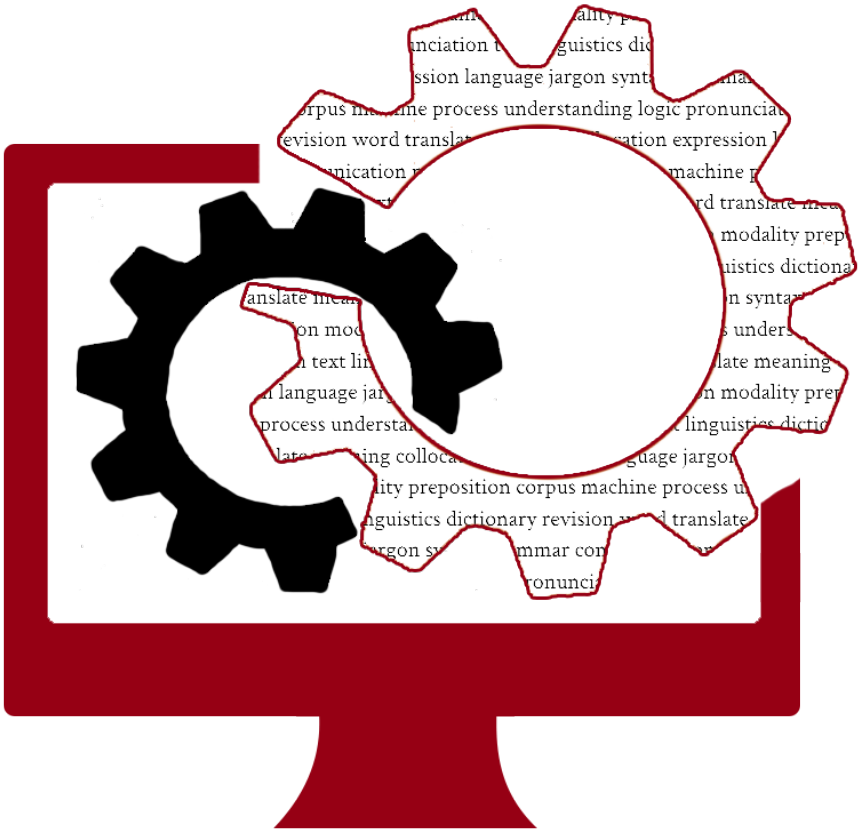

CILC 2025

SALAMANCA
15-17 DE MAYO DE 2025

BOOK OF ABSTRACTS
LIBRO DE RESÚMENES

aelinco
Asociación Española de Lingüística de Corpus

Organising Committee / Comité Organizador

Lucía Bausela Buccianti
Maria Bîrlea
Sara Casco Solís
Izaskun Elorza (Chair)
Alejandro Guevara Guevara
Nely Iglesias Iglesias
Vasilica Mocanu
Elisa Pérez García
René Pérez Tissens
Nora Rodríguez Loro
Javier Ruano-García (Chair)
Pilar Sánchez García
Paula Schintu Martínez
Carmen Sumillera Iglesias
Agata Żelachowska

cilc2025.usal.es **@cilc2025**

DING
English Dialectology and the History of the English Language

Lindes
Linguistic Descriptions of English
Lingüística Descriptiva del Inglés

Contents / Contenidos

Plenary Lectures / Ponencias Plenarias 1

Round Tables / Mesas Redondas 8

Workshops / Talleres 18

Parallel Sessions / Sesiones Paralelas 20

Posters 135

PLENARY LECTURES / PONENCIAS PLENARIAS

‘up yours at the end of life’:
Opposition, emotion, and overlap in a corpus of Scottish debates on assisted dying

Marc Alexander, James Balfour

University of Glasgow

In this plenary, we report on findings from a recent corpus-based study examining public discourse surrounding the Assisted Dying for Terminally Ill Adults (Scotland) Bill, which proposes to legalize medically assisted death for mentally competent terminal patients. With public opinion sharply divided—faith groups generally opposing while public polls show over 70% support—the study analyzes both public consultation responses and media coverage to understand key attitudes and arguments.

The research comprises two complementary studies. The first analyzes public responses to government consultations, drawing from two datasets: 12,314 written submissions from 2022 (2.1 million words) and 7,236 responses from 2024 (1.8 million words). The second examines the media narrative around assisted dying between 2022-2024 in the UK (6,360 texts). By examining language patterns in the two datasets in tandem we reflect on the complex interaction between public attitudes towards a sensitive and contentious topic and the role media framing plays in shaping public debate. To examine oppositional discourse in the public responses, we first identify unique lexical choices exclusive to each group—supporters used terms like "hideous," "abject," "urine," and "linger," while opponents employed words such as "eroded," "burden," "wedge," and "shalt." Second, tagging the corpus using WMatrix, we compare key semantic domains between groups, revealing that supporters' responses contained more emotional content (particularly in the "Sad" domain), while opponents' responses were more analytical in nature. Third, we conduct n-gram studies to identify areas of common ground between opposing viewpoints and detect potential copy-pasted responses within each group.

For the second part of the study, we investigate media influence on the debate by analyzing 6,360 news articles published between 2022-2024 that explicitly referenced assisted dying. Using keyword analysis, n-grams, and concordance analysis, we examine how different politically-affiliated newspapers framed the debate. Our findings suggest significant overlap between media narratives and public consultation responses, with some phraseology being nearly identical across both datasets.

The research extends beyond the specific assisted dying debate to address broader methodological questions about analyzing oppositional discourse in public life. The study revealed distinct rhetorical patterns: supporters of the bill tended to employ more emotionally charged language and personal narratives, while opponents favored analytical and consequence-based arguments. The media analysis demonstrated how news coverage might reinforce these polarized positions through consistent framing patterns. The bill's consideration in Scotland occurs against the backdrop of similar debates in England, where recent legislative efforts have faced setbacks. The research notes the particular challenges of analyzing representative corpora in highly polarized debates, where responses may range from deeply personal experiences to organized campaign submissions.

