de movimiento *ir* (*ġadi/maxi* + verbo), y con una alta frecuencia de empleo.

Palabras clave: COMMAR, sociolingüística, contacto de lenguas, perífrasis verbales, árabe marroquí.

Bibliografía

Blas Arroyo, J. L. (2004). "El español actual en las comunidades del ámbito lingüístico catalán". En Rafael Cano Aguilar (ed.). *Historia de la lengua española*. Madrid: Arco Libros, 1065-1086.

__ (2007). "El contacto de lenguas como factor de retención en procesos de variación y cambio lingüístico. Datos sobre el español en una comunidad bilingüe peninsular". En *Spanish in Context*, 4 (2): 263-291.

__ (2008). "The variable expression of future tense in Peninsular Spanish: The present (and future) of inflectional forms in the Spanish spoken in a bilingual region". En *Language Variation and Change*, 20: 85-126.

__ (2012). *Sociolingüística del español. Desarrollos y perspectivas en el estudio de la lengua española en contexto social*. Madrid: Cátedra.

Enrique-Arias, A. y Méndez Guerrero, B. (2020). "On the effects of Catalan contact in the variable expression of Spanish future tense. A contrastive study of Alcalá de Henares, Madrid and Palma, Majorca". En Luis A. Ortiz-López, Rosa E. Guzzardo Tamargo y Melvin González-Rivera (eds.). *Hispanic Contact Linguistics. Theoretical, methodological and empirical perspectives*. Amsterdam: John Benjamins Publishing Company, 315-334.

Garachana Camarero, M. (2021). "La evolución de *ir a* + INF en zonas de contacto lingüístico. El caso del español de Barcelona". En Azucena Palacios y María Sánchez Paraíso (eds.). *Dinámicas lingüísticas de las situaciones de contacto*. Berlín: De Gruyter, 321-344.

Heine, B. y Kuteva, T. (2005). *Language Contact and Grammatical Change*. Cambridge: Cambridge University Press.

Hernández Campoy, J. M. y Almeida, M. (2005). *Metodología de la investigación sociolingüística*. Albolote (Granada): Editorial Comares.

Labov, W. (1966). *The social stratification of English in New York City*. Washington D.C.: Center for Applied Linguistics.

— (1972). *Sociolinguistic Patterns*. Filadelfia: University of Pennsylvania Press. Trad. al español. *Modelos sociolingüísticos*. Madrid: Ediciones Cátedra.

López Morales, H. (1993). *Sociolingüística*. Madrid: Gredos.

Matras, Y. (2009). *Language Contact*. Cambridge, New York, Melbourne, Cape Town, Singapore, São Paulo, Delhi: Cambridge University Press.

Moreno Fernández, F. (1990). *Metodología sociolingüística*, Madrid: Gredos.

— (2009). *Principios de sociolingüística y sociología del lenguaje.* Barcelona: Ariel.

Palacios Alcaine, A. (2017). *Variación y cambio lingüístico en situaciones de contacto*. Madrid/Frankfurt: Iberoamericana/Vervuert.

Silva-Corvalán, C. (1989). *Sociolingüística. Teoría y análisis*. Madrid: Alhambra.

Thomason, S. G. (2001). *Language contact. An introduction*. Edinburgh: Edinburgh University Press.

Thomason, S. G. y Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics.* Berkeley, Los Angeles, Oxford: University of California Press.

Weinreich, U. (1953). *Languages in contact*. The Hauge: Mouton.

Jorge Agulló
(Universität Wien)

The Definiteness Effect at the crossroads: Spanish and Catalan

1. OVERVIEW AND GOALS

Existential sentences in Spanish are sensitive to what Milsark (1974) referred to as *definiteness* or *quantification* restriction or effect (henceforth DE). The DE prevents personal pronouns (1a), proper nouns (2a), and definite nouns (3a) —following Aissen's (2000) Definiteness Scale (cfr. Farkas 2000)—, but also quantificational constituents (4a) from occupying the pivot position:

(1) a. *Sp.* *Hay él en la habitación.
b. *Cat.* Hi ha ell a l' habitació.
is-LOC him in ART room
'There is him in the room'

(2) a. *Sp.* * Hay Juan en la habitación.
b. *Cat.* Hi ha en Juan a l' habitació.
is-LOC Juan in ART room
'There is Juan in the room'

(3) a. *Sp.* *Hay el niño en la habitación.
b. *Cat.* Hi ha el nen a l' habitació.
is-LOC ART boy in ART room
'There is the boy in the room'

(4) a. *Sp.* * Hay cada libro en la habitación.
b. *Cat.* Hi ha cada llibre a l' habitació.
is-LOC every book in ART room
'There is every book in the room'

The same sequences are grammatical in Catalan, as shown in examples (1b, 2b, 3b, 4b). The DE is known to be subject to cross linguistic variation: it is robust in languages like Spanish or French (Bouchard 1997), whereas Catalan or Italian (see Leonetti 2008) have weaker versions of it. Data from contact varieties, nonetheless, remains largely unexplored. The aim of this study is to cast some light onto the weakening of the DE in a linguistic contact setting between Spanish, a robust language as regards the effect, and Catalan, which has a weaker version of it.

2. METHODOLOGY AND EVIDENCE

The research underlying this proposal is based on observation and reflection upon a large dataset of existential constructions (N = 5000) gathered via the *Audible Corpus of Spoken Rural Spanish* (Fernández-Ordóñez dir.). A dialectal and hence quantitative approach is privileged here.

3. EXPECTED RESULTS AND THEORETICAL IMPLICATIONS

Hitherto unnoticed data, as in (5), will be brought to the fore in support of the hypothesis that Spanish in contact with Catalan amnesties the definiteness effect: even possessive determiners (5a) and definite articles (5b) are found in these varieties.

(5) a. Hasta cuando había mi hermano vivo.
even when there.was my brother alive
'Even when there was my brother alive.' (COSER-2712_01)

b. Menos mal que había el vecino.
less bad COMP there.is ART neighbor
'Luckily, there was the neighbor.' (COSER-2711_01)

An explanation will be developed that assumes the *borrowability* or *transferability* of syntax, along the lines of Thomason & Kaufman (1988), Campbell (1993), Bowern (2008) or Thomason (2014). I will demonstrate that the *locus* of syntactic variation at stake in (5) is the transfer of the weak version of the DE from Catalan to Spanish. New *interdialect* variants, in the sense of Trudgill (1989a, 1989b, 1992), will be argued to arise that are immune to the DE.

Keywords: existentials, definiteness effects, Spanish, contact, dialectology.

References

Aissen, J. (2000). *Differential Object Marking: Iconicity vs. Economy*. Ms.: UCSC.

Bouchard, D. (1997). L'effet existentiel. In Julie Auger & Yvan Rose (eds.), *Explorations du lexique*. Québec : CIRAL, pp. 31-45.

Bowern, C. (2008). Syntactic change and syntactic borrowing in generative grammar. In G. Ferraresi & M. Goldbach (Eds.), *Principles of syntactic reconstruction*. Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 187-216.

Campbell, L. (1993). On proposed universals of grammatical borrowing. In H. Aertsen & R. J. Jeffers (Eds.), *Historical Linguistics 1989. Papers from the 9th International Conference on Historial Linguistics. Rutgers University, 14-18 August 1989*. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Farkas, D. F. (2000). Varieties of Definites. Ms. University of California Santa Cruz, en línea: <https://people.ucsc.edu/~farkas/papers/definites.pdf> [accedido: 20/09/2023].

Leonetti, M. (2008). Definiteness effects and the role of the coda in existential constructions. In Müller, Henrik Høeg, & Alex Klinge (eds.), *Essays on nominal determination. From morphology to discourse management*. Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 131-162.

Milsark, G. L. (1974). *Existential Sentences in English*. PhD Dissertation, Massachusetts Institute of Technology.

Thomason, S. G. (2014). Contact-induced language change and typological congruence. In J. Besters-Dilger, C. Dermarkar, S. Pfänder, & A. Rabus (Eds.), *Congruence in contac-induced language change. Language families, typological resemblance, and perceived similarity*. Berlin: Mouton de Gruyter, pp. 201-218.

Thomason, S. G., & Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. California: University of California Press.

Trudgill, P. (1989a). Contact and isolation in linguistic change. In *Language change: Contributions to the study of its causes* (pp. 227-237). Berlin: Mouton de Gruyter.

Trudgill, P. (1989b). *Language contact and simplification*. Nordlyd, 15, 115-121.

Trudgill, P. (1992). Dialect typology and social structure. In E. H. Jahr (Ed.), *Language contact: Theoretical and empirical studies* (pp. 195-211). Berlin: Mouton de Gruyter.

Marta Albelda Marco
(Universitat de València)

Desafíos en la compilación de materiales naturales de la comunicación conflictiva entre íntimos

Uno de los principales desafíos para la caracterización lingüística del conflicto en conversaciones privadas es la escasez de materiales de estudio, dada su sensibilidad. El panorama actual de la lingüística de corpus es altamente satisfactorio en lo relativo a la compilación, procesamiento y almacenaje de conversación natural espontánea amistosa y distendida, gracias al desarrollo de numerosos corpus de lengua hablada. No ocurre lo mismo, sin embargo, en el caso de la conversación conflictiva natural entre personas íntimas y cercanas.

Se presenta en esta comunicación el camino seguido para la elaboración de un corpus de estas características, el corpus ESPRINT, y se discuten los problemas jurídicos, técnicos, metodológicos y experimentales a los que nos hemos enfrentado y las soluciones que se han aplicado. En la actualidad se dispone de en torno a 750 horas de grabación de ocho parejas españolas, de las cuales se ha transcrito en torno a un 7% del total de horas grabadas, que corresponde aproximadamente a 50 horas de conversación.

Challenges in compiling natural conversations of conflict among intimates

One of the main challenges for the linguistic characterization of conflict in private conversations is the scarcity of study materials, given their sensitivity. The current landscape of corpus linguistics is highly satisfactory regarding the compilation, processing, and storage of friendly and relaxed spontaneous natural conversation, thanks to the development of numerous spoken language corpora. However, the same does not apply to natural conflictive conversation among intimate and close individuals.

This presentation outlines the path taken for the development of the corpus ESPRINT with these characteristics, and discusses the legal, technical, methodological, and experimental issues we have faced and the solutions that have been applied. Currently, there are around 750 hours of recordings from eight Spanish couples, of which approximately 7% of the total recorded hours have been transcribed, corresponding to roughly 50 hours of conversation.

Johnatan E. Bonilla (Universiteit Gent) &
Miriam Bouzouita (Humboldt-Universität zu Berlin)

Gaming for Dialects - Annotating Corpus for Spoken Spanish (COSER-UD)

This presentation introduces the development of a Gold Standard (GS) corpus for part-of-speech (PoS) tagging tailored to the spoken dialectal variations of European Spanish, aiming to support the creation of a future treebank. Named COSER-PoS, this benchmark leverages transcriptions from the Corpus Oral y Sonoro del Español Rural (COSER; "Audible Corpus of Spoken Rural Spanish," initiated by Fernández-Ordóñez in 2005) and follows the Universal Dependencies (UD) project guidelines (Nivre et al. 2016).

This benchmark and methodology aim to enhance the understanding and processing of geographical linguistic variations by NLP tools, providing a robust framework for studying the diverse dialects of spoken Spanish.

The methodology involved several key steps:

1. Preprocessing: The first step involved preprocessing the raw data from the COSER corpus. This included normalizing the text, handling original transcription marks, and addressing transcription particularities. Following this, 500 to 600 turns of conversation from each region were sampled, ensuring a representative subset of the spoken dialectal variations.
2. Automatic PoS tagging: In the automatic PoS tagging phase, the spaCy trf model was utilized, which has been reported to achieve an accuracy of 99% for standard written Spanish. The output was formatted in the CoNLL-U[1] format, focusing on two key parts: UPOS (universal part-of-speech tags) and FEATS (morphological features).
3. Human Annotation and Verification: A comparative analysis involved human annotators who verified PoS tags through interactive games, called "Juegos del Español"[2] (Bonilla, 2023). This analysis underscored the crucial role of human intuition in PoS tagging, especially for categories unique to spoken communication such as interjections, proper nouns, and incomplete words.

[1] https://universaldependencies.org/format.html
[2] www.juegosdelespanol.com

4. Knowledge Transfer Process: This step involved the extrapolation of verified PoS tags by human annotators to unverified, automatically tagged tokens. This approach significantly accelerates the enrichment of the COSER-PoS, facilitating more accurate PoS tagging across the large dataset.
5. PoS Taggers Evaluation: State-of-the-art PoS taggers, initially trained on written Spanish data, were evaluated with the developed GS. The taggers evaluated included spaCy (Montani et al., 2022), Stanza NLP (Qi et al., 2020), and UDPipe (Straka, 2018). The accuracy of these taggers, when assessed with spoken data, dropped from a range of 98-99% to 94-95%.
6. Fine-Tuning a PoS tagger: A model was fine-tuned with the GS using the transformers spaCy training pipeline and BETO (Cañete et al., 2020). The model demonstrated a promising accuracy of 98% for UPOS and 97% for FEATS

The final COSER-UPOS GS includes a total of 13,219 sentences and 196,372 tokens, distributed among seventeen regions (see Table 1) and has been made publicly available at Github[3].:

Table 1 COSER-UD's distribution of sentences and tokens per region

	COSER-UPOS Benchmark	
Region	**Sentences**	**Tokens**
Andalusia	982	14217
Aragon	635	8264
Asturias	787	12454
Balearic Islands	617	11215
Canary Islands	1170	17272
Cantabria	664	8746
Castile	722	10542
Catalonia	892	12122
Extremadura	1118	16912
Galicia	942	16315
Madrid	596	9161
Castilla-La Mancha	731	10326
Murcia	612	7954
Navarre	839	10650
La Rioja	600	9841
Valencian Community	692	11241
Basque Country	620	9150
Total	13219	196372

[3] https://github.com/johnatanebonilla/COSER-PoS.v2

References

Bonilla, J. E., Diaz, R. L. S., & Bouzouita, M. (2023). Using GWAPs for verifying PoS tagging of spoken dialectal Spanish. In 2023 10th International Conference on Behavioural and Social Computing (BESC) (pp. 1–7). https://doi.org/10.1109/BESC59560.2023.10386542

Cañete, J. M., Chaperon, G., Fuentes, R., Ho, J., Kang, M., & Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. Papers presented at the SEPLN 2020 Conference.

Fernández-Ordóñez, I. (2005-present). Corpus Oral y Sonoro del Español Rural (COSER). Retrieved from http://www.corpusrural.es

Montani, I., Honnibal, M., Van Landeghem, S., Boyd, A., Peters, H., O'Leary McCann, P., Samsonov, M., Geovedi, J., O'Regan, J., Altinok, D., Orosz, G., Kristiansen, S. L., Miranda, L., de Kok, D., Roman, Explosion Bot, Fiedler, L., Howard, G., Edward, Wannaphong, P., Tamura, Y., Bozek, S., murat, Daniels, R., Amery, M., Böing, B., Vanroy, B., & Tippa, P. K. (2022). explosion/spaCy: v3.3.0: Improved speed, new trainable lemmatizer, and pipelines for Finnish, Korean, and Swedish. Zenodo. https://doi.org/10.5281/zenodo.6504092

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., & others. (2016). Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1659–1666).

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (pp. 197–207).

Amalia Canes-Nápoles & Eric Engel
(Universität zu Köln)

From *ninguno* to *nadie*: A variationist approach to negative quantifiers in Spanish

Speakers of Spanish dispose of two options for [+human] negative existential quantifiers, corresponding to English 'no one, nobody': *nadie* and *ninguno/a*. While the two forms have co-existed in the language for centuries (Malkiel 1945), it has been observed that *nadie* is gaining ground in the functional space traditionally occupied by *ninguno/a*, potentially leading to the eventual disappearance of the latter (RAE 2009: §48.1c). In this presentation, we analyse this putative change in progress in Peninsular Spanish using the methodology of variationist sociolinguistics, with the aim of identifying the functional and extralinguistic factors shaping the use of both forms.

Prior research in the descriptive grammatical tradition links the choice between *nadie* and *ninguno/a* to partitivity (Sánchez López 1999: 1034; RAE 2009: 20.1m). According to these authors, *nadie* —in contrast to *ninguno/a*— is dispreferred, if not ungrammatical, when followed by a partitive modifier restricting its domain to a superset, as in (1a) (SUPERSET domain). Constructions featuring prepositional modifiers of *nadie*, as in (1b), are considered pseudo-partitives, since they denote origin or class membership instead of instantiating a true subset–superset relation (ORIGIN domain). In this view, *nadie* is thus expected to occur only with ORIGIN domains, but excluded from contexts introducing SUPERSET domains, where speakers resort to using *ninguno/a*.

(1) a. ?*Nadie* / *Ninguno* de nosotros sabe la respuesta.
'None of us knows the answer.'
b. *Nadie* / ?*Ninguno* de esta clase sabe la respuesta.
'No one in this class knows the answer.'

However, analyses of authentic spoken Spanish reveal partitivity to be an unsatisfactory explanatory criterion, since we find *nadie* and *ninguno* both with SUPERSET and with ORIGIN domains, cf. (2)–(5).

(2) *Nadie* restricted by a SUPERSET domain [COSER-1401-01, line 40]
en mi casa *nadie* de los que estábamos matábamos el cerdo
'(literally:) In my house, no one among us did we kill pigs.'

(3) *Nadie* restricted by an ORIGIN domain [PRESEEA, BARC_M21_087, line 214] yo no conozco a *nadie* de mi barrio

'I don't know anybody from my district.'

(4) *Ninguno* restricted by a SUPERSET domain [PRESEEA, BARC_H21_085, lines 177–179]

para mí en mi vida las situaciones importantes fueron el nacimiento de mis hijas [...] tuve la mala desgracia de que no pude ver nacer a *ninguna* de las dos

'For me, the important moments in my life was the birth of my daughters [...] unfortunately I couldn't be there for the birth of either of the two.'

(5) *Ninguno* restricted by an ORIGIN domain [COSER-0109-02, line 359]

o sea que normalmente son todos de por aquí cerca los chicos que se han casao con las chicas de aquí, porque como..., pos eran de estos pueblos y aquí venían, pos tampoco no hay *ninguno* de lejos

'So usually they're all from around here, the guys married women from here, because well, they were from these villages and they came here, so there isn't really anyone from far away.'