This comprehensive analysis of both public and media discourse provides valuable insights into how contentious healthcare policy debates are framed and argued across

different platforms and stakeholder groups. The methodological approach developed here offers a framework for analyzing other polarized public debates, while the findings contribute to our understanding of how public opinion forms and expresses itself on complex ethical issues.

References

McArthur, L. (2022). *Proposed Assisted Dying for Terminally Ill Adults Bill: Consultation Document*. Edinburgh: Scottish Parliament.

Health, Social Care and Sport Committee (2024). *CitizenSpace Call for Views on Assisted Dying for Terminally Ill Adults*. Edinburgh: Scottish Parliament.

Rayson, P. (2008). From key words to key semantic domains. International Journal of *Corpus Linguistics*, *13*(4), 519-549.

Rayson, P. (2009). *Wmatrix: A Web-based Corpus Processing Environment*. Lancaster: Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/

La inteligencia artificial generativa en el laberinto de la traducción especializada

Pascual Cantos

Universidad de Murcia

En este estudio se analiza la eficacia de la inteligencia artificial (IA) generativa como herramienta de traducción automática aplicada a dominios especializados. La investigación se fundamenta en un enfoque metodológico que combina la lingüística de corpus con técnicas cuantitativas avanzadas, empleando corpus paralelos en los campos biomédico, jurídico y técnico. Estos corpus permiten evaluar la capacidad de los sistemas de IA generativa para abordar tareas complejas como la precisión léxica, la cohesión discursiva y la adaptación contextual, aspectos críticos en la traducción especializada.

El diseño de la investigación se centra en la implementación de análisis cuantitativos exhaustivos. Entre las técnicas empleadas destaca el análisis de frecuencia léxica y de patrones de n-gramas, ambos dirigidos a identificar la consistencia terminológica y a detectar posibles incoherencias en el uso del vocabulario técnico. Para la evaluación de la calidad de las traducciones, se aplican métricas ampliamente reconocidas en el campo de la traducción automática, como BLEU (*Bilingual Evaluation Understudy*) y COMET (*Crosslingual Optimized Metric for Evaluation of Translation*). BLEU evalúa la correspondencia entre las traducciones generadas y las traducciones de referencia, mientras que COMET incorpora un enfoque más matizado basado en modelos neuronales que predicen la calidad de la traducción tomando como referencia juicios humanos.

Asimismo, el estudio adopta un análisis de varianza para explorar la adecuación contextual de las traducciones generadas en los diferentes dominios. Este enfoque permite medir cómo los modelos de IA gestionan variaciones en el contexto lingüístico y aseguran la coherencia discursiva, un requisito clave en los textos especializados. La metodología incluye, además, una comparación entre los resultados obtenidos por los sistemas de IA y las traducciones humanas, considerando parámetros como la precisión semántica y la adaptabilidad estilística.

Un aspecto esencial del diseño metodológico es la preparación y curación de los corpus empleados. Estos corpus paralelos se construyen a partir de fuentes autorizadas en los tres dominios seleccionados, asegurando una representación adecuada de la terminología y los estilos discursivos específicos de cada área. Se prioriza la inclusión de textos auténticos que reflejen un uso realista del lenguaje técnico, lo que facilita una evaluación más precisa de las capacidades de los modelos generativos. La elección cuidadosa de estos textos es clave para garantizar la validez de las conclusiones extraídas.

El objetivo principal de este estudio es proporcionar una evaluación integral de la IA generativa en la traducción automática especializada. La investigación no solo examina las capacidades actuales de estos modelos, sino que también sienta las bases para optimizar su rendimiento mediante la incorporación de corpus de entrenamiento más específicos y estrategias de evaluación mejoradas. Además, busca establecer un marco metodológico que pueda ser replicado en estudios futuros sobre traducción automática en otros dominios especializados. Esta aproximación metodológica, que combina análisis lingüísticos detallados con técnicas de evaluación cuantitativa, contribuye a un entendimiento más profundo del potencial de la IA generativa en entornos de traducción profesional.

Referencias

Al-Onaizan, Y., y Papineni, K. (2006, julio). Distortion models for statistical machine translation. En *Proceedings of the 21st International Conference on Computational*

Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (pp. 529-536).

Baxi, V., Edwards, R., Montalto, M., y Saha, S. (2022). Digital pathology and artificial intelligence in translational medicine and clinical practice. *Modern Pathology*, *35*(1), 23-32.

Birch, A. (2021). Neural Machine Translation 2020, by Philipp Koehn, Cambridge, Cambridge University Press, ISBN 978-1-108-49732-9, pages 393. *Natural Language Engineering*, *27*(3), 377-378.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., ... y Zampieri, M. (2016, August). Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation* (pp. 131-198). Association for Computational Linguistics.

Bojar, O., Graham, Y., Kamran, A., y Stanojević, M. (2016, August). Results of the wmt16 metrics shared task. En *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers* (pp. 199-231).

Fu, L., y Liu, L. (2024). What are the differences? A comparative study of generative artificial intelligence translation and human translation of scientific texts. *Humanities and Social Sciences Communications*, *11*(1), 1-12.

Jooste, W., Haque, R., y Way, A. (2021). Philipp Koehn: Neural Machine Translation: Cambridge University Press. *Machine Translation*, *35*(2), 289-299.

Liu, K., y Afzaal, M. (2021). Artificial Intelligence (AI) and translation teaching: A critical perspective on the transformation of education. *International Journal of Educational sciences*, *33*(1-3), 64-73.

Papineni, K., Roukos, S., Ward, T., y Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

Rei, R., De Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., ... & Martins, A. F. (2022, diciembre). COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 578-585).

Rei, R., Stewart, C., Farinha, A. C., y Lavie, A. (2020). COMET: A neural framework for MT evaluation. *arXiv preprint arXiv:2009.09025*.

Rodríguez del Rosario, C. (2021). *Creación de motores de traducción automática (estadística y neuronal) inglés-español especializados en el campo de la aviación con la herramienta MTUOC*. Universitat Oberta de Catalunya.

Tiedemann, J. (2012, may). Parallel data, tools and interfaces in OPUS. In *Lrec* (Vol. 2012, pp. 2214-2218).

The relevance of large, structured corpora in the age of Large Language Models

Mark Davies

Brigham Young University

I will provide a summary of the in-depth data from several "white papers" at English-Corpora.org, on how well the predictions of two prominent Large Language Models (LLMs) match the actual data from several robust corpora, including corpora from Sketch Engine, and several corpora from English-Corpora.org (COCA, COCA, GloWbE, NOW, iWeb, the TV and Movie corpora, and more). I will also provide limited data from the three corpora in the Corpus del español and the three corpora in the Corpus do português.

In terms of strengths, the LLMs arguably provide:

- Much richer collocational data than even 40-50 billion word corpora from Sketch Engine (especially for low frequency words). This is due to the advanced word embeddings in high-dimensional space in LLMs, which are much more powerful than the simplistic surface level association measures used in corpus linguistics.
- Better comparisons of contrasting words (e.g. *entire / complete*, *nuance / subtlety*, *perceive / discern* for English; we will also provide data from Spanish)
- Much more insightful analyses (generated by the LLMs themselves) of what the collocates tell us about the meaning and usage of words

The LLMs are surprisingly good (perhaps at the level of some of the best corpora) at:

- Estimating word and phrase frequency (such as rank ordering a list of 10-20 words)
- Categorizing words and phrases by dialect, historical period, and dialect
- Analyzing variation in word meaning across genres, historical periods, and dialects
- Predicting syntactic variation – between genres, historical periods, and dialects.

However, there LLMs have the following significant limitations, as far as providing language data and carrying out linguistic analyses:

- They are much worse at *generating* word and phrase lists (such as those at WordFrequency.info) than in analyzing / categorizing existing lists
- We can never be sure if they are actually *generating* useful linguistic data themselves (for example, actual data on syntactic variation between genres, time periods, or dialects), or whether they are simply "parroting" something that they have scraped from an article or a web page.
- They provide "static data", whereas "full-featured" corpus sites like English-Corpora.org and Corpusdelespanol.org allow us to see and use links between different words, phrases, and constructions
- Most importantly, LLMs do not allow us to "check the data" (via KWIC entries, metadata, etc) in the same way that we can with structured corpora.

At the end of the day, it is not an either/or proposition (either LLMs *or* structured corpora). LLMs are best used *in conjunction with* reliable corpus data. Corpus linguists can make use of the rich lexical data from LLMs, and AI/ML researchers can use corpus data for fine-tuning, distillation, and Retrieval Augmented Generation (RAG) with LLMs.

From Big Data to Smart Data: Building better datasets for Human-Centric AI with meaning in mind

Rebekah Wegener

Paris Lodron University Salzburg

Recent developments in artificial intelligence, particularly those surrounding large language models, have sparked a renewed interest in foundational questions about the nature of human language and how machines process and generate language. These questions echo Halliday's (2003) early insights about language as meaning potential and what this means for computational approaches to language. However, these advances also highlight fundamental questions about meaning, context, and the relationship between quantity and quality of data. As Dingemanse and Liesenfeld (2022) argue, creating more representative and meaningful datasets requires going beyond text collection to capture the complexity of human communication.

To explore these questions further, I want to consider human-centric AI systems, focusing specifically on how such systems are designed and built in industry and academia. These systems have explicit requirements for understanding meaning making in context, for understanding abstract concepts such as importance and for understanding multimodal interaction. Drawing on previous work (e.g. Cassens & Wegener 2018 and Wegener, in press), I will demonstrate how such systems showcase the importance of high quality datasets for AI - particularly human-centric AI - and show how strong theoretical frameworks can inform the design of "smart" datasets that capture the complexity of human meaning-making.

Such endeavours are not without their challenges, and in many respects these challenges mirror long-standing questions in corpus linguistics about context, annotation, and the nature of meaning itself. These problems become particularly apparent when working with multimodal data, where meaning emerges not just from individual modes, but from their integration and interaction in context (Bateman, Wildfeuer & Hiippala, 2017; O'Halloran, Tan & Wignell, 2019). The development of tools and methods for handling such complexity requires careful consideration of both theoretical and practical concerns.

While some of these questions are destined to remain core philosophical debates, other are the driving force behind new tool development. Such tools provide opportunities for addressing methodological challenges within corpus linguistics and in particular, hold the potential to assist in the study of meaning. I will briefly consider how new tools can be recruited for corpus linguistics and how human-centric AI can also benefit from tools and methods already popular within corpus linguistics (Driess et al. 2023; Henlein et al. 2024).

As Dingemanse & Liesenfeld (2022) argue, "corpora represent an important and mostly untapped resource for language technology. " Understanding meaning and the process of meaning making requires more than just collecting large amounts of data - amongst many other things, it requires theoretically informed approaches to dataset design, representation, annotation and analysis. For meaning-focused corpus research, "data comes in levels of granularity. A well-curated corpus...harbour(s) important insights about human interactional infrastructure" (Dingemanse & Liesenfeld 2022).

References

Bateman, J., Wildfeuer, J., and Hiippala, T. (2017). *Multimodality: Foundations, Research and Analysis – A Problem-Oriented Introduction*. Berlin: Mouton de Gruyter.