This functional overlap between the two forms serves as the basis of our analysis. We extracted all 3496 occurrences of *nadie* and *ninguno/a* from two corpora of sociolinguistic interviews in Peninsular Spanish: COSER (Fernández-Ordóñez 05) and PRESEEA (PRESEEA 14). In line with the Principle of Accountability (Tagliamonte 2011: 9–11), we excluded from our sample all contexts in which the two forms are not functionally parallel or variable, i.e. instances of *ninguno/a* denoting non-human entities, non-pronominal uses, *nadie* in idiomatic or fixed expressions, or excluding mankind in general (*nadie es perfecto* 'nobody's perfect'), as well as immediate repetitions (*hacían caso ninguno, ninguno*).

We propose to replace the notion of partitivity by a multivariable analysis of factors comprising MOD (presence *vs.* absence of a modifier), the semantic type of the DOMAIN of quantification, as well as the level of determination of the domain (DET). We expect that *nadie* is favoured whenever the domain is only loosely specified (modifier absent; restriction by location, time or class types), while *ninguno/a* is favoured with higher levels of determination of the domain (modifier present, restriction by individual or clearly determined groups of individuals). With regard to extralinguistic factors, we will test the hypothesis of a change in progress by inspecting the effect of speaker AGE on rates of choosing *ninguno/a* over *nadie*.

The results of this study will help understand if the system of [+human] negative existential quantifiers is indeed undergoing reduction with *nadie* emerging as the dominant form, as well as the factors conditioning this change.

Keywords: negative indefinites, language variation and change, Spanish

References

Fernández-Ordóñez, I. (2005–). Corpus oral y sonoro del español rural.

Malkiel, Y. (1945). Old Spanish *nadi(e)*, *otri(e)*. *Hispanic Review*, 13(3):204–230.

PRESEEA (2014–). Corpus del Proyecto para el estudio sociolingüístico del español de España y de América.

RAE (2009). *Nueva gramática de la lengua española*, volume 2: Sintaxis. Espasa Calpe, Madrid, Spain.

Sánchez López, C. (1999). Los cuantificadores: Clases de cuantificadores y estructuras cuantificativas. In Bosque, I. and Demonte, V., editors, *Gramática descriptiva de la lengua española*, volume 1: Sintaxis básica de las clases de palabras, pages 1025–1128. Espasa Calpe, Madrid.

Tagliamonte, S. A. (2011). *Variationist Sociolinguistics: Change, Observation, Interpretation*. Number 40 in Language in society. Wiley-Blackwell, Chichester.

Jeanne Marie Debaisieux
(U. Sorbonne Nouvelle)

Corpora y una plataforma: Orfeo

Cualquier corpus es en sí mismo una fuente de investigación, pero como desarrollan Egbert J., Larsson T. & Biber D. (2020), la pregunta que le surge al investigador ante la diversidad de datos es la relevancia del corpus según el objeto de estudio pero también sus límites. Así es como nos gustaría aplicar al ámbito francés la observación de los autores: « *Paradoxically, doing corpus linguistics is both easier and harder than it has ever been before* »2.

El objetivo realista de la plataforma Orfeo: https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo era constituir un corpus de estudio a partir de un conjunto de corpus existentes: el Corpus para el Estudio del francés contemporáneo (CEFC). Tiene diferentes niveles de anotaciones y puede ser consultado de diversas formas en la plataforma operativa Orfeo. La contribución de esta plataforma es única en el campo de la lingüística de corpus francesa. Proporciona acceso gratuito a un corpus de estudio que comprende 4 millones de palabras orales alineadas automáticamente a nivel de fonemas. Este corpus de estudio es comparable a las subpartes con herramientas de los grandes corpus realmente disponibles en código abierto para los estudios actuales. El corpus tiene diferentes niveles de anotaciones: lemas, morfosintaxis (P.O.S.) y dependencias (funciones sintácticas) La plataforma ofrece dos herramientas operativas: un concordancia que permite “búsquedas simples” a partir de cadenas de caracteres y una herramienta de “búsqueda avanzada” que permite realizar consultas por lema, categoría y funciones.

La información sobre el tamaño del contexto y las propiedades de la estructura buscada se puede controlar mediante un retorno directo al texto en el que aparece el suceso a analizar, así como el acceso a la señal sonora. Cada archivo del corpus va acompañado de metadatos: naturaleza de los subcorpus, situaciones, fecha y ubicación de las grabaciones, duración, calidad del sonido y metadatos sobre los hablantes: sobre el tema de los intercambios, sobre el número de hablantes y sobre las relaciones que mantienen, así como una mínima información sobre la edad y su situación sociocultural en el momento del registro. Algunos de estos metadatos se pueden consultar para permitir que cada usuario cree un cuerpo de trabajo específico. También se puede acceder directamente a todos los metadatos. La

plataforma permite acceder directamente al contenido real de las transcripciones con el fin de comprobar si la naturaleza de los textos se corresponde con los comportamientos lingüísticos que interesan al investigador o al área de estudio que desea explorar. Los códigos fuente son gratuitos y ya se han utilizado para un corpus oral en italiano.

Palabras clave: Corpus, francés, métodos de interrogatorio, interfaz, anotaciones de sintaxis

Bibliografía

Debaisieux J.M. & Benzitoun C. (dir) (2020) Orfeo : un corpus et une plateforme pour l'étude du français contemporain, Langages, 219, Armand Colin

Egbert J., Larsson T. & Biber D. (2020), *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User.* Cambridge University Press; 2020.

Imelda Chaxiraxi Díaz Cabrera & Carolina Jorge Trujillo
(Universidad de La Laguna)

El contorno nuclear circunflejo en la región andina y oriental venezolana a partir del contraste corpus formal vs. espontáneo

El macroproyecto AMPER (*Atlas Multimédia Prosodique de l'Espace Roman*) comienza su andadura en el año 2002 con el objetivo de estudiar la prosodia de las lenguas y variedades románicas del espacio europeo; no obstante, muy pronto se vio la necesidad de abarcar Latinoamérica y, en general, todos los países del dominio románico. Una de estas variedades es la venezolana para cuyo estudio prosódico se sigue la división dialectal propuesta por Mora (Los Andes, Los Llanos, Centro, Oriente y región de Zulia, 1996 y 1997).

En el presente trabajo estudiamos las características melódicas de un conjunto de oraciones declarativas e interrogativas emitidas por hombres venezolanos procedentes de Mérida (zona Los Andes) y Bolívar (zona Sur-Oriental). El propósito de este trabajo es realizar un estudio entonativo fonético-fonológico (Dorta Ed. 2018) del corpus SVO diseñado en AMPER (corpus fijo) y contrastarlo con el corpus de habla semiespontáneo *Map Task* para validar o no las características del corpus formal que es el que representará a AMPER en el atlas se difundirá por internet. En estos dos informantes, el acento nuclear del corpus formal de ambas modalidades se caracteriza generalmente por un contorno circunflejo aunque en declarativas este pico final raramente tiene relevancia perceptiva. En otras regiones venezolanas, como la central (Caracas y Aragua), sin embargo, solo se realiza como circunfleja la entonación interrogativa, puesto que la declarativa comparte, con la mayor parte de las variedades del español (Prieto y Roseano, 2010), la realización con un patrón nuclear descendente. El contrate, por tanto, del corpus formal con el semiespontáneo permitirá ver si la mayor espontaneidad propicia que la configuración del pico adquiera importancia.

El análisis acústico se realizó con MatLab (López et al.) y Praat (Boersma y Weenink) y los valores absolutos de Fo, extraídos en el núcleo de la sílaba, fueron relativizados en semitonos, determinando su importancia perceptiva a partir del umbral diferencial de 1,5 semitonos.

Bibliografía

Boersma, P., & Weenink,D. (1992–2022). Praat: dooing phonetics by computer [Computer program].

Version 6.2.06, retrieved 23 January 2022 from https://www.praat.org

Dorta, J. (Ed.). (2018). La entonación declarativa e interrogativa en cinco zonas fronterizas del español: Canarias, Cuba, Venezuela, Colombia y San Antonio de Texas. Peter Lang Edition.

López Bobo, M. J., Muñiz Cachón, C., Díaz Gómez. L.; Corral Blanco, N., Brezmes Alonso D., y Alvarellos Pedrero, M. (2007). Análisis y representación de la entonación. Replanteamiento metodológico en el marco del proyecto AMPER. En J. Dorta (Ed.), *La prosodia en el ámbito lingüístico románico* (pp. 17-34.). La Página Ediciones, Colección Universidad.

Mora, E. (1996). Caractérisation prosodique de la variation dialectale de l'espagnol parlé au Vénézuéla (Tesis de Doctorado). Université de Provence.

Mora, E. (1997). División prosódica dialectal de Venezuela. *Omnia,* 2, 93-99.

Prieto, Pilar, y Roseano, Paolo (Eds.). (2010). *Transcription of Intonation of the Spanish Language.* Lincom Europa.

Josefa Dorta
(Laboratorio de Fonética, SEGAI ULL, Universidad de La Laguna)

La variación dialectal de Cuba partir del estudio de la duración en un corpus de datos oral

Para el estudio del español cubano se han establecido tradicionalmente tres zonas geolectales (Occidente, Centro y Oriente) y se han señalado pequeñas diferencias prosódicas regionales entre el occidente y el oriente de la isla (García Riverón, 1988). La inclusión de esta variedad del español dentro de AMPER (*Atlas Multimédia Prosodique de l'Espace Roman*)[1] permitió que se impulsara el análisis de la entonación en las tres zonas de la isla (*v. gr.* Dorta ed., 2013; Dorta ed., 2018). Estos trabajos de tipo descriptivo o comparativo con otras variedades del español, como la venezolana o la canaria, evidencian que, en relación con el parámetro Fo, las declarativas mantienen el patrón más general en español, pero las interrogativas se caracterizan por compartir un mismo patrón final alto-descendente o circunflejo. Por otra parte, el estudio de la duración también ha sido abordado aunque solo en habla femenina (Dorta ed., 2013), considerando el umbral del 36% definido para el español (Pamies Bertrán y Fernández Planas, 2006).

La presente propuesta tiene como objetivo comparar la duración de tres puntos de encuesta cubanos (La Habana, Santa Clara y Santiago de Cuba) desde dos perspectivas a partir de los datos acústicos obtenidos con AMPER2006 (Brezmes Alonso, 2007): la primera, el etiquetaje para tratar de reflejar las relaciones entre la tónica y sus adyacentes (pretónica y postónica) en los distintos puntos de encuesta siguiendo la propuesta de Muñetón Ayala, Díaz y Dorta (2018) y, la segunda, dialectométrica para establecer las relaciones de distancia y proximidad entre las diferentes zonas dialectales cubanas y comprobar la coherencia de los resultados obtenidos en el análisis acústico. El corpus de análisis es el denominado en el proyecto AMPER *formal sin expansión en los sintagmas de frontera,* emitido mediante elicitación textual, del tipo SN + SV + SPrep como, por ejemplo: *el saxofón se toca con obsesión* o *la cítara se toca con pánico*.

Bibliografía

Brezmes Alonso, D. (2007). *Desarrollo de una aplicación software para el análisis de características fundamentales de la voz*. (Proyecto de fin de carrera). Universidad de Oviedo. Oviedo.

Dorta, J. (ed.) (2013). *Estudio comparativo preliminar de la entonación de Canarias, Cuba y Venezuela*. Madrid–Santa Cruz de Tenerife: La Página ediciones S/L, Colección Universidad.

Dorta, J. (ed.) (2018). *La entonación declarativa e interrogativa en cinco zonas fronterizas del español: Canarias, Cuba, Venezuela, Colombia y San Antonio de Texas.* Peter Lang Edition. Studien zur romanischen sprachwissenschaft und interkulturellen kommunikation. Herausgegeben von Gerd Wotjak.

García Riverón, R. (1988). *Atlas Lingüístico de Cuba. Cuestionario*. La Habana: Academia de Ciencias de Cuba.

Muñetón Ayala, M., Díaz, C., y Dorta, J. (2018). La duración en oraciones sin expansión en la voz femenina de dos países fronterizos: Colombia (Bogotá -Medellín) y Venezuela (CaracasMérida)", *Literatura y Lingüística*, 37, 401-423.

Pamies Bertrán, A., y Fernández Planas, A. Mª. (2006). "La percepción de la duración vocálica en español". En J. D. Luque Durán (Ed.), *Actas del V Congreso Andaluz de Lingüística General. Homenaje al Profesor José Andrés de Molina Redondo* I (pp. 501-512). Granada: Lingvistica-Ediciones Método.

---

[1] AMPER nace en el año 2002 en el Centre de Dialectologie de l'Universitè Stendhal-Grenoble III (Francia) con el propósito de realizar un atlas multimedia en el que plasmar la prosodia del espacio románico. AMPER fue coordinado por Michel Contini (Universitè Stendhal-Grenoble III) y por Antonio Romano (Università di Torino) hasta principios de 2015. Este último es su coordinador en la actualidad.

Jorge Fernández Jaén
(Universidad de Alicante)

"Ya me lo imagino": análisis pragmático de las formas epistémicas del verbo *imaginar(se)* a partir de su uso oral

El verbo *imaginar* posee en el español actual, sobre todo en su dimensión oral, numerosos usos de carácter epistémico, como *me lo imagino*, *ya imagino* o el imperativo de carácter discursivo *imagínate*. Todos estos usos son el resultado de un proceso de epistemización que tiene su origen en el significado original del verbo, basado en la noción general de ´crear imágenes con la mente`. En esta comunicación, presentaremos un análisis de los empleos epistémicos de la primera persona del verbo –*imagino* y todas sus variantes- utilizando (de forma complementaria, ya que las ocurrencias son limitadas) los datos que ofrecen el corpus del español coloquial de Val.Es.Co, el corpus audiovisual GestINF y el Corpus MEsA. A partir del examen de los datos de los tres corpus, intentaremos demostrar las siguientes hipótesis:

- El valor epistémico del verbo *imaginar* se ha desarrollado siguiendo las pautas establecidas por Halliday y Hasan (1975) y Traugott (1982) en sus estudios sobre (inter)subjetivización y gramaticalización; así, *imaginar* parte de un contenido proposicional basado en una situación interna (imaginarse a uno mismo) para después pasar a una situación externa (imaginarse algo exterior) y acabar, en el final del proceso, en un contenido subjetivo basado en creencias totalmente desconectadas de cualquier figuración imaginística. Los corpus orales demostrarán, por ejemplo, que la ausencia del pronombre átono *me* (*me imagino* frente a *imagino*) puede implicar un descenso de la subjetividad y la introducción de especulaciones epistémicas que no atañen directamente al hablante.
- De acuerdo con Maldonado (2019), cuando *imaginar* se utiliza con un valor reflexivo, pone en marcha siempre dos espacios mentales, de modo que el hablante se desdobla y aparece tanto en el espacio base -la realidad-, como en un espacio mental abstracto -las imágenes creadas en la mente-. Los datos de corpus evidenciarán que esos desdoblamientos semánticos también se producen en el discurso oral, ya que resultan muy útiles para la elaboración de argumentos.
- El verbo *imaginar* presenta en los usos orales una función clara de reciclaje discursivo, relacionada con el concepto de información compartida (Clark, 1996; Rodríguez Rosique, 2021; Heine, 2023). De forma más específica, los datos

mostrarán que *imaginar* funda sus especulaciones epistémicas en la información previa (información que actúa como antecedente, en los términos que propone Maldonado (2010)), sobre la cual el hablante elabora subjetivamente sus opiniones.

- Los usos orales de *imaginar* admiten una interpretación evidencial. Los ejemplos de corpus permitirán defender la tesis de que este verbo puede considerarse como un evidencial léxico asociado a las evidencias reportadas e inferidas; el hablante recaba de la comunicación inmediata información y, tras un proceso de reflexión, construye contenidos nuevos que ayudan a dinamizar el discurso y a reforzar la retroalimentación que se establece entre hablante e interlocutor/es.

En definitiva, esta comunicación ampliará el análisis de los recursos de que dispone la lengua española para generar información espontáneamente en el ámbito oral y ofrecerá reflexiones originales sobre el uso del verbo *imaginar* en sus variaciones de tipo epistémico.

Palabras clave: (inter)subjetividad, epistemización, español oral, evidencialidad

Jorge Fernández Jaén (Universidad de Alicante) &
Herminia Provencio Garrigós (Universidad de Murcia)

Mi coche es caro no, lo siguiente: análisis cognitivo de una estructura cuantificativa del español actual

Desde hace unos pocos años, ha empezado a documentarse en la lengua española -tanto oral como escrita- una construcción cuantificativa de carácter superlativo que responde al patrón construccional ´X no, lo siguiente`. Esta estructura ha desatado numerosos debates en el ámbito de la gramática normativa y de la sociolingüística, ya que expresa el grado superlativo y la cuantificación escalar de un modo muy llamativo. El propósito de nuestra comunicación es ofrecer una explicación lingüística tanto de esta estructura como de la versión menos conocida ´X no, lo anterior`, con objeto de poner de manifiesto sus características gramaticales y las motivaciones semánticas que están detrás de su aparición. De forma más específica, analizaremos desde un planteamiento cognitivo-funcional los siguientes puntos: a) los componentes sintácticos de ambas construcciones, b) su significado semántico de base espacial, c) su prototipo construccional y sus variaciones y d) sus rasgos específicos en tanto que estructura cuantificativa (tipos de base cuantificada, propiedades icónicas, etc.). Todo el estudio se basará en un amplio corpus de más de mil ocurrencias de las dos construcciones, recopilado a partir de corpus lingüísticos (Corpus del Español Web/Dialectos, Corpus del Español NOW y CORPES XXI) y de redes sociales (Instagram, YouTube, Facebook y Twitter). Nuestro trabajo probará, en definitiva, que las expresiones ´X no, lo siguiente` y ´X no, lo anterior`, lejos de ser una mera ocurrencia pasajera, constituyen una innovación -originada en la lengua oral- muy interesante, que responde a pautas de eficacia comunicativa sumamente sofisticadas.

Palabras clave: español oral, cuantificación, grado superlativo, metáfora cognitiva

Bibliografía

Fernández Jaén, J. y Provencio Garrigós, H. (2020): "Interesante no, lo siguiente: análisis cognitivo de una construcción cuantificativa del español actual", *LEA*, 42 (1), 41-59.

Raffaella Gambardella
(Universidad de Salerno)

Potenciar la anotación pragmática con los grandes modelos de lenguaje (LLM): aceleración del proceso y contextualización con GPT-4

En los últimos años, los Grandes Modelos de Lenguaje (LLM) han tenido un enorme éxito en diversos campos de investigación. La arquitectura básica de la mayoría de los Grandes Modelos de Lenguaje (LLM) se basa en los denominados *tranformers* (Vaswani et al., 2017). Aunque los detalles exactos de la arquitectura de GPT-4 aún no están bien delineados, sabemos que el modelo subyacente es un *transformers* entrenado capaz de predecir el siguiente *token* dentro de una secuencia (OpenAI, 2024). Esta investigación específica pretende explorar las posibilidades que ofrece la inteligencia artificial y, gracias al apoyo de los Grandes Modelos de Lenguaje (LLM) y del modelo GPT-4, se está intentando proceder a la anotación pragmática de diálogos en lengua española (*corpus DiEspa - Diálogos en Español*) (ParlarItaliano, 2008). En la base del proyecto se encuentra el esquema de anotación de diálogos Pr.A.T.I.D. (*Pragmatic Annotation Tool for Italian Dialogues*) (Savy, 2010), que en el momento de su creación requería el uso de un software de anotación manual (XGate), ahora obsoleto. Dado que ya no es posible anotar con el software XGate (Cutugno, D'Anna, 2006), se decidió intentar entrenar el modelo GPT-4 con el objetivo de acelerar la anotación pragmática de diálogos y refinar su memoria. Tradicionalmente, la anotación pragmática requiere una intensa actividad humana para identificar y clasificar los actos lingüísticos y los fenómenos pragmáticos complejos, un proceso que puede resultar especialmente complejo y estar sujeto a múltiples interpretaciones. El modelo GPT-4 se entrena con un amplio corpus de textos que abarcan una extensa gama de temas, lenguas y estilos comunicativos, lo que permite generar respuestas contextualizadas y comprender los matices del lenguaje humano.

Mediante un enfoque iterativo, el modelo se adaptó para reconocer y anotar fenómenos pragmáticos específicos basados en el esquema de anotación de diálogos Pr.A.T.I.D. La metodología adoptada implica el preprocesamiento de los diálogos para adaptarlos a los requisitos del esquema Pr.A.T.I.D. (Castagneto, 2012), seguido del uso del modelo GPT-4 para la identificación preliminar de los *moves* dialógicos (Savy, Solís García, 2009) esperados que deben detectarse en el texto del diálogo. Posteriormente, las anotaciones generadas automáticamente fueron revisadas y corregidas por expertos humanos, garantizando así un alto nivel de precisión y

fiabilidad. El trabajo de revisión permitió no sólo identificar posibles errores, sino corregirlos con el objetivo de perfeccionar el sistema y permitir que el modelo GPT-4 mejorara su rendimiento. Este enfoque "híbrido" combina las capacidades de contextualización del modelo GPT-4 con la experiencia humana manteniendo altos niveles de calidad y coherencia.

En conclusión, este estudio demuestra el potencial de los Grandes Modelos de Lenguaje (LLM) y del modelo GPT-4 para superar algunas de las limitaciones asociadas a la anotación pragmática manual. El objetivo es allanar el camino para futuras investigaciones sobre la aplicación de modelos generativos del lenguaje en otras áreas de la lingüística computacional y la posibilidad de ampliar este enfoque a otras lenguas y tipos de texto. Las implicaciones de este estudio se extienden más allá del campo de la anotación pragmática, sugiriendo nuevas direcciones para la integración de los Grandes Modelos de Lenguaje (LLM) en los procesos de análisis lingüístico.

Palabras clave: grandes modelos de lenguaje, anotación pragmática, GPT-4, diálogos

Bibliografía

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). *"Attention is All you Need."* In Advances in Neural Information Processing Systems. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://arxiv.org/abs/1706.03762

OpenAI, (2024), Arxiv, URL: https://arxiv.org/abs/2303.08774

ParlarItaliano, Studium Dipsum, (2006), https://parlaritaliano.studiumdipsum.it/it/792-corpus-diespa-dialogos-en-espanol

Savy Renata, (2010) *"Pr.A.T.I.D: a coding scheme for pragmatic annotation of dialogues"*, Department of Linguistic and Literary Studies University of Salerno, Fisciano (Salerno), Italy

Cutugno Francesco, D'Anna Leandro, (2006) *"XGate e XRG: strumenti per l'editing visuale, l'interrogazione e il benchmarking di annotazioni linguistiche XML"*, Dipartimento di Fisica - Gruppo NLP, Università 'Federico II' di Napoli, Italia, Dipartimento di Linguistica e Letteratura, Università di Salerno, Italia

Castagneto Marina, (2012), *"Il sistema di annotazione Pra.Ti.D tra gli altri sistemi di annotazione pragmatica. Le ragioni di un nuovo schema"*, in ANNALI del Dipartimento di Studi Letterari, Linguistici e Comparati Sezione Linguistica, Università degli Studi di Napoli "L'Orientale", Napoli, Italia

Savy Renata, Solís García Inmaculada (2008), *"Strategie pragmatiche in italiano e spagnolo a confronto: una prima analisi su corpus"*, Università degli Studi di Salerno, Fisciano (Salerno), Italia.

Davide Garassino[1], **Lorenzo Filipponio**[2] & **Dalila Dipino**[3]
([1]Zurich University of Applied Sciences, [2]University of Genoa, [3]University of Zurich)

Contrastive Vowel Length in Ligurian. Constructing a corpus for micro-dialectological research on non-standard Northern Italo-Romance varieties

The main purpose of this contribution is to present a corpus of spoken data collected from three non-standard Italo-Romance language varieties and to illustrate its potential for exploring relevant issues in contemporary research on phonology. Such varieties belong to the Ligurian group (Northern Italo-Romance) and are: Genoese and the dialects of Porto Maurizio, representing Western Ligurian, and Ventimiglia, representing Intemelian Ligurian (for a classification of Ligurian dialects, see Forner, 1988 and Toso, 1995).

The corpus was compiled as part of a research project hosted at the University of Zurich (20182022) utilizing the production data of 62 native speakers: 21 informants from Genoa, 20 from the Western Ligurian area, and 21 from the Intemelian coast (comprising 19 females and 42 males aged between 21 and 88, averaging 67 years). Each participant engaged in several tasks designed to elicit information about dialect usage across various speech styles (in the spirit of Wagner et al., 2015), including (contrastive) carrier sentences, SVX sentences, Map Task, and spontaneous conversations.

Beyond documenting severely endangered language varieties, the primary research objective of the project was to investigate contrastive vowel length (CVL) in Ligurian by analyzing its main phonetic cues: vowel and consonant duration. The production tests utilized as target items (sub)minimal pairs such as /ˈkaːze/ 'case' vs. /ˈkaze/ 'fall', /ˈnaːzu/ 'nose' vs. /ˈmazu/ 'may', /ˈpɔːku/ 'little' vs. /ˈtɔku/ 'piece', /ˈfryːtu/ 'fruit' vs. /ˈbrytu/ 'dirty, ugly', etc.

Two main findings are presented in this contribution, primarily based on carrier sentences and SVX data, so as to highlight the role of the corpus in exploring micro-dialectological and individual variation. Firstly, CVL is still present in Genoese and Western Ligurian, although with variations in its phonetic implementation, while it has completely disappeared in Intemelian Ligurian, thus empirically confirming traditional dialectological observations (e.g., Azaretti, 1982; for other experimental work, see Garassino, Loporcaro & Schmid, 2017; Filipponio & Garassino, 2019; Dipino, Filipponio & Garassino, 2022). Secondly, the extensive participation of native speakers in the dialectological survey allows for a detailed investigation of inter-

speaker variation. This analysis reveals that even in Genoese, where CVL is robustly observed at the group level, this phonological feature is not consistently present across all speakers and appears to be declining. Additionally, it will be shown that this 'fading' trajectory of CVL deviates from that of other northern Italo-Romance varieties (Loporcaro, 2015), suggesting a possible language contact explanation due to the pressure of standard Italian (Cerruti, Crocco & Marzo, 2017).

References

Azaretti, E. (1982) [1977]. *Etimologia dei dialetti liguri attraverso l'evoluzione del ventimigliese*. Casablanca.

Cerruti, M., Crocco, C., & Marzo, S. (Eds.) (2017), *Towards a New Standard. Theoretical and Empirical Studies on the Restandardization of Italian*. De Gruyter.

Dipino, D., Filipponio, L., & Garassino, D. (2022). Manifestazioni della quantità vocalica nella Liguria centro-occidentale: tipologia e metodologia. In L. Baranzini, S. Christopher, & M. Casoni (Eds.), *Linguisti in contatto 3. Ricerche di linguistica italiana in Svizzera e sulla Svizzera* (pp. 17-37). Osservatorio linguistico della Svizzera italiana.

Filipponio, L., & Garassino, D. (2019). Center and Periphery in Phonology: a "stress-test" for two Ligurian Dialects. *Italian Journal of Linguistics, 31(2),* 141-168.

Forner, W. (1988). Areallinguistik I. Ligurien / Aree linguistiche I. Liguria. In G. Holtus, M. Metzeltin, & C. Schmitt (Eds.), *Lexikon der Romanischen Linguistik Vol. IV* (pp. 453469). Max Niemeyer Verlag.

Garassino, D., Loporcaro, M., & Schmid, S. (2017). La quantità vocalica in due dialetti della Liguria. In C. Bertini, C. Celata, G. Lenoci, C. Meluzzi, & I. Ricci (Eds.), *Fattori biologici e sociali nella variazione fonetica / Social and Biological Factors in Speech Variation* (pp. 127-144), Officinaventuno.

Loporcaro, M. (2015), *Vowel Length from Latin to Romance*. Oxford University Press.

Toso, F. (1995). *Storia linguistica della Liguria*. *Vol. I, Dalle origini al 1528*. Le Mani.

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics, 48,* 1-12.

Mar Garachana Camarero (Universitat de Barcelona) &
Daniel Cuní Díez (Universitat de Barcelona & Universitat Internacional de Catalunya)

Las perífrasis verbales del español en el corpus multimodal GRADESBAR. Frecuencia de empleo y productividad en la lengua oral

El objetivo de esta comunicación es llevar a cabo una investigación empírica acerca de la frecuencia de empleo y la productividad de 100 perífrasis verbales del español en el corpus oral GRADESBAR (Corpus Oral del español de Barcelona), compilado por el grupo GRADIA de la *Universitat de Barcelona*. Con este trabajo aspiramos a responder a las siguientes preguntas de investigación:

1. ¿Cuál es la distribución funcional de las perífrasis verbales del español en la lengua oral?
2. ¿Cuál es la frecuencia de empleo y la productividad real de las perífrasis verbales en el español hablado? ¿Existe una correlación entre el elevado número de perífrasis verbales que tiene el español y su empleo efectivo en la lengua?
3. ¿Cuál es la frecuencia y la productividad de las cadenas de auxiliares en la lengua hablada? ¿Es su empleo marginal? ¿Existe algún patrón ligado a la tradicionalidad discursiva que condicione su empleo?
4. ¿Cuáles son las diferencias de frecuencia y de productividad de perífrasis sinónimas?
5. ¿Existen diferencias en el empleo de las perífrasis verbales entre el español estándar y el español hablado en Barcelona que puedan explicarse por contacto de lenguas?

Los resultados de esta investigación permitirán sentar las bases para futuros estudios sobre el contacto de lenguas (español-catalán) en el ámbito de las perífrasis. Asimismo, las conclusiones relacionadas con la frecuencia y la productividad posibilitarán futuras comparaciones entre las muestras lingüísticas propias de la proximidad y la distancia comunicativas.

Esta investigación se inserta en una aproximación al estudio lingüístico basada en el uso (Bybe 1985), en la línea de los trabajos colostruccionales y de gramática de construcciones (Rosemeyer 2016, Enghels 2018, Cuní 2013, Garachana y Sansiñena 2023). El estudio parte de un vaciado completo de todas las perífrasis verbales contenidas en el corpus GRADESBAR. Con los ejemplos obtenidos se realizará un análisis de colexemas (Gries & Stefanowitsch 2004) y se estudiará la productividad

sintáctica realizada de todas las perífrasis (Baayen 2009), atendiendo a la *Aktionsart* y al significado de los verbos auxiliados.

Palabras clave: perífrasis verbales, cadenas de auxiliares, productividad, frecuencia, GRADESBAR

Bibliografía

Baayen, H. R. (2009). Corpus linguistics in morphology: morphological productivity. En: A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An international handbook*. Berlín: Mouton De Gruyter, 900–919.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Cuní, D. (2023). *Un estudio diacrónico de "estar + gerundio"* (Tesis doctoral). Barcelona: Universitat de Barcelona.

Enghels, R. (2018). "Towards a constructional approach to discourse-level phenomena: The case of the Spanish interpersonal epistemic stance construction". *Folia Linguistica*, 52(1): 107-138.

Garachana, M. y Sansiñena, S. (2023). "Combinatorial Productivity of Spanish Verbal Periphrases as an Indicator of Their Degree of Grammaticalization". *Languages*, 8(3): 1-25.

Gries, S. & Stefanowitsch, A. (2004). Extending collostructional analysis A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics,* 9(1): 97-129.

Rosemeyer, M. (2016). "Modeling frequency effects in language change". En: H. Behrens y S. Pfänder (Eds.), *Experience Counts: Frequency Effects in Language*. Berlín, Boston: De Gruyter, 175-208.

Mar Garachana & Magdalena Rosková
(Universitat de Barcelona)

El corpus multimodal GRADESBAR: recogida, codificación y manejo de datos audiovisuales en la investigación lingüística

La presente contribución tiene por objetivo presentar el corpus audiovisual GRADESBAR (GRADIA del español de Barcelona) y mostrar algunas de sus aplicaciones en el terreno de los estudios lingüísticos. GRADESBAR es un corpus multimodal, que se está confeccionando desde el año 2018 en el seno del grupo de investigación GRADIA (Gramática y Diacronía). El corpus GRADESBAR forma parte de la colección SOFA TALKS, un tipo de corpus en el que se graban a dos personas que charlan acerca de temas libres. Actualmente, el corpus contiene un total de 42 horas de grabaciones de conversaciones diádicas en el español coloquial entre hablantes nativos que residen en Barcelona. Los datos se presentan en formato de vídeo, de audio y de transcripciones de los diálogos. En esta comunicación se va a exponer el procedimiento metodológico que se ha seguido para la confección del corpus GRADESBAR, desde su concepción hasta su materialización en grabaciones transcritas. En concreto, se describirá cómo se llevó a cabo la recogida de datos mediante herramientas de grabación, su codificación a través de programas como Praat (Boersma y Weenink 2013) o ELAN (Wittenburg *et al.* 2006), siguiendo las normas de transcripción GAT2 (Ehmer *et al.* 2019).

El corpus viene a llenar dos vacíos en la lingüística hispánica actual. Por un lado, GRADESBAR es el banco de datos más extenso que existe en la actualidad para abordar el estudio del español de Barcelona, donde el español convive con el catalán. De esta manera, el corpus permite analizar cómo el contacto lingüístico del español con el catalán se materializa en las producciones lingüísticas de los hablantes. De este modo, el corpus es esencial para abordar un análisis de las características del español hablado en Barcelona, en la línea de los trabajos de Vila (2005, 2007).

Por otro lado, al tratarse de grabaciones en vídeo, el corpus constituye una base de datos idónea para las investigaciones empíricas del español contemporáneo que tratan de llevar a cabo análisis de la conversación desde el enfoque multimodal, tratando de analizar la confluencia de diferentes modalidades comunicativas en la conversación en persona (Streeck et al. 2011). Para mostrar las posibles aplicaciones de este corpus, se hará referencia a estudios sobre contacto lingüístico relacionados con el empleo de las perífrasis verbales del español, así como a estudios

multimodales actualmente en marcha. Concretamente, se presentarán investigaciones relacionadas con el empleo de gestos como estrategia comunicativa destinada a completar semánticamente oraciones sintácticamente incompletas (Rosková y Satti en prensa). También se analizará el empleo de la mirada como recurso comunicativo por parte de los hablantes españoles y marroquíes (Rosková y Aderdouch Derdouch en preparación). Dado que los trabajos sobre el español de Barcelona y sobre la multimodalidad del español constituyen un área de investigación todavía en desarrollo, el corpus GRADESBAR adquiere relevancia metodológica, empírica y teórica dentro de la lingüística hispánica actual.

Palabras clave: corpus multimodal, datos audiovisuales, metodología lingüística, multimodalidad, español coloquial.

Bibliografía

Boersma, P. y Weenink, D. (2013). *Praat: doing phonetics by computer* (5.3.39) [software]. Recuperado de http://www.fon.hum.uva.nl/praat/download_win.html

Ehmer, O., Satti, L. I., Martínez, A. y Pfänder, S. (2019), "Un sistema para transcribir el habla en interacción: GAT 2", G*esprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 20.

Rosková, M. y Aderdouch Derdouch, S. (en preparación), "El árabe marroquí y el español en contacto: análisis contrastivo del empleo de la mirada como recurso comunicativo".

Rosková, M. y Satti, I. (en prensa), "Oraciones multimodales: la compleción gestual como recurso comunicativo", Signo y seña, 45.

Streeck, J.; Goodwin, Ch. y LeBaron, C. (eds.) (2011), Embodied Interaction: Language and Body in the Material World. Cambridge: Cambridge University Press.Vila, R. (2005), "Corpus para el estudio de las interferencias lingüísticas: los corpus de Barcelona, Lérida y Bilbao", Oralia: Análisis del discurso oral, 8, 213-242.

Vila, R. (2007), "Sociolinguistics of Spanish in Catalonia". International journal of the sociology of language,184, 59-77.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. y Sloetjes, H. (2006), "ELAN: a professional framework for multimodality research," en *5th International Conference on Language Resources and Evaluation* (LREC 2006) (Genoa), 1556-1559.

Tajana Gavon
(University of Heidelberg)

Investigating endogenous norms of French: a spoken corpus of Togolese French

Studying diatopic variety in worldwide spread languages is a continuing concern within linguistics. These languages are subject to various sociocultural environments and to language contact situations. Therefore, the African continent, where multilingual diversity is particularly rich, represents a research area of great interest within the field of variational linguistics. For example, a recent project moved the focus on the situation of Romance languages in Africa (cf. https://www.geku.uni-passau.de/en/romance-languages/research/rola).