Cassens, J., and Wegener, R. (2018). Supporting Students Through Notifications About Importance in Academic Lectures. In *Proceedings of AmI 2018 -- International Joint Conference on Ambient Intelligence*, pages 227-232, Springer, Larnaca, Cyprus LNCS, 2018.

Dingemanse, M., and Liesenfeld, A. (2022). From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) 5614–5633 (Association for Computational Linguistics, 2022).

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., ... and Florence, P. (2023). Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378. (Google Brain)

Halliday, M.A.K. (2003). On the Architecture of Human Language. In J. Webster (ed.). *On Language and Linguistics. Volume 3: The collected works of M.A.K. Halliday*. London: Continuum, 2003.

Henlein, A., Bauer, A., Bhattacharjee, R., Ćwiek, A., Gregori, A., Kügler, F., ... and von Eiff, C. I. (2024, June). An outlook for AI innovation in multimodal communication research. In *International Conference on Human-Computer Interaction* (pp. 182-234). Cham: Springer Nature Switzerland.

O'Halloran KL, Tan S, and Wignell P. (2019). SFL and Multimodal Discourse Analysis. In Thompson G, Bowcher WL, Fontaine L, and Schönthal D., (eds.). *The Cambridge Handbook of Systemic Functional Linguistics*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, pages 433-461.

Wegener, R. (in press). *Multimodal Interaction and Technology in Context: The Semiotic Machine*. Routledge.

ROUND TABLES / MESAS REDONDAS

Entre datos y discurso: IA y análisis lingüístico para combatir desinformación, odio y radicalización en redes sociales

Convenor / Moderadora: Encarna Hidalgo Tenorio (Universidad de Granada)

En esta mesa redonda, crearemos un espacio abierto para examinar cómo la inteligencia artificial y el análisis del discurso contribuyen a detectar y mitigar la desinformación y los discursos extremistas en redes sociales. Presentaremos investigaciones sobre detección automática de noticias falsas, narrativas xenófobas sobre migración, discurso incel en español y análisis de palabras clave en la radicalización de la célula terrorista del 17-A. A continuación, se ofrecen los resúmenes extendidos de cada intervención, en los que la noción de "aboutness" resulta fundamental de diferente forma.

- **Inteligencia artificial en la detección de noticias falsas**

Juan Luis Castro Peña (Universidad de Granada)

Detectar noticias falsas en redes sociales es fundamental para garantizar la fiabilidad de la información difundida. El aprendizaje automático (AA) se posiciona como la técnica principal para evaluar en tiempo real el contenido y detectar desinformación mediante el análisis de grandes volúmenes de datos. En este estudio entrenamos distintos modelos de AA con bases de datos públicas de noticias falsas en español, obteniendo altos niveles de precisión según métricas estándar de validación. Sin embargo, al evaluar estos modelos con noticias externas a los conjuntos de entrenamiento, la eficacia disminuye, aunque sigue siendo aceptable. Estos resultados evidencian la complejidad de detectar desinformación utilizando únicamente el texto aislado, dado el sesgo inherente en los datos de entrenamiento y la influencia de factores externos no reflejados en la noticia. Concluimos que, si bien el AA ofrece una herramienta poderosa para mitigar la propagación de noticias falsas, se requieren modelos más sofisticados (por ejemplo, híbridos, dinámicos y basados en análisis de discurso avanzado o aprendizaje por refuerzo) para reducir sesgos y mejorar su robustez.

- **Discurso incel en X: Un análisis con topic modeling en contextos hispanohablantes**

Aritz Gorostiza Cerviño (Universidad de Málaga)

Los incel (involuntary celibates) son una subcultura masculina con ideología misógina y racista ampliamente estudiada en contextos angloparlantes, pero poco investigada en el ámbito hispanohablante. Este estudio analiza el discurso de líderes de opinión incel en X (Twitter) entre el 26 de enero de 2023 y el 10 de febrero de 2024, recopilando 10.581 publicaciones. Mediante Topic Modeling (LDA) y un riguroso filtrado de palabras, se identificaron 24 temas organizados en nueve categorías principales: misoginia, hombres, inmigración, política, comunidad incel, fútbol, redes sociales, relaciones sexoafectivas y expresiones humorísticas/metafóricas. Los hallazgos revelan patrones discursivos distintivos en el contexto hispanohablante y permiten comparar sus énfasis temáticos con los observados en estudios previos. Este análisis contribuye a comprender mejor la

naturaleza y particularidades del discurso incel en español, ofreciendo una base empírica para futuras investigaciones sobre radicalización en línea y estrategias de mitigación del discurso de odio.

- **Análisis de palabras clave en el discurso de la radicalización**

Encarna Hidalgo Tenorio (Universidad de Granada)

En agosto de 2017, un grupo de jóvenes de Ripoll llevó a cabo dos atentados en Barcelona y Cambrils, que causaron la muerte de 16 personas y dejó más de un centenar de heridos. En un intento de entender cómo se llegó hasta ahí, este trabajo explora la conducta verbal de la célula terrorista del 17-A a través de los textos hallados en sus ordenadores y dispositivos móviles. Adoptando una perspectiva de análisis del discurso basado en corpus, nuestro objetivo principal es determinar sus características discursivas. Nuestro interés no sólo se centra en lo que dicen, a quién se dirigen, de quién hablan, y cuáles son las reacciones que se infieren de ello, sino también, y, sobre todo, en cómo lo hacen. Mostramos los resultados de un primer acercamiento a los datos de naturaleza inductiva. A fin de tener una visión panorámica de esos materiales, hemos hecho uso de una de las herramientas imprescindibles en la lingüística de corpus, a saber, el análisis de palabras clave presentes en los mismos, ya fueran audios, mensajes cortos o conversaciones de WhatsApp.