In the *LingCor Workshop* I propose to present a spoken corpus dealing with the variety of French spoken in the West African Republic of Togo. According to the latest report of the International Organisation *La Francophonie* (*OIF*) nearly half of the world's French speakers are found in regions of sub-Saharan Africa and Indian Ocean. In view of the worldwide strongest increase of francophones in this area within the last years, the *OIF* underlines the significance of the African countries for the development of the French language (cf. OIF report *La langue française dans le monde* 2022, 6). Despite the importance of sub-Saharan regions for the francophone world, there remains a paucity of recent corpus-based research for several countries and research domains. The article dealing with Togo in the recently published *Manual of Romance languages in Africa* concludes that full-scale projects on (socio-)linguistic aspects of Togolese French represent a desideratum (cf. Essizewa / Kpoglu / van den Berg 2024, 410). Up to now, far too little attention has been paid specially to spoken African varieties of French – a research gap that is due to challenges in spoken corpus construction (cf. Diao-Klaeger 2018, 14).

The poster presents a corpus that was built as basis for my doctoral research on the endogenous norm of Togolese French. It contains authentically produced spoken data extracted from local media. The corpus is divided in two subcorpora in order to give the possibility to compare linguistic aspects in different styles. Based on the model of a continuum between the *language of distance* and the *language of immediacy* established by Koch / Oesterreicher one of the subcorpora consists of material from rather formal situations (e.g. news broadcast) while the other subcorpus contains data from informal language contexts. It is to mention that parts of the corpus also represent a resource for codeswitching and language contact studies between

French and Ewe/Mina. The poster introduces the selection and organization of the data, presents the transcriptions, and considers the supporting software.

Keywords: Diatopic variety, French, Togo, spoken corpus, media corpus

References

Diao-Kläger, Sabine (2018): *Diskursmarker. Eine Studie zum gesprochenen Französisch in Burkina Faso*, Stauffenburg Verlag, Tübingen.

Essizewa, Komlan / Kpoglu, Promise Dodzi / van den Berg, Margot (2024): "Togo", in: Reutner, Ursula (ed.): *Manual of Romance languages in Africa*, De Gruyter, Berlin / Boston, 391–412.

Koch, Peter / Oesterreicher, Wulf ²(2011): *Gesprochene Sprache in der Romania*, De Gruyter, Berlin / New York.

Project *Romance Languages in Africa*: https://www.geku.uni-passau.de/en/romance-languages/research/rola

Report 2022 *Organisation International de la Francophonie*: https://www.francophonie.org/sites/default/files/2022-03/Synthese_La_langue_francaise_dans_le_monde_2022.pdf

Katharina Gerhalter
(University of Graz)

Tu madre coser cose bien, pero dibujar, dibuja mal. Recopilación y análisis de infinitivos topicalizados en el español hablado

La construcción [TOPIC infinitivo topicalizado] + [COMMENT verbo conjugado del mismo lema], como en el ejemplo *comer no come mucho*, muestra un comportamiento pragmático complejo, ya que desencadena inferencias como, por ejemplo, un contraste implícito de tipo *pero bebe un montón* (Bastos 2001; Valenzuela et al. 2005; Reich 2011; Vicente 2007; Hein 2020; Verdecchia 2021; Muñoz Pérez & Verdecchia 2022). La construcción se limita a contextos muy específicos en los que el infinitivo topicalizado explicita la *Question under Discussion* (= QUD), de acuerdo con la propuesta de Verdecchia (2021) y Muñoz Pérez & Verdecchia (2022). Se trata de un fenómeno típico de la lengua hablada y, especialmente, de diálogos coloquiales; por ejemplo, aparece predominantemente como réplica o respuesta (Narbona Jiménez 2015).

A pesar de que la interpretación de la construcción [TOPIC infinitivo topicalizado] + [COMMENT verbo conjugado del mismo lema] es altamente sensible al contexto, no existen hasta el momento estudios que hayan analizado los infinitivos topicalizados dentro de su contexto discursivo en corpus del español hablado. Los estudios fundamentales de Vicente (2007), Valenzuela et al. (2005), Verdecchia (2021) y Muñoz Pérez & Verdecchia (2022) se basan en la introspección y en el análisis de unos pocos ejemplos aislados y construidos. Otros autores mencionan la construcción sólo de manera anecdótica (NGLE 2009, Hernanz Cabó 1999, Narbona Jiménez 2005). Al ser una estructura muy poco frecuente, las ocurrencias en los corpus son muy escasas en comparación con otros tipos de topicalización, sobre todo, de elementos nominales (Hidalgo Downing 2003).

Por lo tanto, en nuestro proyecto de investigación pretendemos aportar nuevos datos acerca de la construcción del infinitivo topicalizado gracias a la recopilación de una base más extensa de ejemplos auténticos sacados de dos fuentes principales:

- Recopilación de un corpus propio de ejemplos televisivos (provenientes de rtve.es): 43 ejemplos hasta la fecha.
- Búsqueda sistemática de la estructura en corpus orales ya establecidos: 33 ejemplos del COSER (hasta la fecha) y futuras búsquedas en otros corpus como Val.Es.Co y CORAL-ROM.

Los ejemplos recopilados nos permiten analizar la amplia gama de valores pragmáticos de dicha construcción dependiendo de su contexto de uso: por ejemplo, ¿siempre desencadenan inferencias contrastivas o adversativas? ¿siempre son réplicas que retoman explícitamente un verbo ya mencionado anteriormente? Además, nos permiten detectar patrones sintácticos recurrentes como la secuencia de sujeto topicalizado seguido de un infinitivo topicalizado:

(1) *Tu madre **coser** cose bien, pero **dibujar**, dibuja mal.*
(https://www.rtve.es/alacarta/videos/maestros-de-la-costura-3/maestros-costura-3-programa-10-final/5548674/ (min. 2:43:55))

Por último, nuestra colección de ejemplos es lo suficientemente grande para sacar también algunas conclusiones cuantitativas sobre los valores pragmáticos y los patrones de uso de esta construcción (aserción enfática; adversatividad como implicatura conversacional; división de una QUD mayor en sub-QUDs). La finalidad del trabajo consiste en discernir entre la función básica inherente a esta construcción y sus funciones contextuales.

Palabras clave: infinitivos topicalizados, topicalización de predicados, pragmática, discurso, *question under discussion*

Bibliografía

Bastos, A. C. P. (2001). *Fazer, eu faço! Topicalização de constituintes verbais em português brasileiro*. Master's Thesis, Universidade Estadual de Campinas, Campinas, SP.

COSER = Inés Fernández-Ordóñez (dir.) (2005-): *Corpus Oral y Sonoro del Español Rural*. www.corpusrural.es

Hein, J. (2020). *Verb Doubling and Dummy Verb: Gap Avoidance Strategies in Verbal Fronting*. De Gruyter. https://doi.org/10.1515/9783110635607

Hernanz Carbó, M. L. (1999). El infinitivo. In I. Bosque & V. Demonte Barreto (Eds.): *Gramática descriptiva de la lengua española* (Vol. 2, pp. 2197–2356). Espasa Calpe.

Hidalgo Downing, R. (2003). *La tematización en el Español hablado. Estudio discursivo sobre el Español peninsular*. Gredos.

Muñoz Pérez, C., & Verdecchia, M. (2022). Predicate doubling in Spanish: On how discourse may mimic syntactic movement. *Natural Language & Linguistic Theory* (40), 1159–1200. https://doi.org/10.1007/s11049-022-09536-3

Narbona Jiménez, A. (2015). *Sintaxis del español coloquial*. Editorial Universidad de Sevilla.

Reich, U. (2011). *Frontalizaciones de la semántica verbal en español y portugués*. Talk presented at 18. Deutscher Hispanistentag, Passau.

Valenzuela, J., Hilferty, J., & Garachana-Camarero, M. (2005). On the reality of constructions: The Spanish reduplicative-topic construction. *Annual Review of Cognitive Linguistics* (3), 201–215. https://doi.org/10.1075/arcl.3.11val

Verdecchia, M. (2021). Impossible Presuppositions. On factivity, focus, and triviality. *Glossa* 6(1), Article 92, 1–29. https://doi.org/10.16995/glossa.5879

Vicente, L. (2007). *The Syntax of Heads and Phrases: A Study of verb (phrase) Fronting*. PhD. dissertation, Landelijke-LOT, Leiden University.

Karolina Grzech
(University of Valencia)

Comparative research on minority and dominant languages: a view from spoken corpora

For languages with established literary traditions, large speaker populations, and normative grammars, spoken corpora are one of many available data sources for linguistic research. For minority and endangered languages, on the other hand, such corpora often constitute the only available source of primary data. At the same time, spoken corpora of majority and minority languages differ significantly in their design, structure, and analytical potential. These discrepancies have acute consequences for the scope and quality of possible comparative and typological studies.

This is particularly relevant in Romance linguistics within Latin America, where Spanish and Portuguese interact with a super-diverse array of indigenous languages (Van Gijn 2018). This talk explores the theoretical, analytical, and methodological implications arising from differences in available spoken corpora across language types. Specifically, it focuses on evidentiality—a linguistic category studied extensively in Romance and indigenous languages of Latin America.

Evidentiality is most often defined as indicating the 'source of information' for what is being said (cf. Aikhenvald, 2004). The example from Cuzco Quechua (Quechuan, Peru, Faller 2002:

122) demonstrates how it can work when it is expressed by dedicated morphemes:

*Parashan**mi*** 'It is raining' [the speaker sees the rain]

*Parashan**chá*** 'It is raining' [the speakers conjectures it without observing the rain]

*Parashan**si*** 'It is raining' [the speaker was told by another person]

As these examples suggest, the use of evidentials does not change the propositional content of the utterance. Rather, it adds an additional layer of meaning (cf. Faller 2002; Boye 2012), the precise nature of which varies between languages and remains an object of study and debate. In line with the established definition of evidentiality, this additional layer is widely assumed to indicate 'how the speaker knows.' However, a growing body of descriptive research shows that evidential expressions do more than that, or that their function is different altogether. This mismatch is particularly evident when we analyse language-in-use, rather than isolated

sentences, and thus can be appreciated most clearly in research based on spoken corpora.

It has already been established that interactional uses of grammatical evidentials, attested e.g. in indigenous languages of Latin America, and those of 'evidential strategies' (cf. Aikhenvald, 2004) attested e.g. in Romance languages, are functionally very similar (Mushin, 2013). This implies that to be accurate, any cross-linguistic generalisations about evidentiality should incorporate data on both types of systems, so as to cover grammatical and lexical evidentiality. This, in turn, means that we should be able to consider and compare data from large spoken corpora, created, curated and annotated by entire teams of native speakers (KiParla, Val.Es.Co. Ameresco, etc.), with small corpora created by a sole researcher, often an outsider to the speech community (e.g. those available in language archives AILLA or ELAR). Based on previous studies of evidentials in Romance and Amerindian languages, this talk covers the issues we need to consider to make comparative corpus-based research possible, including, but not limited to the size of corpora, granularity and levels of annotation, issues of inter-annotator agreement, and representativeness of the collected data.

Keywords: corpus linguistics, evidentiality, Amerindian, Romance

References

Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford University Press.

Boye, K. (2012). *Epistemic Meaning, A Crosslinguistic and Functional-Cognitive Study*. De Gruyter Mouton.

Faller, M. T. (2002). *Semantics and Pragmatics of Evidentials in Cuzco Quechua*. Stanford University.

Mushin, I. (2013). Making knowledge visible in discourse: Implications for the study of linguistic evidentiality. *Discourse Studies*, *15*(5), 627–645. https://doi.org/10.1177/1461445613501447

Van Gijn, R. (2018). *Linguistic diversity and the South American perspective*. University of Zurich lecture. Retrieved November 25, 2019, from https://www.comparativelinguistics.uzh.ch/dam/jcr:82ccd60b-0531-42e8-949565ccd1476b82/Antrittsvorlesung_final.pdf

Carolina Julià Luna[1], Alba Aguete Cajiao[2], Soraya Almansa Ibáñez[1], Borja Alonso Pascua[2], Beatriz Blecua Falgueras[3], Joseph García Rodríguez[1], César Gutiérrez Miguel[4], M.ª Jesús Machuca Ayuso[5], Assumpció Rost Bagudanch[5] & Natalia Terrón Vinagre[5]

([1]Universidad Nacional de Educación a Distancia, [2]Universidad de Salamanca, [3]Universitat de Girona, [4]Wake Forest University, [5]Universitat Autònoma de Barcelona)

Avatares en el proceso de creación de un corpus geolectal para el estudio del español oral

Los atlas lingüísticos del español constituyen las primeras fuentes de datos que nos permiten documentar muestras de lengua oral recogidas mediante criterios científicos. En los inicios de la aplicación del método geolingüístico a las lenguas de la península ibérica, no todas las encuestas se llevaron a cabo con apoyo tecnológico de forma sistemática (González González 1992: 175); por ello, las transcripciones georreferenciadas de estas obras son las únicas informaciones que han permanecido de las entrevistas que se realizaron.

En el ámbito del español europeo, el análisis, la explotación y la difusión que se ha hecho de los datos de la geografía lingüística en su conjunto ha sido escaso debido a las limitaciones que ha impuesto la consulta y el formato de publicación de los atlas. La digitalización y sistematización de estos materiales es indiscutiblemente necesaria para poder paliar esta situación. En el caso del *Atlas Lingüístico de la Península Ibérica* (*ALPI*), este proceso se inició hace más de una década en el CSIC (García Mouton 2017, 2022). Sin embargo, en la geolingüística regional, para la que el español dispone de un importante número de atlas, se ha contado con escasas iniciativas de digitalización hasta la creación del *Corpus de los atlas lingüísticos* (*CORPAT*, www.corpat.es). Esta herramienta, de consulta libre en la red, categoriza y repertoriza las formas lingüísticas de los atlas regionales del español en una base de datos relacional georreferenciada.

Los objetivos de este corpus son, por un lado, poner a disposición de la comunidad científica los datos de los atlas regionales para que puedan ser estudiados desde múltiples perspectivas y, por otro, conservar y dar a conocer el patrimonio lingüístico de la España de la segunda mitad del siglo XX. Ello permitirá examinar con detalle los fenómenos desde una perspectiva tanto sincrónica como diacrónica, ya que, además de facilitar la obtención de una foto del perfil lingüístico de las zonas

de encuesta en el momento en el que se realizaron las entrevistas, también se podrá comparar la información con los datos del *ALPI* y de otros corpus orales más recientes para observar y analizar con detalle el cambio lingüístico a través del espacio.

Actualmente, este corpus se encuentra en un estadio inicial de desarrollo y es posible consultar parcialmente algunos de los datos que contiene. En el póster que se presenta, se expondrán las características del corpus (Julià 2023), la metodología (Rost *et al.* 2022), los problemas y dificultades que han surgido en las primeras fases de trabajo (Blecua *et al.* 2024), los primeros resultados (Julià 2022, Gutiérrez 2023, Terrón 2023) y las líneas futuras de investigación derivadas de su explotación como herramienta para el estudio de la lengua oral (Alonso 2023).

Palabras clave: geolingüística, dialectología, atlas lingüísticos, español

Bibliografía

Alonso, B. (2023). Las hablas de Salamanca en el continuo lingüístico noroccidental: una puesta al día. *Revista de Investigación Lingüística*, 26, 15-34. https://doi.org/10.6018/ril.555881

Blecua, B., Machuca, M.ª J., Aguete, A., Elvira, W., Garrido, J. M.ª, Julià, C., Marrero, V., Quijada, C., Roseano, P. y Rost, A. (2024). *Criterios para la conversión a AFI de los símbolos fonéticos en los atlas lingüísticos de España*. LII Simposio de la Sociedad Española de Lingüística, Centro de Ciencias Humanas y Sociales del CSIC, Madrid, 22-25 de enero de 2024. http://sel.edu.es/liisimposio-sel-2023/

García Mouton, P. (2017). El *Atlas Lingüístico de la Península Ibérica* (*ALPI*) en línea. Geolingüística a la carta. *Estudis romànics*, 39, 335-343.

García Mouton, P. (2022). El *Atlas Lingüístico de la Península Ibérica* (*ALPI*) de Tomás Navarro Tomás y nuestra geografía lingüística. En I. Molina Martos y P. García Mouton (Eds), *Geolingüística en la Península Ibérica* (pp. 33-53). Madrid, CSIC, Anejos de la Revista de Filología Española.

González González, M. (1992): «Metodología de los atlas lingüísticos en España». En G. Aurrekoetxea y X. Videgain (Eds.), *Nazioarteko dialektologia biltzarra. Agiriak = Actas del Congreso Internacional de Dialectología = Actes du Congrès international de dialectologie = Proceedings of International Congress on Dialectology (Bilbao, 1991)* (pp. 151-177), Bilbao, Euskaltzaindia.

Gutiérrez, C. (2023). Notas etimológicas sobre columpio y sus variantes en las lenguas de la Península Ibérica. *Cuadernos del Instituto Historia de la Lengua*, 16, 13-42.

Julià Luna, C. (2022). Geolingüística digital y bases de datos: una aproximación al estudio de la variación y el cambio léxico en español. *Revista Internacional de Lingüística Iberoamericana (RILI)*, 40/2, 13-31.

Julià Luna, C. (2023). Desarrollo de un corpus de atlas lingüísticos. In A. Grajales Ramírez, J. Mauricio Molina Mejía y P. Valdivia Martín (Eds.), *Digital Humanities, Corpus and Language Technology / Humanidades Digitales, Corpus y Tecnología del Lenguaje: A look from diverse case studies / Una*

mirada desde diversos casos de estudio (pp. 123-142), Groningen /Antioquia: University of Groningen Press / Universidad de Antioquia.

Rost, A., Blecua, B., Aguete, A., Elvira, W., Garrido, J. M.ª, Julià, C., Machuca, M.ª, Marrero, V., Quijada, C. y Roseano, P. (2023). *Una propuesta unificada basada en AFI para la transcripción fonética de los atlas regionales de España*. IX Congreso Internacional de Fonética Experimental, Universidad de Vigo (21-23 de junio de 2023). https://cife2023.webs.uvigo.es/es/presentacion/

Terrón Vinagre, N. (2023). Las designaciones de *llevar a cuestas* en la geolingüística regional. *Revista Internacional de Lingüística Iberoamericana (RILI),* XXI/1 (41), 167-189.

Christian Koch
(Universität Siegen)

Desafíos y estrategias en la creación de un corpus de español L2: ¿Cómo unificar transcripciones en un proyecto internacional?

El proyecto presentado en esta ponencia se centra en la creación de un corpus de español L2 hablado, con énfasis en el uso de marcadores del discurso en diferentes niveles de competencia lingüística, desde A2 hasta C1. Este corpus comprende narraciones de aprendices con diversas L1 (alemán, francés, inglés, italiano, neerlandés y ruso), con 40 grabaciones por cada L1. Además, el corpus se complementa con 40 grabaciones de hablantes nativos de español sobre el mismo tema: narrar una situación peligrosa (Payrató & Fitó, 2008). El equipo del proyecto consta de 6 subequipos, cada uno de los cuales trabaja en el contexto de un L1. Los subequipos crearon las partes respectivas del corpus y las transcribieron inicialmente utilizando los sistemas que les parecían más adecuados.

Con el objetivo de llevar a cabo análisis comparativos de la adquisición y uso de marcadores del discurso en diferentes niveles de español L2, planeamos reunir todas las transcripciones en un repositorio abierto como el *TalkBank*/*CHILDES*, utilizando el programa *CLAN* para el procesamiento de datos (MacWhinney, 2000). Al mismo tiempo, buscamos realizar una anotación funcional de los marcadores del discurso y de los procesos de elaboración discursiva.

Uno de los principales desafíos es la transformación de las transcripciones, provenientes de diferentes sistemas (por ejemplo, el GAT2 para la parte alemana, Ehmer et al., 2019, y CAP para la parte neerlandesa, Payrató & Fitó, 2008), al formato CHAT, lo que puede implicar la reducción o adición de detalles según los estándares de CHAT. En la práctica, surgen dificultades en varias partes del corpus que presentan diferentes necesidades de adaptación. Este proceso se ilustrará con ejemplos concretos, como los siguientes:

(1.a) *Versión GAT2*

L: y algún (.) como en diez minutos después °h
[ahm]: (-)
M: [hm]
L: llegaron y °hh

era como:: ehm: (.)

tre:s o (.) cuatro per[sonas] y °h

M: [hm_hm]

L: también ahm (.)

eran involvido otras ehm otros co[ches con otr]as

M: [ah otros coches]

L: personas

(1.b) *Versión CHAT*

*DT001: y algún [///] (.) como en diez minutos después h@fp <ahm@fp> [>] llegaron .

*INV:<hm> [<] .

*DT001: y h@fp era como:: ehm@fp (.) tre:s o cuatro per<sonas> [>] .

*INV:<hm_hm> [<] .

*DT001: y h@fp también ahm@fp (.) eran involvido [: involucrados] otras ehm@fp [//] otros co<ches con otr> [>] as personas .

*INV:<ah@i otros coches> [<] .

(2.a) *Versión CAP*

21. o yo yo no mmm
22. sepa/
23. no, no es eso
24. yo no sabía? dónde era/
25. {(AC)Yo no sabía dónde era el sol}
26. o el el el aire
27. es que yo
28. yo:_
29. yo me me encon:tré
30. completamente desorientada/

(2.b) *Versión CHAT*

*DUT46: o yo [/] yo no, ahm@fp, (.) sepa?

*DUT46: no, no es eso.

*DUT46: yo no (.) sabía?

*DUT46: dónde era.

*DUT46: yo no sabía.

*DUT46: dónde era el sol.

*DUT46: o el [/] el [/] el aire.

*DUT46: es_que@i yo [/] yo: [///], ahm@fp, yo me [/] me encon::tré completamente desorientada.

Como se puede observar, los sistemas de transcripción difieren no solo en términos de simbología —por ejemplo, la indicación con "@fp" ('*filled pause*')—, sino que CHAT también hace mayor hincapié en marcar los procesos de verbalización —como el uso de "[///]" para señalar reformulaciones— y en la identificación de desviaciones de la norma. Estas marcas pueden ser especialmente interesantes en la transcripción del lenguaje de los aprendices, pero también requieren un mayor grado de interpretación como parte del procesamiento de los datos. Basándonos en ejemplos concretos del proceso de transformación, discutiremos los problemas y cuestiones que aún están abiertos en este aspecto del proyecto, destacando así la importancia de abordar los desafíos técnicos y estrategias en la creación de un corpus de español L2.

Palabras clave: transcription systems; Spanish; language acquisition; discourse markers

Bibliografía

Ehmer, O. et al. (2019). Un sistema para transcribir el habla en la interacción: GAT 2. In: *Gesprächsforschung* 20, 64–114.

MacWhinney, B. (2000). *The CHILDES Project: tools for analyzing talk. Transcription format and programs*. Lawrence Erlbaum Associates.

Payrató, L. & Fitó, J. (2008). *Corpus audiovisual plurilingüe*. Universitat de Barcelona.

Svenja Krieger
(Universität Konstanz)

When pragmatics meets prosody: the use of *alors*, *bon*, and *donc* in heritage speakers of French

Linguistic phenomena at the interfaces are often considered to be vulnerable to language influence (CLI) in heritage language research, especially if pragmatics is involved (Sorace, 2004). A large number of studies have shown such a vulnerability for the syntax-pragmatics interfaces (see e.g., Casalicchio & Moroni, 2023 on the positioning of discourse markers, i.e., words that are used on a discursive level instead of a referential or textual level, Bayer, 2009). Compared to this interface, the pragmatics-prosody interface, in particular the prosodic realization of discourse markers in heritage speakers (HSs) has received considerably less attention.

The present study aims at filling this gap by focusing on the pragmatic meaning of the French discourse marker *alors* ('so'), *bon* ('well'), and *donc* ('thus') as well as their prosodic properties. In French, the discourse marker *alors* is most often used to express introduction, additional information, and conclusion, *bon* usually expresses additional information, and *donc* additional information and conclusion (e.g., Lee et al., 2019). Concerning the prosodic characteristics, *alors* is mostly produced with a pause right before it regardless of the pragmatic function. The same holds for the *donc* with a conclusive meaning. By contrast, *bon* and *donc* introducing additional information usually occur without pauses (Lee et al., 2020). Lee et al. (2020) found that French discourse markers are mostly produced with medium Fo level and a plateau, especially if they function as an introduction marker. By contrast, English (as well as German) usually show a rising-high-falling Fo (Lee et al., 2020). This difference between French and German make discourse markers with the function of an introduction potentially susceptible to CLI.

In this study, I examine: i) which pragmatic meanings *alors*, *bon*, and *donc* encode in the speech of HSs and ii) which prosodic properties HSs use in their production of discourse markers. For this purpose, a study on the corpus HABLA ("Hamburg Adult Bilingual Language" Kupisch et al., 2012) is carried out. This corpus includes semi-structured interviews with German-French

HSs. The corpus contains 12 interviews with HSs who grew up in Germany and consequently French as their heritage language (HL) and 9 interviews with HSs who grew up in France with French as their majority language (ML). Preliminary results

of five HSs with French as their HL reveal that HSs use *alors*, *bon*, and *donc* to express the same pragmatic functions as French monolinguals (see Table 1 and compare Lee et al., 2020). A qualitative analysis of HSs' prosody revealed that HSs mostly produce the same kind of pauses as reported for French monolinguals. Even though HSs mostly use the medium-plateau contour, a closer analysis of the contours at the individual level show that at least two speakers transfer the contour from German, suggesting CLI from the ML into the HL. Therefore, the preliminary results only confirm partially the vulnerability of the pragmatics-prosody interface in HL acquisition.

In the presentation, I will provide results for all HSs including a detailed analysis of the pragmatic functions and prosodic characteristics of the three discourse markers.

	alors	*bon*	*donc*
Total number	**30**	**64**	**67**
Introduction	14 (46%)	2 (3%)	1 (1%)
Additional information	6 (20%)	59 (92%)	33 (50%)
Conclusion	8 (27%)	3 (5%)	29 (43%)
Other	2 (7%)	0	4 (6%)

Table 1: Number of occurrences of discourse markers, and percentage of associated pragmatic function.

Keywords: heritage speakers, pragmatics, prosody, interfaces

References

Bayer, J. (2009). Discourse particles in questions. In: Proceedings of GLOW Asia.

Casalicchio, J., & Moroni, M. C. (2023). The Syntax–Pragmatics Interface in Heritage Languages: The Use of anche ("Also") in German Heritage Speakers of Italian. *Languages*, *8*(2), 104.

Kupisch, T., Barton, D., Bianchi, G., & Stangen, I. (2012). The HABLA-Corpus (German-French and German-Italian). *Multilingual corpora and multilingual corpus analysis*, 163179.

Lee, L., Bartkova, K., Jouvet, D., Dargnat, M., & Keromnes, Y. (2019). Can prosody meet pragmatics? Case of discourse particles in French. In: *ICPhS 2019-International Congress of Phonetic Sciences*.

Lee, L., Jouvet, D., Bartkova, K., Keromnes, Y., & Dargnat, M. (2020). Correlation between prosody and pragmatics: case study of discourse markers in French and English. In: *INTERSPEECH 2020*.

Sorace, A. (2004). Native language attrition and developmental instability at the syntaxdiscourse interface: Data, interpretations and methods. *Bilingualism: Language and cognition*, *7*(2), 143-145.

Ana Llopis Cardona
(Universitat de València)

Hacia un corpus de películas y series de televisión: el *Corpus oral de coloquialidad fingida*

En estudios recientes se ha reconocido que la representación del lenguaje coloquial es bastante fiable en películas y series de televisión (Jucker & Landert 2023; Tonetti & Landert 2023); no obstante, en la actualidad solo existe un corpus de películas y series de televisión en lengua inglesa (*TV Corpus/Movies Corpus* de Mark Davies, 2021). El español no cuenta con un corpus similar ni los grandes corpus actuales han incluido este tipo de material (CORDE; CREA; CdH; CdE, CORPES XXI), al igual que sucede en el resto de lenguas románicas. Ante este vacío, surge el *Corpus oral de coloquialidad fingida* (COCF), compuesto por una selección de películas y series de televisión desde los años cincuenta del siglo pasado hasta la actualidad. Este corpus se inició en 2020 con el proyecto "Difusión del cambio lingüístico en el español coloquial durante los últimos cincuenta años" (GV/2020/157) y se está desarrollando en la actualidad gracias a la Beca Leonardo a Investigadores y Creadores Culturales 2023 de la Fundación BBVA (LEO23-2-10560). Esta comunicación se propone comentar aspectos cruciales del proceso de elaboración de este corpus, tales como la selección de las obras y los metadatos registrados de las obras y de los personajes. En cuanto a la selección de las obras, se escogieron películas y series de televisión que se ajustaran a los rasgos coloquializadores y rasgos de registro coloquial en la medida de lo posible (Briz 1998), a saber: dramas sociales del neorrealismo español, comedias sociales, cine quinqui, etc. En la comunicación se ofrecerán datos generales y muestras representativas para ilustrar los aspectos anteriores, y se disertará sobre las dificultades surgidas en los metadatos y las soluciones que se han ido tomando para sortearlas. Finalmente, se esbozarán posibles aplicaciones de este corpus, que permitirá indagar fenómenos lingüísticos y pragmáticos desde la segunda mitad del siglo XX hasta la actualidad (cf. Llopis y Pons 2024; Llopis y Jansegers 2024), así como explorar los diálogos del guion cinematográfico y contrastar la simulación del registro coloquial con conversaciones coloquiales reales.

Bibliografía

Briz Gómez, Antonio. 1998. *El español coloquial en la conversación. Esbozo de pragmagramática*. Barcelona: Ariel.

Davies, Mark (2002-2021). *Corpus del español.* https://www.corpusdelespanol.org/

Davies, Mark (2019). *TV and Movie Corpora.* https://www.english-corpora.org/iweb/

Instituto de Investigación Rafael Lapesa de la Real Academia Española, *Corpus del Nuevo diccionario histórico*, <https://apps.rae.es/CNDHE/view>

Jucker, Andreas H. & Daniela Landert. 2023. The diachrony of im/politeness in American an British movies (1930–2019). *Journal of Pragmatics* 209. 123–141.

Llopis Cardona, Ana & Marlies Jansegers. 2024. Del *rollo* como sustantivo comodín contracultural al *rollo* aproximador en el español coloquial actual. *Spanish in Context 21.1.*

Llopis Cardona, Ana & Salvador Pons Bordería. 2024. Fases y factores socioculturales en la difusión de *tío/tía* como vocativos: "juvenilización" del español coloquial actual. *Spanish in Context 21.1.*

Real Academia Española. Banco de datos (CORDE) [en línea]. *Corpus diacrónico del español*, http://www.rae.es

Real Academia Española. Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual*, <http://www.rae.es>

Real Academia Española: Banco de datos (CORPES XXI) [en línea]. *Corpus del Español del Siglo XXI (CORPES).* http://www.rae.es

Tonetti Tübben, Ilenia & Daniela Landert. 2023. Uh and Um as Pragmatic Markers in Dialogues: A Contrastive Perspective on the Functions of Planners in Fiction and Conversation. *Contrastive Pragmatics* 4.2: 350–381.

Anna De Marco[1], **Stefania Ferrari**[2], **Samuele Giordano**[2] & **Marianna Ceravolo**[1]
([1]Università della Calabria, [2]Università del Piemonte Orientale)

Developing a pragmatically annotated task-based corpus for spoken Italian

In the field of second language teaching there is an increasing emphasis on the pragmatic dimension of spoken communication and the development of interactional language skills. Despite its central role in language learning, pragmatics struggles to be systematically integrated into language teaching and assessment. One reason for this is the lack of resources that are representative of real language use. Corpus linguistics could potentially provide practical tools for accessing searchable authentic and pragmatically relevant spoken input (Bardovi-Harlig, Mossman 2023; Romero-Trillo 2008). Nevertheless, the interaction between corpus linguistics and pragmatics is still under-researched, especially for Italian. This is partly due to a number of methodological challenges, especially with regard to the corpus collection unit and the labelling and annotation criteria (Rühlemann, Aijmer 2015). If the aim is to explore the pragmatic dimension of communication, it is important to be able to study interaction by considering the communicative event as a whole, thus referring to context as determined by both external contextual variables, and internal organisational sequences of task-based interaction.

The present work aims at discussing the development process of a first prototype of a pragmatically annotated task-based corpus of spoken Italian. Starting from a need analysis of adult immigrants and international students as target learners, a first set of 100 spontaneous and pedagogical spoken task-based interaction samples was collected. The two tasks considered are "ordering in a restaurant" and "shopping in a small store". The data were transcribed and annotated using the PraTiD system (Savy, Castagneto 2009), a hierarchically organised multilevel annotation system that takes into account the different levels of sequential organisation of the interaction. The combination of task-based corpora and pragmatic annotation systems should facilitate the balance between form-based and function-based approaches, allowing a corpus interrogation starting from communicative functions and speech acts realization within tasks, and then exploring the variety of forms and interactional strategies through which these functions are realised.

Keywords: pragmatics, spoken interaction, corpora, TBLT, Italian L1, L2,

References

Bardovi-Harlig K., Mossman S. (2023), "Corpora in instructed second language pragmatics", in Jablonkai R.R., Csomay E. (eds.), *The Routledge Handbook of Corpora and English Language Teaching and Learning*, Routledge, New York-London, pp. 71-88.

Romero-Trillo J. (2008), *Corpus Linguistics and Pragmatics: A Mutualistic Entente,* Walter de Gruyter, Berlin.

Rühlemann C., Aijmer K. (2015), "Corpus pragmatics: Laying the fundations", in Aijmer K., Rühlemann C. (eds.), *Corpus Pragmatics: A Handbook*, Cambridge University Press, Cambridge, pp. 1-26

Savy R., Castagneto M. (2009), "Funzioni comunicative e categorie d'analisi pragmatica: dal testo dialogico allo schema xml e viceversa", in G. Ferrari, R. Benattu, M. Mosca (a cura di), *Linguistica e Modelli tecnologici di Ricerca*. *Atti del XL Congresso SLI,* Roma, Bulzoni, pp. 569-579.

Caterina Mauri, **Silvia Ballarè**, **Beatrice Beransconi**, **Massimo Cerruti**, **Eugenio Goria** & **Eleonora Zucchini**

(Università di Bologna)

The KIParla corpus of Spoken Italian

The aim of this paper is to describe the KIParla corpus (www.kiparla.it), i.e. a new, freely accessible resource for the study of spoken Italian characterized by (i) access to metadata on interactions and speakers, (ii) orthographic and conversational transcriptions aligned with audio files, (iii) NoSketch Engine search interface. The KIParla corpus consists of data collected in different situations and involving speakers with different socio-geographical characteristics. It is designed as a modular and expandable resource: each module focuses on different dimensions of sociolinguistic variation, and new modules can be added over time. Such a dynamic nature makes the KIParla corpus suitable to document spoken language over time.

After describing the methodology employed during the phases of data collection and resource construction, we will provide the details of the actual KIParla modules, together with examples of the types of phenomena that can be observed in the data (Mauri et al. 2019; Ballarè et al. 2022):

1) KIP - educated speakers living in Bologna and Turin, variety of ages, and kinds of interaction (recorded at university: lessons, free conversations, exams, office hours, semi-structured interviews): ca. 661.000 tokens, 69 hrs, 150 speakers involved. Published.
2) ParlaTO – semi-structured interviews collected in Turin, speakers with varying age, educational degree and employment: ca. 550.000 tokens, 50 hrs, 88 speakers involved. Published.
3) ParlaBO - semi-structured interviews collected in Bologna, speakers with varying age, educational degree and employment: ca. 630.000 tokens, 66 hrs, 158 speakers involved. Completed, will be published in June 2024.
4) KIPasti – dinner-table conversations collected in 13 regions of Italy, speakers with varying age, educational degree and employment: ca. 490.000 tokens, 43 hrs, 149 speakers involved. Published.
5) Stra-ParlaBO and Stra-ParlaTO – semi-structured interviews and dinner-table conversations, speakers with foreign-origin with complex multilingual repertoires. Data collected in Bologna and Turin. Under construction: expected 128 hrs. Planned publication in June 2025.

Once the development of the last modules will be completed, the KIParla corpus will consist of ca. 350 hrs of spoken interactions (ca. 3.500.000 tokens), yielding a particularly rich and detailed view of existing variation in contemporary spoken Italian, destined to grow over time. Thanks to the integration with the other modules of the KIParla corpus, foreign-origin speakers are considered on a par with "native" speakers, in the belief that very concept of "native speaker" is inherently blurred and disaggregated (Berruto 2003; Dewaele 2018).