CONSTRIDIOMS y CREA-CONSTRIDIOMS: Construcciones fraseológicas del español y del alemán"

Convenor / Moderadora: Carmen Mellado (Universidade de Santiago de Compostela)

Pedro Ivorra Ordines (Universidad de Zaragoza)
Nely M. Iglesias Iglesias (Universidad de Salamanca)
Maricel Esteban-Fonollosa (Universitat de València)
Caterina Chinellato (Universidade de Santiago de Compostela)

En nuestra **mesa redonda**, el debate se articulará en torno a tres ejes temáticos principales:

1. Presentación de nuestro proyecto de investigación **CONSTRIDIOMS** y de la plataforma asociada.
2. Potencial de aplicación didáctico-metdológica de la plataforma.
3. La construcción del significado y el concepto de creatividad lingüística: **CONSTRIDIOMS** y **CREA-CONSTRIDIOMS**.

Nuestro proyecto de investigación **CONSTRIDIOMS** (*Gramática de Construcciones y Fraseología. Las construcciones fraseológicas del alemán y el español en contraste a través de los corpus*, PID2019-108783RB-100) tiene por objeto el análisis contrastivo de las construcciones fraseológicas en alemán y español. Para ello, nos planteamos los siguientes objetivos:

1. **Elaboración de un corpus de construcciones fraseológicas** o **semiesquemáticas** en alemán y sus equivalentes en español, con especial atención a las construcciones de naturaleza intensificadora.
2. **Descripción holística de las construcciones fraseológicas** objeto de estudio mediante el análisis de su comportamiento en grandes corpus comparables del español y el alemán: básicamente *deTenTen20* y *esTenTen18* de *Sketch Engine* (https://www.sketchengine.eu/).
   Se han considerado tanto los datos cuantitativos (frecuencia *token* y *type*) como sus características estructurales y pragmáticas, con énfasis en el potencial ilocutivo de cada construcción. Destacan, además, los fenómenos de fijación cognitiva (*entrenchment*) y productividad, determinados a partir de diferentes tipos de frecuencia y el número de *hápax* en los *slots*. En el análisis, el continuum léxico-gramática juega un papel crucial, especialmente en aquellas construcciones licenciadas tanto por locuciones como por instancias de baja frecuencia *token*.
3. **Profundización en el concepto de creatividad lingüística**, analizando el funcionamiento de las construcciones más productivas y el surgimiento de nuevas construcciones fraseológicas o semiesquemáticas.
4. **Desarrollo de una metodología contrastiva basada en corpus paralelos** del alemán y el español (*PaGeS*: https://www.corpuspages.eu/), aplicando el *método contrastivo unilateral*. Este método ha permitido describir por niveles las construcciones fraseológicas del alemán y sus equivalentes en español, facilitando la identificación de relaciones interconstruccionales y, por tanto, la definición de familias de construcciones con significados pragmáticos afines.
5. **Identificación de relaciones verticales y horizontales** entre las distintas construcciones analizadas dentro del conjunto del *constructicón* del alemán y el español.
6. **Creación de una plataforma en línea de acceso gratuito**, en la que se presentan las construcciones analizadas. Se destaca especialmente la consideración del potencial ilocutivo, es decir, el significado pragmático de cada una de las unidades lematizadas. Asimismo, la plataforma introduce un enfoque innovador en la

presentación de los datos relativos a los elementos (*slot-fillers*) que ocupan los diferentes *slots* de las construcciones.

En definitiva, la plataforma **CONSTRIDIOMS** ofrece una aproximación novedosa y contrastiva al estudio de las construcciones fraseológicas en español y alemán, proporcionando herramientas de análisis avanzadas y recursos aplicables tanto a la investigación lingüística como a la didáctica y metodología de la lengua en general y de las lenguas extranjeras en particular. El proyecto contribuye a un conocimiento más profundo del lenguaje (*construcicón* mental), prestando especial atención a los procesos cognitivos subyacentes a la construcción del significado en contexto – tema central de CILC 2025.

Referencias

Esteban-Fonollosa, M. (2023). *El inútil de su hijo*: análisis de la construcción intensificadora [DET$_{\{det\}}$ ADJ *de* SN]. *Romanica Olomucensia*, *35*(2), 299-311. https:/dx.doi.org/10.5507/ro.2023.023

Esteban-Fonollosa, M., y Holzinger, H. (2025). Die Darstellung von UWV in Lehrwerken der Niveaustufen A1–C1. Vorschläge für neue Wege im Licht von Korpusanalyse und Konstruktionsgrammatik. *Deutsch als Fremdsprache* 1, 3-14.

Holzinger, H., e Iglesias Iglesias, Nely M. (en prensa). Musterhaftigkeit und Kreativität: Zwei gegenläufige Tendenzen im Sprachgebrauch. En Mansilla, Ana / Strohschen, Carola (eds.): *Formen der Interkulturalität und Mehrsprachigkeit im Kontext der Germanistik*. Berlin et al.: Peter Lang.

Iglesias Iglesias, Nely M. (2021). Produktivität und Kreativität sprachlicher Muster. Am Beispiel der Phrasemkonstruktion [DET *nächste* N *kommt bestimmt*]. *Beiträge zur Fremdsprachenvermittlung*, Sonderheft 28: 21-40.

Iglesias Iglesias, Nely M., y López Meirama, B. (2024). "La expresión hiperbólica de las sensaciones en español y alemán: análisis de las construcciones fraseológicas [morir(se) de (ART) S$_{sing\{sensación\}}$] y [vor N$_{Sg\{Gefühlsempfindung\}}$ sterben]". *Quaderns de Filologia: Estudis Lingüístics* XXIX: 157-176. doi: 10.7203/QF.29.28891. https://turia.uv.es/index.php/qfilologia/article/view/28891

Ivorra Ordines, P. (2024). *Un hambre que da calambre*. Creativity and extravagance in the context of a family of consecutive constructional idioms. *Cognitextes. Special Issue Constructions and Context(s)*.