For all the modules, the metadata concerning both speakers and interactions are accessible and available as search filters. This makes it possible to analyze each module individually, e.g. focusing on specific varieties, or jointly, e.g. selecting spontaneous interactions, including speakers under-40 and with a low level of educational achievement - independently of their being native or non-native speakers of Italian.

We will conclude by highlighting the problems, and the solutions we found, regarding (i) data transcription of non-standard variants, (ii) GDPR and audio sharing, (iii) project sustainability over time.

References

Ballarè, S., Goria, E., Mauri, C. (2022). Italiano parlato e variazione lingusitica. Teoria e prassi nella costruzione del corpus KIParla. Bologna: Pàtron.

Berruto, G. (2003) 'Sul parlante nativo (di italiano)', in H.I. Radatz and R. Schlosser (eds) Donum grammaticorum. Festschrift fur Harro Stammerjohann. Tubingen: Niemeyer, pp. 1–14.

Dewaele, J.-M. (2018). Why the dichotomy 'L1 versus LX user' is better than 'native versus non-native speaker'. Appl. Linguis. 39, 236–240.

Mauri, C., Ballarè, S., Goria, E., Cerruti, M., & Suriano, F. (2019). KIParla Corpus: A New Resource for Spoken Italian. Proceedings of the Sixth Italian Conference on Computational Linguistics. Bari, Italy, November 13-15, 2019

Johanna Miecznikowski, **Elena Battaglia**, **Christian Geddo**, **Nina Profazi**
(Università della Svizzera Italiana)

TIGR: Towards a FAIR audio/video corpus of spoken Italian

The TIGR corpus of spoken Italian was gathered in the Swiss cantons Ticino and Grisons in 2021-2022 and is currently being prepared for sharing via the repository LaRS @ SWISSUbase. We present (a) the corpus design; (b) workflows of data processing in view of FAIR data sharing (Wilkinson et al. 2016); (c) a lab blog dedicated to these topics.

TIGR is a special corpus (Teubert/Čermáková 2004:119) that was collected within a research project focused on epistemic aspects of talk (InfinIta, SNSF grant no. 192771). At the same time, it was designed to increase the diversity of available resources for spoken Italian (for an overview see Mauri et al. 2019) in the perspective of "opportunistic" corpus re-use (cf. Teubert/Čermáková 2004:120). It includes 23.5h of video recordings documenting 23 face-to-face interactions. These vary as to genre and as to external criteria (Sinclair & Ball 1996), more specifically event-related parameters (Deppermann/Hartung 2011:423-424) such as institutionality, the number of participants, speaker roles and the presence of multi-activity (Mondada 2009): table conversations (6h5'), food preparation (1h40'), tutoring encounters (4h40'), lessons and practical instruction (7h20'), interviews (3h40'). The 115 speakers are 10-70 years old (most represented range: 20-29 years) and about 3/4 of them finished a higher secondary school. They declared their consent to data use and re-use for scientific purposes and expressed some de-identification demands. The technical set-up included two camcorders and 2-4 pocket audio recorders equipped with clip-on microphones, all synchronized through timecode generators. The A/V files were aligned and cut to equal length in Adobe Premiere. The team then transcribed them in ELAN (Sloetjes/Seibert 2016) using an adapted version of the GAT 2 conventions (Selting et al. 2011). Proper names were pseudonymized.

To enhance interoperability and reusability, we plan to provide two transcript versions in addition to the EAF file generated by ELAN. By now, we have implemented a script-assisted workflow to produce TXT transcripts that are optimized for the human eye and preserve a reduced amount of timecode stamps. Later we intend to create tokenized XML transcripts readable by corpus linguistic software. In A/V files we will mask faces and voices, where so required, and replace proper names by noise.

For each event, we will edit a single compact, easy-to-use movie file with split screen and mixed audio.

Once ready, the corpus will be uploaded to LaRS, completed by metadata and documentation. Users could have the following download options: event by event, either a full version (A/V files, compact movie, EAF file, transcripts) or a light version (compact movie, transcripts); transcripts only for all events at once, TXT or XML. A desirable further step is to make the corpus accessible on a platform that allows for on-line use; yet, suitable infrastructure still needs to be developed in Switzerland.

While preparing the data, we are step by step building a webpage to present the corpus. In parallel, we use a lab blog (sharetigr.usi.ch) to publicly report on our experience and discuss issues we are facing, thus developing a case study of open research data practices in linguistics.

References

Mauri, C., Ballarè, S., Goria, E., Cerruti, M., & Suriano, F. (2019). KIParla corpus: A new resource for spoken Italian. In R. Bernardi, R. Navigli & G. Semeraro (eds.), *Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it,* CEUR-WS.

Mondada, L. (2009). Multimodalità e multi-attività nelle conversazioni a tavola. In M. Fatigante, L. Mariottini & M. E. Sciubba (Eds.), *Lingua e società. Scritti in onore di Franca Orletti* (pp. 88-106). Bologna, Il Mulino.

Selting, M., et al. (2011). A system for transcribing talk-in-interaction: GAT 2 translated and adapted for English by Elizabeth Couper-Kuhlen and Dagmar Barth-Weingarten. *Gesprächsforschung, 12*, 1-51.

Sinclair, J., & Ball, J. (1996). *EAGLES Text typology. https://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html*

Sloetjes, H., & Seibert, O. (2016). Measuring by marking; the multimedia annotation tool ELAN. In A. J. Spink, G. Riedel, L. Zhou, L. Teekens, R. Albatal & C. Gurrin (Eds.), *Measuring Behavior 2016, 10th International Conference on Methods and Techniques in Behavioral Research* (pp. 492-495). Dublin, Dublin City University.

Teubert, W., & Čermáková, A. (2004). Directions in corpus linguistics. In M. A. K. Halliday, W. Teubert, C. Yallop & A. Cermáková (Eds.), *Lexicology and Corpus Linguistics. An Introduction* (pp. 113–165). London/New York, Continuum.

Wilkinson, M. D. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*(1), 1-9. 10.1038/sdata.2016.18.

https://sharetigr.usi.ch/en/news-events/blog: Lab blog.

Language Repository of Switzerland LaRS: https://www.lars.uzh.ch/en.html

The categorization of information sources in face-to-face interaction: a study based on the TIGR-corpus of spoken Italian (short title: InfinIta).

https://data.snf.ch/grants/grant/192771

Oana Niculescu
(Romanian Academy Institute of Linguistics "Iorgu Iordan – Al. Rosetti")

Pause annotation schema in an emerging spoken Romanian corpus

The aim of this presentation is to report on a currently under development Romanian speech corpus, with a specific emphasis on pausal contexts annotation. The presentation showcases the stages of corpus recording and data annotation in relation to silent and breath pauses.

In comparison to other Romance languages, Romanian speech corpora are still scarce (Mîrzea-Vasile 2017), especially those providing time-aligned data. Since most conversational corpora are distributed in text version only (CORV, IVLRA, ROVA, a.o.), including the recently developed multimodal CLRV corpus, accessing the audio and video recordings proves rather challenging.

In response to these challenges, this presentation showcases the development of a monologue speech corpus for Romanian, recorded and manually aligned using Praat. Our project aims to address the shortage of linguistic resources for Romanian by providing open-access corpus for research purposes, thus licensing an in-depth examination of language variation. The goals of this project are in accordance with current national research initiatives focused either on developing resources and tools for natural language processing or building learner corpora, as in the case of the LECOR project (Barbu et al. 2023).

The corpus consists of ten monologues pertaining to native monolingual adult speakers without any speaking or hearing impairments. All participants (5 male, 5 female) are representative of the Southern dialect on which the standard language is based on. The monologues share the same three main conversational topics related past (memories from childhood, life lessons), present (activities, likes and dislikes, travelling), and future endeavours (personal and professional). Participants had the option of moving freely from one topic to another, always addressing the same female researcher performing the recordings which were carried out in a sound attenuated room. Speakers were highly invested in the task, producing monologues of up to 70 minutes. The data were transcribed and manually aligned in Praat via TextGrids.

In terms of annotating speech pauses, firstly, we distinguished between articulatory pauses (Hieke et al. 1983), typically occurring in the closure phases of voiceless stops and thus excluded from transcription, and pauses produced in connected speech,

the focus of our transcription. Secondly, our annotation schema includes a separate category for *silent pauses* (Figure 1) in a phonetic sense, which do not contain any phonetic particles or events (Belz & Trouvain 2019), and *breath pauses* (Figure 2), typically containing inhalation and exhalation noises (Trouvain & Werner 2022; Werner 2023). Moreover, *pause-internal particles* (Werner 2023) such as clicks (Ogden 2013), filler particles (Belz 2023) and vegetative sounds (e.g., swallowing, coughing, sneezing etc.) were also included in the annotation as separate categories alongside other nonverbal vocalisations (Trouvain 2014) like onomatopoeia or laughter.

Our preliminary results derived from a corpus-based analysis of a 53 min monologue pertaining to a male speaker reveal that silent pauses account for 47% (N = 381) of the total pausal contexts, while breath pauses amount to 53% (N = 425) of the data. Among breath pauses, inbreaths have the highest frequency of occurrence (89%). In terms of pause-internal particles, our data show that swallowing episodes occur in 62% of the cases, vastly outnumbering filler particles (24%), clicks (13%) and throat clearing episodes (1%).

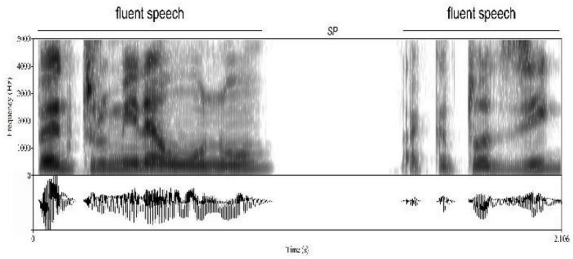

Figure 1. Silent pause between two stretches of fluent speech
Broadband spectrogram and waveform of the utterance
pentru mine unu(l) # (0.497) dacă tot zic
"for me <silent pause> if I say"

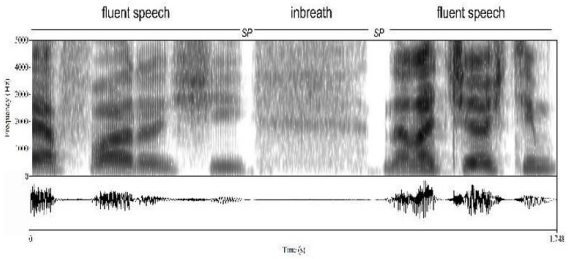

Figure 2. Audible breath intake between two stretches of fluent speech
Broadband spectrogram and waveform of the utterance
de afară şi⊥ # (0.036) ⇓ (0.382) # (0.055) da(r)-(î)n acelaşi timp
"from outside and <abandoned phrase> <silent pause> <audible inbreath> <silent pause> but at the same time"

Keywords: Romanian speech corpus, annotation schema, silence, breath pause

References

Barbu, A. M., Irimia, E., Mîrzea Vasile, C., & Păiș, V. (2023). Designing the LECOR Learner Corpus for Romanian. in G. Angelova, M. Kunilovskaya & Ruslan Mitkov (eds.), *Deep Learning for Natural Language Processing Methods and Applications (Proceedings of International Conference Recent Advances in Natural Language Processing, RANLP 2023, Varna, 4–6 September, 2023)*, 143-152.

Belz, M. (2023). Defining filler particles: A phonetic account of the terminology, form, and grammatical classification of "filled pauses". *Languages* 8.1: 57.

Belz, M., & Trouvain, J. (2019). Are 'silent' pauses always silent?. *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 2744-2748.

CORV = Dascălu Jinga, L. (2002). *Corpus de română vorbită. Eșantioane* [Corpus of Spoken Romanian. Samples]. București, Oscar Print.

Hieke, A. E., Kowal, S.. & O'Connell, D. C. (1983). The trouble with 'articulatory' pauses. *Language and Speech*, 26, 203-214.

IVLRA = Ionescu-Ruxăndoiu, L. (2002). *Interacțiunea verbală în limba română actuală. Corpus (selectiv). Schiță de tipologie* [Verbal Interaction in Present Day Romanian. (Selective) Corpus. Outline of a Typology]. București, Editura Universității din București.

Mîrzea-Vasile, C. (2017). Corpusurile de limba română și importanța lor în realizarea de materiale didactice pentru limba română ca limbă străină [Romanian corpora and their importance in developing didactic resources for Romanian as a foreign language]. *Romanian Studies Today*, I, 74-95.

Ogden, R. (2013). Clicks and percussives in English conversation. *Journal of the International Phonetic Association*, 43(3), 299-320.

ROVA = Dascălu Jinga, L. (2011). *Româna vorbită actuală (ROVA): corpus și studii* [Present-Day Spoken Romanian: Corpus and Studies]. Bucureşti, Editura Academiei.

Trandabat, D., Irimia, E., Mititelu, V., Cristea, D., & Tufis, D. (2012). *The Romanian Language in the Digital Age*. META-NET White Paper Studies, Springer.

Trouvain, J. (2014). Laughing, breathing, clicking - The prosody of nonverbal vocalisations. *Proceedings of Speech Prosody*, Dublin, Ireland, 598-602.

Trouvain, J., & Werner, R. (2022). A phonetic view on annotating speech pauses and pause-internal phonetic particles. *Transkription und Annotation Gesprochener Sprache und Multimodaler Interaktion: Konzepte, Probleme, Lösungen*, 64, 55-73.

Werner, R. (2023) *The phonetics of speech breathing: pauses, physiology, acoustics, and perception*. PhD Thesis, der Universität des Saarlandes.

Andrea Pešková
(Freie Universität Berlin)

Phonology Corpus for L2 Italian: Research on Geminate Consonants

This paper introduces the main features of the phonology corpus for L2 Italian of an ongoing DFG-funded project (2024-2027), serving as a new resource for studying the acquisition of geminate consonants. Although "long consonants" in Italian play a crucial role in distinguishing meanings (e.g., nono 'ninth' vs. nonno 'grandpa'), mastering their pronunciation poses challenges for learners (e.g., Sorianello 2004; Altmann et al 2012). By comparing languages with and without phonological length contrast, this study seeks to uncover the sources of difficulties in both geminate production and perception among 20 Czech, 20 Finnish, 20 German, and 20 Spanish adult learners of Italian. The project aims to document the pronunciation of Italian as foreign language and to contribute to second language acquisition research in general (e.g., Piske et al. 2001; Colantoni et al. 2015; Derwing & Munro 2015; Flege & Bohn 2021). Additionally, it also bridges the gap between linguistic research and language teaching by integrating empirical findings into language classrooms.

At this stage, the project is running the first data collection. Hence, in the workshop, the contribution addresses theoretical background, research questions, and methods utilized in building the L2 phonology corpus. It will discuss the advantages and disadvantages of experimental corpus phonology and compare its data with other spoken corpora.

The project includes both production and perception experiments. The production experiment consists of semi-spontaneous and controlled tasks. Semi-spontaneous data collection involves conversational interviews on predefined topics to test learners' phonological awareness. Controlled tasks include repetition and reading tasks featuring Italian geminates and singletons. The experiment encompasses real-word and pseudo-word stimuli strategically positioned to assess geminate perception and production. The main benefits of controlled tasks over spontaneous data for phonological research will be discussed in greater detail during the workshop. The perception experiment, comprising discrimination and identification tests, investigates learners' ability to differentiate between singleton and geminate

consonants. Prerecorded stimuli, manipulated in duration, together with measuring reaction times aid in determining learners' perceptual thresholds.

In sum, the study aims to contribute to the understanding of L2 Italian acquisition, shedding light on geminate consonant production and perception. During the workshop, presenting and discussing the methodology will be essential.

References

Altmann, H.; Berger, I. & B. Braun. 2012. Asymmetries in the perception of non-native consonantal and vocalic contrasts. Second Language Research 28(4), 387–413.

Colantoni, L.; Steele, J. & P. Escudero. 2015. Second language speech. Theory and practice. Cambridge: CUP.

Derwing, T. & M. Munro. 2015. Pronunciation Fundamentals. Evidence-based perspectives for L2 Teaching and Research. Amsterdam/Philadelphia: John Benjamins.

Flege, J. E. & O-S. Bohn. 2021. The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.): Second Language Speech Learning: Theoretical and Empirical Progress. Cambridge: CUP, 3–83.

Piske, T.; MacKay, I. R. A. & J. E. Flege. 2001. Factors affecting degree of foreign accent in an L2: A Review. Journal of Phonetics 29, 191–215.

Sorianello, P. 2014. Italian geminate consonants in L2 acquisition. In L. Costamagna & C. Celata (Eds.): Consonant gemination in first and second language acquisition. Perugia: Pacini Editore SpA, 25–46.

Pekka Posio
(University of Helsinki)

CoLaGe: Corpus for the Study of Language and Gender in two varieties of spoken Spanish

Language and gender is a well-established field of research in English linguistics, and there is a wide variety of theoretical and methodological approaches acknowledging the diversity, complexity and performativity of the notion of "gender" in this field. However, in quantitative corpus linguistics, speakers' gender is still typically operationalized as a binary, discrete explanatory variable. This is particularly true of research focusing on the Spanish-speaking world. The research project *Gender, society, and language use: Evidence from Mexico and Spain* (Posio, Kachel, Uclés-Ramada, Carcelén-Guerrero, 2019-2025), funded by the Kone Foundation, aims precisely at discussing and problematizing the notion of gender in corpus linguistics and quantitative sociolinguistics, and also providing new tools to analyse it.

Within the project, we have developed the *CoLaGe Corpus for the Study of Language and Gender in Spanish,* an oral bidialectal corpus collected in Guadalajara (Mexico) and Valencia (Spain). It consists of three subcorpora, one from Valencia (*CoLaGe-V*) and two from Guadalajara (*CoLaGe-G* and *CoLaGe-GD*) each containing different types of linguistic data: sociolinguistic interviews, roleplays, and picture description tasks to elicit phonetic data. The informants are divided into two age groups (30–40 and 60–70) and two genders, except for the *CoLaGe-GD* subcorpus where the informants are members of gender and/or sexual minorities. By standardizing the data collection procedures and the socioeconomic background of the informants, we aim at creating a database permitting sociolinguistic comparisons across dialects, age groups, and genders. In total, the *CoLaGe* contains data from 127 informants and will be published in the beginning of 2025.

In addition to general sociolinguist metadata, we have collected extensive social psychological data from the informants, in particular regarding their self-concepts, values and attitudes towards gender and sexuality. For instance, we have studied the informants' gender role self-concepts using different social-psychological surveys such as the Traditional Masculinity-Femininity Scale (Kachel et al. 2016). This allows using scalar variables such as the informants' self-assessed gender role self-concept or their conformity with traditional gender roles alongside their societal gender. The

research project thus aims at renewing the methodology of gender and language research by combining sociolinguistic and social psychological methods and theories.

In my talk, I will present the *CoLaGe* corpus focusing on both theoretical and practical aspects of the project. I will also discuss some case studies we have realized with the corpus data so far. In a pilot study (Posio, Kachel, Ucles-Ramada, 2024) we examined stereotypical beliefs about the use of certain morphosyntactic features by women, men, lesbian women and gay men in Mexico and Spain. Using the role play data simulating conflictive situations, we have analyzed gendered variation the amount of talk and interactivity in conflictive situations. We are also using the sociolinguistic interview data to study variation in the use of different grammatical persons in speech, such as the first-person singular which has been claimed to be more frequently used by women than men. These studies highlight the importance of the context and societal expectations towards speakers depending on their gender.

References

Kachel, Sven; Steffens, Melanie C.; Niedlich, Claudia. (2016). Traditional masculinity and femininity:Validation of a new scale assessing gender roles. *Frontiers in Psychology* 7(956). Available at: https://doi.org/10.3389/fpsyg.2016.00956

Posio, Pekka (PI); Kachel, Sven; Uclés-Ramada, Gloria; Carcelén-Guerrero, Andrea. (2019-2025). *Gender, society, and language use: Evidence from Mexico and Spain.* Research project funded by the Kone Foundation.

Posio, Pekka; Kachel, Sven; Uclés-Ramada, Gloria. (2024). Morphosyntactic stereotypes of speakers with different genders and sexual orientations: an experimental investigation. *Linguistics* (published online on May 3 2024). https://doi.org/10.1515/ling-2022-0143.

Leticia Rebollo Couto (Universidade Federal de Río de Janeiro, Brasil) &
Albert Rilliard (Université Paris Saclay, CNRS, LISN, Francia)

Variación pragmática y expresividad negativa: análisis multimodal en datos de doblaje

El doblaje es una modalidad de la traducción audiovisual que nos permite observar la dimensión multimodal de actos de habla, considerando estrategias verbales (lingüísticas) y no verbales (auditivas y visuales) en la realización de actos de habla expresivos. La expresividad negativa, en críticas y desacuerdos, se puede presentar intensificada o atenuada a través de diferentes estrategias visuales, auditivas, léxicas y discursivas. En este trabajo nos proponemos analizar comparativamente las estrategias de intensificación o atenuación de la expresividad negativa (en actos descorteses de críticas y desacuerdos) a partir de un corpus de animación, doblado en tres variedades del español (argentino, mexicano y español). Tres de estas películas son originales en inglés americano, *The Incredibles* (2004), *Cars* (2006) y *Ratatouille* (2007), producidas por Disney/Pixar y dobladas en estas tres variedades del español. La cuarta es una producción argentina *Metegol* (2013), doblada por *Universal Pictures* en su distribución internacional, al español mexicano (latino) para su distribución americana, al español español para su distribución europea y doblada al inglés inglés, por la localización de la temática de fútbol, más popular en Inglaterra que en Estados Unidos.

A partir de estas cuatro animaciones y sus respectivos doblajes nos proponemos comparar las estrategias de intensificación y atenuación de conflictos que se utilizaron en la traducción audiovisual, considerando elementos de variación pragmática desde una perspectiva multimodal. Desde una perspectiva verbal observamos estrategias convergentes y divergentes entre las tres variedades del español referentes a variación morfosintáctica, por un lado: uso de formas de tratamiento verbo pronominales, uso de pronombres clíticos átonos de tercera persona y selección de tiempos verbales o a variación léxico conversacional, por otro lado: uso de diminutivos, léxico coloquial indexador, marcadores del discurso, formas de tratamiento nominales, ajustes articulatorios, construcciones de mitigación o de intensificación.

Los datos de doblaje se caracterizan por la oralización de un texto escrito, dialogado, que imita o representa la conversación coloquial. El texto escrito (guion original) pasa por un proceso complejo de oralización dramatizada en el que se hacen los ajustes

verbales finales en función de la sincronía, restricción de tiempo fundamental para este tipo de traducción audiovisual. Desde un punto de vista acústico, observamos convergencias y divergencias prosódicas y de ajustes de calidad de voz que funcionan como estrategias vocales de intensificación o atenuación del conflicto o enfrentamiento, combinadas a las pistas visuales de la animación: alargamientos, tensión vocal, modalidades interrogativas o imperativas que intensifican el desalineamiento de la postura del que habla hacia su interlocutor.

Los actos expresivos negativos como críticas y desacuerdos se analizan desde el modelo de las emociones de Fontaine et alii (2007), para quienes la función expresiva se organiza a partir de grados en cuatro dimensiones: valencia, activación, dominancia e imprecisión. El concepto de evaluación y alineamiento conversacional divergente (Kiesling, 2022), desde una perspectiva socio interaccionista también es útil para analizar la toma de posición de los hablantes (*stancetaking*) en críticas y desacuerdos, con sus usos variables codificados lingüísticamente en distintos grados de convencionalización simbólica y sus respectivas traducciones en diálogos de animación.

Palabras clave: variación pragmática; expresividad; multimodalidad; desalineamiento; críticas y desacuerdos.

Bibliografía

de Moraes, J. A., & Rilliard, A. (2014). Illocution, attitudes and prosody: A multimodal analysis. In T. Raso & H. Mello (Eds.), Studies in Corpus Linguistics (Vol. 61, pp. 233–270). John Benjamins Publishing Company. https://doi.org/10.1075/scl.61.09mor

Fontaine, J R-J., Scherer, K., Roesch, E. B., & Elsworth, P.C. (2007). The World of Emotions Is Not Two-Dimensional. Psychological Science, 16(12), 1050–1057. https://doi.org/10.1111/j.1467-9280.2007.02024.x

Kiesling, S.F. (2022). Stance and Stancetaking. Annual Review of Linguistics, 8, 409–426. https://doi.org/10.1146/annurev-linguistics-031120-121256

Rilliard, A., Shochi, T., Martin, J.-C., Erickson, D., & Aubergé, V. (2009). Multimodal Indices to Japanese and French Prosodically Expressed Social Affects. Language and Speech, 52(2–3), 223–243. https://doi.org/10.1177/0023830909103171

Rilliard, A., Erickson, D., de Moraes, J. A., & Shochi, T. (2017). Perception of expressive prosodic speech acts performed in USA English by L1 and L2 speakers. Journal of Speech Sciences, 6(1), 27–45. https://doi.org/10.20396/joss.v6i1.14981

Romero, Lupe. When Orality Is Less Pre-fabricated: An Analytical Model for The Study of Colloquial Conversation in Audiovisual Translation. In: MCLOUGHLIN, Laura Incalcaterra, Marie Biscio y Maire Aine Ni Mhainnin (eds). Audiovisual Translation. Subtitles and Subtitling. Lausanne, Peter Lang, 2023.

Sánchez-Mompeán, Sofía. *The Prosody of Dubbed Speech*. Cham (Switzerland), Palgrave Macmillan, 2020.

Schneider, Klaus & Anne Barron (eds.). *Variational Pragmatics*. Amsterdam/Philadelphia, John Benjamins Publishing Company, 2008.

Shochi, T., Rilliard, A., & Erickson, D. (2023). Chapter 8. Perceptual changes between adults and children for multimodal im/politeness in Japanese. In A. H. Jucker, I. Hübscher, & L. Brown (Eds.), Pragmatics & Beyond New Series (Vol. 333, pp. 213–249). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.333.08sho

Doina Repede
(Universidad de Granada)

Corpusmigrasev (Corpus oral del español de los migrantes residentes en la ciudad de Sevilla). Construcción, codificación y exploración

Corpusmigrasev es un corpus informatizado constituido por entrevistas semidirigidas, basadas en un guion conversacional, realizadas a hablantes peruanos, colombianos, nicaragüenses y ecuatorianos residentes en la ciudad de Sevilla, grabadas entre los años 2020 y 2023. Las entrevistas realizadas se estructuran "en forma de relato de vida" (Paredes 2020: 60) o historias de vida (Lucca Irizarry y Berríos Rivera 2009; Ruiz Olbuénaga 2012) y buscan reconstruir y transmitir los acontecimientos vividos por los individuos como actores y participantes en la vida social (Chárriez Cordero 2012). Específicamente, en el *Corpusmigrasev*, las entrevistas están orientadas a que los informantes expresen "sus sensaciones y opiniones acerca del proceso migratorio, desde su lugar de origen hasta su situación actual" (Paredes 2020: 60) con la finalidad de tener una visión completa del escenario migratorio en el que están implicados los entrevistados.

El diseño y elaboración del presente corpus tiene como objetivo recopilar y codificar muestras orales representativas del español hablado de las diferentes comunidades hispanohablantes presentes en la ciudad de Sevilla, en consonancia con otros corpus ya recogidos, como el CORDIESIN (*Corpus dinámico del español de la inmigración*) en la comunidad de Madrid. Estos materiales permitirán obtener resultados sobre las características lingüísticas del español hablado por los migrantes, sobre su acomodación e integración sociolingüística, a la vez que posibilitará la realización de estudios contrastivos, por una parte, con el español de hablantes vernáculos, basados en diferentes corpus con características similares a este, y, por otra parte, con otros materiales ya recogidos sobre la comunidad migrante residente en España. En esta presentación expondremos las principales características del *Corpusmigrasev* y los criterios empleados en la construcción del corpus, la metodología de la recogida de los datos y su codificación, así como la consulta de la información que contiene esta base de datos.

Bibliografía

Chárriez Cordero, M. (2012): Historias de vida: Una metodología de investigación cualitativa. *Revista Griot*, 1, 50-67.

Lucca Irizarry, N. & Berríos Rivera, R. (2009) Investigación cualitativa. Fundamentos, diseños y estrategias. Puerto Rico: Ediciones SM.

Paredes García, F. (2020): "Un modelo para el análisis de la integración sociolingüística de la población migrante: fundamentos, dimensiones e instrumentos", *Lengua y migración*, 12(1), Monográfico, pp. 39-81. Disponible en:

https://erevistas.publicaciones.uah.es/ojs/index.php/lenguaymigracion/article/view/65.

Ruíz Olabuénaga, J. I. (2012). Historias de vida. En Metodología de la Investigación Cualitativa. Bilbao: Universidad de Deusto. pp. 267-313.

Malte Rosemeyer
(Freie Universität Berlin)

Fictive orality in a corpus of historical theater plays

It is commonly assumed that language change emerges from specific communicative strategies in interaction such as expressiveness (Detges and Waltereit 2002), discourse organization (Rosemeyer and Grossman 2017, Rosemeyer and Grossman 2021) or economy (e.g., Haspelmath 1999). However, when applied to actual historical data, these accounts of semantic change are difficult to verify. This is due to the fact that written documents that have survived until today can almost never be said to represent data from actual spontaneous linguistic interaction. In order to be able to assess the relevance of these hypotheses for langueg change, we need to develop methods that enable us to describe to which extent the analyzed historical data can be said to reflect routines in spoken interaction at the time in question.

On the basis of prior studies (Rosemeyer 2019, Rosemeyer and Becker In press), I propose that by applying register analysis (Biber and Finegan 1997, Biber and Finegan 2004) to a large corpus of theater plays, it is possible to describe the degree to which these theater plays approximate actual spoken interaction at the time of composition of the play. Focusing on the development of interrogative constructions and the present perfect in French and Brazilian Portuguese, I demonstrate the relevance of this approach. In particular, incorporating register-related aggregate measurements of formality into quantitative longitudinal analyses of language change allows determining which constructional changes can be described as having originated in informal interaction, and which seem to be the result of indirect indexing, corresponding to stylistically marked "changes from above" (Meyerhoff 2006: 222-225).

References

Biber, D. and E. Finegan (1997). Diachronic Relations among Speech-Based and Written Registers in English. *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. T. Nevalainen and L. Kahlas-Tarkka. Helsinki, Finland, Societe Neophilologique: 253-275.

Biber, D. and E. Finegan (2004). Historical drift in three English genres. *Corpus Linguistics: Readings in a Widening Discipline*. G. Sampson and D. McCarthy. London, Continuum: 67-77.

Detges, U. and R. Waltereit (2002). "Grammaticalization vs. reanalysis: a semantic-pragmatic account of functional change in grammar." *Zeitschrift für Sprachwissenschaft* 21(2): 151-195.

Haspelmath, M. (1999). "Explaining Article-Possessor Complementarity: Economic Motivation in Noun Phrase Syntax." *Language* 75(2): 227-243.

Meyerhoff, M. (2006). *Introducing Sociolinguistics*. London, New York, Routledge.

Rosemeyer, M. (2019). "Actual and apparent change in Brazilian Portuguese wh-interrogatives." *Language Variation and Change* 31(2): 165–191-165–191.

Rosemeyer, M. and M. Becker (In press). The Brazilian Portuguese present perfect: from nominal to verbal pluractionality. *Indefinites in Romance and Beyond*. O. Kellert and M. Rosemeyer. Berlin, Language Science Press.

Rosemeyer, M. and E. Grossman (2017). "The road to auxiliariness revisited: the grammaticalization of FINISH anteriors in Spanish." *Diachronica* 34(4): 516-558.

Rosemeyer, M. and E. Grossman (2021). "Why don't grammaticalization pathways always recur?" *Corpus Linguistics and Linguistic Theory* 17(3): 653-681. Ahead of print.

Josep Sarria Navarro
(Universitat de València)

Va i diu...: estudi de la pseudocoordinació en català a partir de corpus orals i escrits

La construcció v-i-v ('va i diu', 'va i resulta', 'agafa i se'n va', etc.) és un fenomen lingüístic que consisteix en una mena de pseudocoordinació entre dos verbs, es podria dir que a mitjan camí entre la mateixa coordinació i les construccions perifràstiques verbals, ja que comparteix trets amb totes dues, com ara la unió amb la conjunció 'i' i la pèrdua del significat lèxic del primer verb, respectivament. En aquest sentit, en seqüències que segueixen l'esquema, el primer verb pertany a una sèrie tancada, ha perdut la pràctica totalitat del seu significat original, i el segon pertany a una sèrie oberta, selecciona l'estructura argumental de l'oració i aporta el significat bàsic de l'estructura. Al seu torn, l'estructura en conjunt es gramaticalitza i adquireix matisos pragmàtics nous de mirativitat, de contrarietat o de brusquedat.

Encara no hi ha hagut un consens definitiu pel que fa a la nomenclatura/classificació d'aquesta construcció (perífrasi verbal, pseudocoordinació, construcció multiverbal, construcció paratàctica, hendíadis...) (Ross, 2014 & 2021; Coseriu, 1977; Jaque *et al*., 2018 & 2019) i la seua descripció tampoc no ha estat gaire aprofundida en la bibliografia, però hi ha referències tant en anglés (Ross, 2021; De Vos, 2005; Stefanowitsch, 1999 & 2000), com en castellà (Garachana, 2022; Bravo 2020; García Sánchez, 2003, 2007; García Fernández *et al*., 2006). Aquesta construcció també és present en català, tal com veiem en els exemples següents, però no disposem d'estudis que l'hagen estudiada:

(1) Però uno / tenia uno i va anar per un atre / **i va i** van bessonar (*Parlars*, Pinet).

(2) **I va** i ma mare **fa** i diu <EDirecte>M'han dit que al cine ny'ha una pel·lícula molt bonica / teníeu que anar / només acabeu de dinar (*Parlars*, Castelló).

(3) Va pegar sarpà pa agarra-me-la però no me va agarrar i jo <EDirecte>Ai ai ai ai ai!</EDirecte> **I va** i ma mare <EDirecte>Què passe?</EDirecte> i jo dic <EDirecte>Un xiqüelo que me volie llevar una queraïlla</EDirecte> (Parlars, Castelló).

El treball que ara presentem, així doncs, és un estudi preliminar sobre la construcció v-i-v en català a partir de la consulta a diversos corpus lingüístics orals del català actual, com ara el corpus *Parlars* o el CCCUB, i també a corpus escrits com el CTILC, el CIVAL o alguns corpus d'accés lliure disponibles en l'eina Sketch Engine. En

conseqüència, es farà servir la lingüística de corpus per a obtenir totes les dades lingüístiques possibles de les seqüències v-i-v i els contextos discursius en què solen aparèixer.

En primer lloc, farem cerques en el corpus i destriarem els casos que corresponen a la construcció objecte de l'estudi i els que no, i tot seguit, farem una proposta de descripció sintàctica, morfològica, semàntica i pragmàtica dels trets generals d'aquestes seqüències a partir de la teoria de la gramàtica de construccions (Goldberg, 1995, 2003; Ibarretxe-Antuñano & Valenzuela, 2012). També tindrem en compte la funció discursiva en els diversos tipus de textos orals (monòlegs, converses) i escrits (teatre, narrativa, poesia, premsa) estudiats i la variació dialectal i els registres en què apareix la construcció, així com la freqüència d'ús. Finalment, extraurem les conclusions pertinents de l'anàlisi i en compararem els resultats amb estudis fets en altres llengües, especialment l'anglés i el castellà, que són les llengües en què més dades tenim sobre aquesta construcció (Ross, 2014, 2016, 2021; De Vos, 2005; Stefanowitsch, 1999 & 2000; Wulff, 2007; García Sánchez, 2004 & 2007; García Fernández *et al*., 2006; Jaque *et al*., 2019, 2022; Bravo, 2020; Covarrubias *et al*., 2020; Orqueda *et al*., 2020; Garachana, 2022; Kornfeld, 2019 & 2022).

Referències

Aikhenvald, A., & Muysken, P. (ed.). (2011). Multi-verb constructions: A view from the Americas. Brill.Bravo, A. (2020). On pseudo-coordination in Spanish. *Borealis (Tromsø)*, *9*(1), 125-180. https://doi.org/10.7557/1.9.1.5365

Coseriu, E. (1977). *Estudios de linguística románica*. Gredos.