Ivorra Ordines, P. (en prensa). Productivity and creativity of constructions. En X. Wen and Ch. Sinha (eds.), *The Cambridge Encyclopedia of Cognitive Linguistics*. Cambridge: Cambridge University Press.

Ivorra Ordines, P., y López Meirama, B. (2024). *Vete a freír cristales*. The interplay of convention and innovation in a constructional idiom of rejection in Spanish. *Review of Cognitive Linguistics*. 1-38.

López Meirama, B., e Iglesias Iglesias, Nely M. (2023). The construction [*a todo* N$_{sing}$] in Spanisch. En Wiesinger, Evelyn / Hennecke, Inga (eds.): *Constructions in Spanish*. Amsterdam: John Benjamins, 129-153.

Mansilla Pérez, A. (2024). Lingüística de corpus y fraseología: El patrón [PREP+ S {arbitrariedad}] en las combinaciones usuales a su antojo, a su gusto, a voluntad. *Revista Signos. Estudios de Lingüística* 57 (114)

Mellado Blanco, C. (2020). *Romanica Olomucensia*. *Número monográfico Nuevas aportaciones de la Gramática de Construcciones a los estudios de fraseología en las lenguas románicas* 32/1. https://romanica.upol.cz/magno/rom/2020/mn1.php

Mellado Blaco, C. (2023). "*Wie Fisch und Fahrrad*. Inkongruenz als konstitutives Merkmal der verneinenden Vergleichskonstruktionen". In: Mollica, F. / Stumpf, S. (Hrsg.): *Konstruktionsgrammatik IX. Konstruktionsfamilien im Deutschen*. Tübingen: Stauffenburg Linguistik, 165-203.

Mellado Blanco, C. (2024a). "*El tiempo vuela*: Metáforas cognitivas y fraseologismos con dominio meta TIEMPO en español y alemán". *Phrasis* 7, 53-81.

Mellado Blanco, C. (2024b). „The ways of phraseology are mysterious: Humour and snowclones in Spanish and German Bibleisms from a Construction Grammar perspective". In: Saša Babič, Anna T. Litovkina, Fionnuala Carson Williams and Christian Grandl (eds.): *"Standing on the shoulders of giants". A Festschrift in honour of Wolfgang Mieder on the occasion of his 80th birthday*. *Proverbium Online Supplement* 3: 2024. Osijek: Faculty of Humanities and Social Sciences at the University of Osijek, 495–512. https://naklada.ffos.hr/knjige/index.php/ff/catalog/book/18. **2024**

Mellado Blanco, C. (in press). Constructional idioms. En X. Wen and Ch. Sinha (eds.), *The Cambridge Encyclopedia of Cognitive Linguistics*. Cambridge: Cambridge University Press.

Mellado Blanco, C., e Iglesias Iglesias, Nely M. (2022). Traducir y descubrir construcciones. En Iris Holl / Beatriz de la Fuente, Marina (eds.): *La traducción y sus meandros: diversas aproximaciones en el par de lenguas alemán-español*. Salamanca: Universidad de Salamanca, 361-378. https://doi.org/10.14201/0AQ0320361378

Mellado Blanco, C., e Iglesias Iglesias, Nely M. (2024). "*In aller Kürze*: die Konstruktion [*in* ALL Nabst] im gesteuerten DaF-Unterricht". *Deutsch als Fremdsprache* 61(2): 67-80.

Mellado Blanco, C., Holzinger, H., Iglesias Iglesias, N., y Mansilla Pérez, A. (eds.) (2020). *Muster in der Phraseologie. Monolingual und kontrastiv*. Hamburgo: Dr. Kovač.

Mellado Blanco, C., Ivorra Ordines, P., y Esteban Fonollosa, M. (2024). Pasado y futuro de la(s) Gramática(s) de Construcciones y los límites de las construcciones. *Quaderns De Filologia - Estudis Lingüístics, 29*, 9–21. https://doi.org/10.7203/QF.29.29771

Mellado Blanco, C., López Meirama, B., Losada Aldrey, M. C. (2020). *LinRed. Lingüística en la Red*. Número monográfico XVII (2020), *Modelos de análisis en la fraseología de las lenguas europeas*. http://www.linred.es/numero17_monografico.html

Corpus-based approaches to meaning in historical sociolinguistics

Convenors / Moderadores:
Carolina Amador-Moreno (Universidad de Extremadura),
Javier Ruano-García (Universidad de Salamanca)

In the past few years, meaning has gained renewed attention in historical sociolinguistics. In line with current third-wave approaches to linguistic variation (Eckert 2012, 2018), recent historical studies have considered "the social meaning of linguistic variation as an intrinsic feature of language, understanding variation as a social semiotic system which conveys the entire spectrum of social issues within a community" (García-Vidal 2023: 3). In fact, as García-Vidal (2023: 5) goes on to explain, variation is conceived not only as a reflection but also as a construction of social meaning by means of deliberate linguistic choices in unfolding discourse (see also Schilling 2013). In this context, key sociolinguistic concepts such as identity and ideology have taken centre stage, as they have been revisited as well as reassessed through the lens of key frameworks in this constructionist and speaker-oriented approach. Research on intra-speaker/writer variation (e.g. Hernández-Campoy and García-Vidal 2018; Oudeluijs and Yáñez-Bouza 2023), enregisterment (e.g. Amador-Moreno and Ruano-García 2023; Schintu 2023), and communities of practice (e.g. Conde-Silvestre 2016; Alcolado-Carnicero 2023), for example, have convincingly demonstrated that, despite the pervasive bad data problem (Labov 1994: 11), new light can be shed on how meaning was negotiated, crafted and circulated in the past. The available evidence has indeed shown that the combination of such third-wave sociolinguistic perspectives with corpus-based methodological approaches proves useful to uncover historical patterns of variation that inform our understanding of speakers' linguistic choices in their construction of identity. Needless to say, the increasing availability of new historical materials that lend themselves to third-wave sociolinguistic readings—e.g. private correspondence and various forms of speech representation—have contributed in significant ways to clarifying an otherwise blurred picture.

This roundtable explores the potential of historical sociolinguistic approaches to various text types such as letters and dialect writing. Combining different theoretical frameworks and research methods, the contributions explore third-wave historical sociolinguistic issues that place an emphasis on identity marking, social status and ideolog(ies), intra-writer variation and the role of the addressee, identity across social networks, and the identification of semantic fields that allow us to explore linguistic choices. The papers focus on spelling, formulaic language, specific dialectal features and the expression of emotions. By looking at dialect representation in writing and a variety of other linguistic features in letter writing, the papers listed below employ different corpus-based approaches to meaning from a historical sociolinguistic perspective.

- **Unveiling intra-writer variation in English historical letters: Upward and downward accommodation patterns in social interaction**

Tamara García-Vida (UNED)

This presentation addresses sociolinguistic meaning by examining intra-writer variation as addressee-based accommodation patterns in historical letters. Two case studies illustrate this phenomenon. Drawing on the *Paston Letters* (1425-1503), the first study analyses how male Paston family members adopted the innovative orthographic variable <th> over <þ> when addressing recipients of higher social ranks. The second study

examines intra-writer variation through the use of comparative forms in historical letters, using data from the *Parsed Corpus of Early English Correspondence* (PCEEC, 1410–1681) and the *Corpus of Early English Correspondence Extension* (CEECE, 1700–1800). Writers favoured synthetic comparatives with short, Germanic adjectives for lower-status recipients and analytic comparatives with long, Romance adjectives for higher-status recipients, reflecting addressee-based accommodation.

- **Identity and sociability in Late Modern English correspondence**

Nuria Yáñez-Bouza (Universidade de Vigo)

The Georgian period in Britain was a time of significant sociocultural, literary and linguistic change, during which increasing social mobility heightened individuals' awareness of identity within and across social networks. Letter writing functioned as a key social practice, and letter-writing manuals served as resources for disseminating linguistic and stylistic conventions. Framed within historical sociopragmatics and third-wave historical sociolinguistics, this paper examines the conventions transmitted in manuals of the long eighteenth century (*c.*1650–1800) and the sociopragmatic nuances they conveyed in constructing, shaping and negotiating meaning and identity in Georgian polite society. The case study focuses on the use of closing formulae (e.g. *yours*, *your servant*), analysing a dataset of 2,500 model letters with attention to variables such as theme, gender, social distance and rank.

- **"I hate Home Rule & the women here are all Home Rulers": Investigating identity and ideology in nineteenth and twentieth-century Irish private correspondence**

Nancy E. Ávila Ledesma, David Sotoca Fernández, Carolina Amador Moreno (Universidad de Extremadura)

This study investigates ideological issues contained in the private discourse of Irish emigrant letters written in the nineteenth century. The study presents a corpus-based analysis of a selection of letters exchanged between Irish emigrants to the USA. It turns to LIWC-22 in combination with SketchEngine to explore how letter writers position themselves regarding political issues in this dataset in order to shed light on the intersections between political ideology, personal identity and social conflicts in the nineteenth and twentieth century. LIWC-22 is used to perform an initial quantitative analysis of the data and to identify the main linguistic patterns that are later explored qualitatively in the study.

- **Meaning-making, indexicality and enregisterment: Evidence from nineteenth-century dialect writing**

Paula Schintu, Javier Ruano-García (Universidad de Salamanca)

Literary representations of dialect can be read as a metadiscursive practice in the typification of dialect and identity. They open windows not only into how non-standard varieties were perceived in the past, but also into the social meanings they indexed and how such meanings became linked to them. In this presentation we look at nineteenth-century dialect writing from the lens of indexicality and enregisterment. We undertake a corpus-based quantitative and qualitative approach to specimens of late modern Lancashire and Derbyshire dialects taken from *The Salamanca Corpus* in an attempt to cast light on how these varieties and the meanings associated with them were shaped and

represented. By exploring writers' linguistic choices in the representation of regional sounds, we show that dialect writing relied on enregistered linguistic repertoires to (re)create contemporary models of linguistic behaviour, while it played an active part in the dissemination of sociolinguistic meaning and the (re)construction of identity.

References

Alcolado-Carnicero, J. M. (2023). A community of practice in the mercers of the City of London. *International Journal of English Studies*, *23*(2), 89-115.

Amador-Moreno, C., & Ruano-García, J. (2023). Linguistic perceptions of Irish English in nineteenth-century emigrant letters. *International Journal of English Studies*, *23*(2), 41-63.

Conde-Silvestre, J. C. (2016). A 'third-wave' historical sociolinguistic approach to late Middle English correspondence: evidence from the Stonor Letters. In C. Russi (Ed.), *Current Trends in Historical Sociolinguistics* (pp. 46–66). Warsaw/Berlin: Open De Gruyter.

Eckert, P. (2012). Three waves of variation study: the emergence of meaning in the study of variation. *Annual Review of Anthropology*, *41*, 87–100.

Eckert, P. (2018). *Meaning and Linguistic Variation*. Cambridge UP.

García-Vidal, T. (2023). Contextualising third-wave historical sociolinguistics. *International Journal of English Studies*, *23*(2), 1-14.

Hernández-Campoy, J. M. & García-Vidal, T. (2018b). Style-shifting and accommodative competence in late Middle English written correspondence: putting audience design to the test of time. *Folia Linguistica Historica*, *39* (2), 383–420

Labov, W. (1994). *Principles of Linguistic Change I: Internal Factors*. Oxford: Blackwell.

Oudeluijs, T., & Yáñez-Bouza, N. (2023). Constructing identities and negotiating relationships in late eighteenth-century England. *International Journal of English Studies*, *23*(2), 15-40.