Covarrubias, M., Guerrero, S., González Vergara, C., Jaque, M., Orqueda, V., & Hasler, F. (2020). Aquí llegas, pero allá coges: Distribución dialectal de los auxiliares de las construcciones multiverbales de verbos finitos coordinados en español. *Itinerarios. Revista de estudios lingüísticos, literarios, históricos y antropológicos*, *31*, 229-250. https://doi.org/10.7311/ITINERARIOS.31.2020.12

De Vos, M. (2004). Pseudo coordination is not subordination. *Linguistics in the Netherlands*, *21*, 181-192. https://doi.org/10.1075/avt.21.20vos

De Vos, M. (2005). *The syntax of verbal pseudo-coordination in English and Afrikaans*. LOT.

Dingemanse, M. (2017). On the margins of language: Ideophones, interjections and dependencies in linguistic theory. En *Dependencies in language: On the casual ontology of linguistic systems*. Language Science Press. https://doi.org/10.5281/ZENODO.573781

Garachana Camarero, M. (2022). Unexpected grammaticalizations: The reanalysis of the Spanish verb ir 'to go' as a past marker. En M. Garachana Camarero, S. Montserrat Buendia, & C. D. Pusch (Ed.), *IVITRA Research in Linguistics and Literature* (Vol. 31, p. 171-188). John Benjamins Publishing Company. https://doi.org/10.1075/ivitra.31.09gar

García Fernández, L., Ángeles Carrasco Gutiérrez, Bruno Bergareche, María Martínez-Atienza, & María de los Ángeles García García-Serrano. (2006). *Diccionario de perífrasis verbales*. Gredos.

García Sánchez, J. J. (2003). «Tomo y me voy». Entre el influjo bíblico y la gramaticalización obvia. En *Studi in memoria di Eugenio Coseriu* (p. 139-150).

García Sánchez, J. J. (2007). «Tomo y me voy». Expresión plena y elipsis. En *Actas del XV Congreso de la Asociación Internacional de Hispanistas: Las dos orillas*. Fondo de Cultura Económica : Asociación Internacional de Hispanistas : Tecnológico de Monterrey : El Colegio de México.

Goldberg, A. E. (1995). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Goldberg, A. E. (2003). *Constructions: A construction grammar approach to argument structure*. The Univ. of Chicago Press.

Ibarretxe-Antuñano, I., & Valenzuela, J. (ed.). (2012). *Lingüística cognitiva*. Anthropos.

Jaque, M., González, C., Guerrero, S., Hasler, F., & Orqueda, V. (2019). Es llegar y llevar: Construcciones multiverbales de verbo finito coordinadas en español. *Lenguas Modernas*, *o*(52), 163-186.

Jaque, M., González, C., Orqueda, V., Guerrero, S., Hasler, F., & Covarrubias, M. (2022). A la altura de las expectativas: Interacciones entre la negación y construcciones multiverbales del tipo llegar y + VF. *Verba: Anuario Galego de Filoloxía*, 1-36. https://doi.org/10.15304/verba.49.7380

Kornfeld, L. M. (2019). Expresión de la sorpresa, miratividad y gramaticalización de verbos inacusativos en español. *Borealis – An International Journal of Hispanic Linguistics*, *8*(2), 165-197. https://doi.org/10.7557/1.8.2.4913

Kornfeld, L. M. (2022). Estructuras seudocoordinadas. En *Universales vernáculos en la gramática del eespañol* (Vol. 85, p. 337). Iberoamericana Vervuert.

Ross, D. (2014). El origen de los estudios sobre la pseudocoordinación verbal. *Diálogo de la lengua*, *VI*, 116-132.

Ross, D. (2021). *Pseudocoordination, Serial Verb Constructions and Multi-Verb Predicates: The relationship between form and structure* (v1.0 (Original version, as deposited with university for degree)) [Zenodo]. https://zenodo.org/record/5546425

Solà, J., Lloret, M.-R., Mascaró, J., & Pérez Saldanya, M. (Ed.). (2002). *Gramàtica del català contemporani*. Editorial Empúries.

Stefanowitsch, A. (1999). THE GO-AND-VERB CONSTRUCTION IN A CROSS-LINGUISTIC PERSPECTIVE: IMAGE- SCHEMA BLENDING AND THE CONSTRUAL OF EVENTS. En D. Nordquist & C. Berkenfield (Ed.), *Proceedings of the Second Annual High Desert Linguistics Society Conference* (Vol. 2). High Desert Linguistics Society.

Stefanowitsch, A. (2000). The English GO-(PRT)-AND-VERB Construction. *Annual Meeting of the Berkeley Linguistics Society*, *26*(1), 259. https://doi.org/10.3765/bls.v26i1.1158

Wulff, S. (2006). Go-V vs. go-and-V in English: A case of constructional synonymy? En S. Th. Gries & A. Stefanowitsch (Ed.), *Corpora in Cognitive Linguistics* (p. 101-126). Mouton de Gruyter. https://doi.org/10.1515/9783110197709.101

Carolina Jorge Trujillo & Imelda Chaxiraxi Díaz Cabrera
(Universidad de La Laguna)

Comparación de Fo y duración en un corpus de español cubano

Los estudios prosódicos del español han estado frecuentemente centrados en el tonema o núcleo oracional, puesto que, entre otros aspectos, registra importantes variaciones de frecuencia fundamental (Fo). Por ejemplo, el esquema melódico final de las oraciones declarativas en nuestra lengua suele ser, en general, descendente, mientras que las interrogativas registran mayor variedad. Es el caso del esquema nuclear circunflejo o ascendente-descendente, presente en el canario, cubano y venezolano, que contrasta con variedades como las del español central o el texano, con final ascendente, o el colombiano, en el que se ha registrado una variación diatópica que da lugar a uno u otro patrón según las distintas zonas geográficas. Por su parte, la carga informativa del pretonema, o prenúcleo, suele ser menor que la del núcleo, por lo que no ha recibido tradicionalmente tanta atención investigadora en este sentido. Sin embargo, en este segmento oracional también se observan fenómenos prosódicos relevantes, como es el caso del pico máximo inicial. Así, diversos estudios han registrado una doble tendencia: esta cumbre tonal puede aparecer alineada con la primera sílaba tónica o desplazada a la postónica, dando lugar en este caso al fenómeno conocido como *overshooting* (Sosa 1995, 1999; Face 2002; Dorta, ed., 2013, 2018).

Otros parámetros acústicos, como la duración y la intensidad, han sido menos estudiados que la Fo, si tenemos en cuenta que los trabajos de prosodia son relativamente recientes (en comparación con otros planos de la lengua) y que ello ha influido en que se comenzara a estudiar principalmente la frecuencia fundamental en sus orígenes (Dorta, ed., 2013, 2018).

Así pues, el objetivo de este trabajo es realizar un análisis comparativo de los dos índices prosódicos mencionados, Fo y duración, para responder a la siguiente pregunta de investigación: ¿el comportamiento temporal en el pretonema puede relacionarse con variaciones de Fo, específicamente con la presencia o no de *overshooting*?

En el contexto del español cubano, partimos de un corpus formal o de laboratorio, emitido por mujeres y hombres en las modalidades declarativa e interrogativa, que se corresponde con la metodología seguida en el proyecto AMPER (*Atlas Multimedia*

de Prosodia del Espacio Románico). Hemos utilizado el umbral psicoacústico de 1,5 St (Rietveld y Gussenhoven 1985; Pamies *et al.* 2002) para determinar las variaciones significativas de F0, y el del 33,33 % (Fernández Planas y Martínez Celdrán, 2003) para estudiar la duración. A partir del análisis acústico, hemos realizado el etiquetaje de la F0 a partir de la propuesta de Dorta (ed., 2013, 2018), y de la duración basándonos en Muñetón Ayala, Díaz Cabrera, Dorta Luis (2018) y Dorta (2019).

Palabras clave: frecuencia fundamental, duración, pico máximo, *overshooting*, variedades del español

Bibliografía

Dorta Luis, J. (ed.) (2013). *Estudio comparativo preliminar de la entonación de Canarias, Cuba y Venezuela*. La Página ediciones S/L, Colección Universidad.

Dorta Luis, J. (ed.) (2018). *La entonación declarativa e interrogativa en cinco zonas fronterizas del español: Canarias, Cuba, Venezuela, Colombia y San Antonio de Texas*. Studien zur romanischen sprachwissenschaft und interkulturellen kommunikation. Herausgegeben von Gerd Wotjak. Peter Lang Edition. DOI: 10.3726/b12056

Dorta Luis, J. (2019). Estructuras tonales y de duración en la entonación del español de hablantes bilingües americanos con ascendencia mexicana. *Onomázein, 45,* 232-258. DOI: 10.7764/onomazein.45.04

Face, T. L. (2002). Local intonational marking of Spanish contrastive focus. *Probus, 14*(1), 71-92. DOI:10.1515/prbs.2002.006

Fernández Planas, A. M.ª, & Martínez Celdrán, E. (2003). El tono fundamental y la duración: dos aspectos de la taxonomía prosódica en dos modalidades de habla (enunciativa e interrogativa) del español. *Estudios de Fonética Experimental, 12,* 165-200.

Muñetón Ayala, M., Díaz Cabrera, Ch., & Dorta Luis, J. (2018). La duración en oraciones sin expansión en la voz femenina de dos países fronterizos: Colombia (Bogotá -Medellín) y Venezuela (Caracas-Mérida). *Literatura y Lingüística, 37,* 401-423. DOI:10.29344/0717621X.37.1389

Pamies Bertrán, A., Fernández Planas, A. M.ª, Martínez Celdrán, E., Ortega, A., Amorós Céspedes, M. C. (2002). Umbrales tonales en español peninsular. *Actas del II Congreso de Fonética Experimental* (pp. 272-278). Universidad de Sevilla.

Rietveld, T., & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics, 13,* 299-308.

Sosa, J. M. (1995). Nuclear and pre-nuclear tonal inventories and the phonology of Spanish declarative intonation. En Elenius, K., & Branderud, P. (eds.), *Proceedings of the XIIIth Internacional Congress of Phonetic Sciences* (pp. 646-649). Stockholm University.

Sosa, J. M. (1999). *La entonación del español, su estructura fónica, variabilidad y dialectología.* Cátedra.

Verena Weiland
(Universität Bonn)

"Las tierras altas se comen las vocales, las tierras bajas se comen las consonantes": Revisión de la dicotomía dialectal de Hispanoamérica mediante el corpus TiAlBa

Desde principios del siglo XX, se han propuesto varias ideas divergentes para dividir Hispanoamérica en zonas dialectales. Dependiendo de los criterios utilizados, estas propuestas han sugerido entre 5 y 250 zonas diferentes (por ejemplo, Henríquez Ureña 1921, Canfield 1962, Resnick 1975). Las discrepancias confirman la complejidad de este tema; hasta la fecha, no hay un modelo estándar establecido. Sin embargo, una regla general ampliamente aceptada en la lingüística hispánica simplifica la zonificación: "las tierras altas se comen las vocales, las tierras bajas se comen las consonantes" (Henríquez Ureña 1921, Rosenblat 1973: 39). Según esta regla, el debilitamiento de las consonantes implica la aspiración o eliminación de /s/, /d/ intervocálica y /r/ final, la neutralización de /r/ y /l/ implosivas, y el debilitamiento de /x/ (correspondiendo al grafema <j>, Rosenblat 1973: 39-40). La reducción de vocales afecta principalmente a /e/ y /o/ en sílabas abiertas y en conexión con oclusivas (/p/, /t/, /k/) o con /s/ final (Noll 2019: 29).

El problema fundamental de esta distinción entre tierras altas y bajas es que carece de una base de datos suficiente (De Crignis 2016: 6). Por lo tanto, para evaluar su validez, se utilizaron datos del corpus *Tierras Altas y Bajas de Hispanoamérica* (TiAlBA). Este corpus incluye grabaciones de más de 20 ciudades y pueblos de Hispanoamérica. La mayoría de estos datos se obtuvieron entre julio de 2023 y abril de 2024, por lo que se trata de datos lingüísticos nuevos y actualizados. Cabe mencionar que también se incluyen datos de regiones de Hispanoamérica para las que actualmente faltan datos, por ejemplo, El Salvador y Panamá en Centroamérica o Venezuela y Bolivia en Sudamérica.

En cada punto de investigación se realizaron grabaciones de voz de 12 personas, copilando datos de unas 250 personas nativas del español. Repartidos en 11 países y distribuidos equitativamente en cuanto a edad y sexo. El protocolo del corpus sigue la estructura del proyecto *Fonología del Español contemporáneo* (Pustka et al. 2018) e incluye la lectura de una lista de palabras y de un texto, así como una entrevista, de modo que se tienen en cuenta diversas formas de control del habla (Labov 1972).

En un primer paso, se transcribieron ortográficamente las grabaciones del corpus TiAlBa. En un segundo paso, se segmentaron las consonantes enumeradas y las vocales para analizarlos en función de criterios perceptivos y acústicos (Praat, Boersma/Weenink 2020). El análisis estadístico (Hothorn/Hornik/Zeileis 2006) confirma la importancia de la variable ZONA ALTA/BAJA para la pronunciación de las consonantes /s/, /d/ y /r/. Además, las variables PROTOCOLO (lista de palabras, texto o entrevista), TIPO DE PALABRA (por ejemplo, verbo, sustantivo) y CONTEXTO FONOLÓGICO (por ejemplo, /s/ intervocálica o en posición final de palabra) explican no solo la pronunciación de /s/, sino también algunas diferencias entre los propios lugares de las tierras altas y bajas. Un estudio piloto sobre las vocales no confirma el debilitamiento vocálico en las tierras altas.

Bibliografía

Boersma, Paul/Weenink, David (2020): *Praat: doing phonetics by computer*. <https://www.fon.hum.uva.nl/praat/>.

Canfield, Lincoln (1962): *La pronunciación del español en América: ensayo histórico-descriptivo*. Bogotá: Instituto Caro y Cuervo.

De Crignis, Patricia (2018): *Vokalschwächung im peruanischen Spanisch*, München: LMU.

Henríquez Ureña, Pedro (1921): „Observaciones sobre el español de América“, en: *Revista de Filología Española* 8, 357-390.

Hothorn, Torsten/Hornik, Kurt/Zeileis, Achim (2006): „Unbiased recursive partitioning: conditional inference framework“, en: *Journal of Computational and Graphical statistics* 15(3), 651-674.

Labov, William (1972): *Sociolinguistic patterns*. Oxford: Blackwell.

Noll, Volker (2019): *Das amerikanische Spanisch. Ein regionaler und historischer Überblick*, Berlin/Boston: de Gruyter.

Pustka, Elissa/Gabriel, Christoph/Meisenburg, Trudel/Burkard, Monja/Dziallas, Kristina (2018): „(Inter-)Fonología del Español Contemporáneo/(I)FEC: metodología de un programa de investigación para la fonología de corpus“, en: *Loquens* 5(1), 1-16.

Resnick, Melvyn C. (1976): „Algunos aspectos histórico-geográficos de la dialectología hispanoamericana“, en: *Orbis* 15, 264-276.

Rosenblat, Angel ([1962] 1973): *El castellano de España y el castellano de América. Unidad y diferenciación*, Madrid: Taurus.

Craig Welker
(Universität Bern)

Convergencia a un mediador no presente - Evidencia del español de Juchitán, México

Según Bell (1984), lxs hablantes suelen acomodarse a su destinatario (addressee), en lo que se denomina diseño de audiencia, aunque pueden, en determinadas circunstancias, acomodarse a un/a mediador/a (referee) que no está presente, con el/la que se identifican y se afilian, en un proceso denominado diseño de mediador. Sin embargo, en la actualidad existe una falta de consenso sobre el grado en que la agencia desempeña un papel en la acomodación. Aunque el modelo de Bell (1984) considera la acomodación como un proceso bastante agentivo, algunxs investigadores, como Trudgill (2008), asumen que la acomodación suele producirse de forma casi automática, independientemente de la identidad de lxs hablantes y del nivel de afiliación con su interlocutor/a. Si Trudgill (2008) tuviera razón, se esperaría que lxs hablantes convergieran también a mediadores con lxs cuales no se afilian y no se identifican. Los resultados de un proyecto sobre la variación en un corpus del español de Juchitán, México, una comunidad indígena bilingüe en el español y el zapoteco del istmo, arrojan luz sobre esta cuestión.

Este proyecto investigó la variación en la /s/ implosiva y el género gramatical usado para hablar de referentes muxes (una identidad local, a menudo considerada un "tercer género" entre los hombres y las mujeres). Tras este análisis, se averiguó que los hombres hispanohablantes, con mucha más frecuencia que lxs otrxs hablantes, utilizan la variante retenida de la /s/ y el género gramatical masculino para hablar de referentes muxes. Se compararon también los patrones de variación que emergen cuando lxs hablantes asumen posturas (stances) asociadas con la ideología

machista y la ideología pro-española local con aquellas que se dan cuando lxs hablantes no mencionan estas ideologías. Ya que la ideología pro-española y la ideología machista perjudican a las mujeres, a lxs muxes y a lxs hablantes del zapoteco, en Juchitán se vinculan comúnmente con los hombres hispanohablantes.

Mientras lxs hablantes en el corpus mencionan las ideologías asociadas con los hombres hispanohablantes, convergen hacia ellos (utilizan más género gramatical masculino y /s/ retenida), aun cuando no están de acuerdo con estas ideologías. Incluso, las mujeres hablantes de zapoteco muestran este patrón de convergencia. Es decir, cuando los comentarios vinculados con los hombres hispanohablantes

forman parte del discurso, lxs participantes se acomodan a un mediador imaginario, masculino e hispanohablante, con el que no necesariamente se identifican ni se afilian. Por ende, sugiero una revisión de la teoría del diseño de la audiencia de Bell (1984) que lleva a la teoría en una dirección más consistente con el argumento de Trudgill (2008), en el sentido de que la acomodación tiende a ocurrir de forma casi automática. Concretamente, parece que lxs hablantes convergen a menudo hacia mediadores en su discurso, aun si no se afilian con ellxs. Al parecer, la tendencia psicológica de lxs hablantes a converger hacia su interlocutor/a es tan marcada que hasta puede llevar a la convergencia con personas imaginarias con las que lxs hablantes quieren desafiliarse.

Bibliografía

Bell, A. (1984). Language style as audience design. Language in society, 13(2), 145-204.

Trudgill, P. (2008). Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation. Language in Society, 37(2), 241-254.