Schilling, N. (2013) Investigating stylistic variation. In J.K. Chambers & N. Schilling (Eds.), *The Handbook of Language Variation and Change* (2nd ed.) (pp. 327–349). Oxford: Blackwell.

Schintu, P. (2023). Dialect in the making: A third-wave sociolinguistic approach to the enregisterment of Late Modern Derbyshire spelling. *International Journal of English Studies*, *23*(2), 65-87.

Redefining "aboutness": The role of NLP and AI in corpus linguistics

Convenor / Moderadora: Chantal Pérez Hernández (Universidad de Málaga)

Carla Fernández Melendres (Universidad de Málaga)
Javier Fernández Cruz (Universidad de Málaga)
María García Gámez (Universidad de Málaga)

In the 1980s, the availability of electronic text led to a paradigm change in linguistic research, transforming the way language samples were analyzed through word counting and computer-aided qualitative observation, marking the era of concordances and collocates. With the rise of the Internet, new data scraping methods and analytical techniques like regression and clustering spread across disciplines, elevating the prestige of data-driven careers and studies. Nowadays, linguistics has undergone a new paradigm change, incorporating data science principles where words are not only simply counted but represented as complex numerical matrices, enabling deeper exploration and interpretation. Large Language Models (LLMs) further advance this shift, offering new ways to process and analyze language through advanced machine learning techniques.

In this evolving landscape, *Corpus Sense* (Moreno-Ortiz 2025) is the first of a new generation of corpus query tools that integrates quantitative, qualitative, and AI features to facilitate comprehensive text analysis. This combination of advanced tools allows for a deeper exploration of *aboutness*, a central concern in corpus linguistics whose study is essential for understanding how language functions and how people use it to communicate while presenting significant methodological challenges for its detection and analysis. Until now, keyword extraction is the one tool that aims at providing a quick overview of the contents of a corpus. Keywords are meant to capture the aboutness of a document or set of documents (Scott & Tribble 2006; Mahlberg 2007; Bondi 2010; Marchi 2018). Keywords are identified by comparing word frequency of a focus corpus with a reference corpus, and assigning a keyness score to each word, thus equating keyness to aboutness (Moreno-Ortiz 2024a). *Corpus Sense* goes beyond this by providing a number of sophisticated tools aimed at the exploration of content. The *Topics* tool helps users explore the main themes within a corpus, while *Insights* leverages LLMs to analyze specific aspects of the content, such as style, readability, or emotions, thereby offering a deeper understanding of the aboutness of the texts. Additionally, *Corpus Sense* also extracts lists of keywords, but it uses a graph-based approach that does not require the use of a reference corpus, and therefore words are not ranked by comparison, which is biased by definition, but by their co-occurrence within the corpus itself.

Language-related AI techniques, such as word embeddings and LLMs, can play a crucial role by making it easier than ever for researchers to dive into large corpora. However, the interpretation of the results still requires the expertise of linguists. Given the challenges that LLMs face, such as alignment, bias, and efficiency (Shen et al. 2024; Yang et al. 2023; Wan et al. 2023), the results generated should only be considered accurate with careful analysis by language experts. Hence, while technology continues to advance, the essential role of the linguist in interpreting, refining, and contextualizing these results remains as important as ever (Moreno-Ortiz 2024). Novel tools like *Corpus Sense* empower linguists by providing sophisticated features that enhance their research capabilities, fostering a productive synergy between human expertise and machine learning.

This roundtable explores the potential of advanced NLP techniques and AI in corpus linguistics. Panelists will showcase the many novel features that *Corpus Sense* has to offer: keyword extraction, entity recognition, semantic search, topic modeling, and LLM insights. This discussion will emphasize the dual role of LLMs as analytical tools and subjects of

inquiry, proposing innovative methodologies for corpus linguistics and reshaping our understanding of aboutness.

References

Bondi, M. (2010). Perspectives on keywords and keyness: An introduction. In M. Bondi & M. Scott (Eds.), *Keyness in Texts* (pp. 1–18). John Benjamins Publishing Company. https://doi.org/10.1075/scl.41.01bon

Mahlberg, M. (2007). Corpus stylistics: Bridging the gap between linguistic and literary studies. *Text, Discourse and Corpora: Theory and Analysis*, *8*, 219–246.

Marchi, A. (2018). Dividing up the data: Epistemological, methodological and practical impact of diachronic segmentation. In C. Taylor & A. Marchi (Eds.), *Corpus Approaches to Discourse*. Routledge. https://doi.org/10.4324/9781315179346

Moreno-Ortiz, A. (2024a). Keywords. In A. Moreno-Ortiz (Ed.), *Making Sense of Large Social Media Corpora: Keywords, Topics, Sentiment, and Hashtags in the Coronavirus Twitter Corpus* (pp. 59–102). Palgrave Macmillan Cham. https://doi.org/10.1007/978-3-031-52719-7_4

Moreno-Ortiz, A. (2024b). The linguist's role in sentiment analysis: From knowledge provider to data annotator. In S. Maci & G. Garofalo (Eds.), *Investigating Discourse and Texts* (pp. 25-54). Peter Lang. https://doi.org/10.3726/B21393

Moreno-Ortiz, A. (2025). *Corpus Sense* (Version 1.0) [Python 3]. Universidad de Málaga. https://corpus-sense.app/

Scott, M., & Tribble, C. (2006). Textual Patterns. Key Words and Corpus Analysis in Language Education. In *Studies in Corpus Linguistics. 22*. John Benjamins Publishing Company.

Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., Zheng, B., Koehn, P., & Khashabi, D. (2024). *The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts*. arXiv. https://doi.org/10.48550/arXiv.2401.13136

Wan, Z., Wang, X., Liu, C., Alam, S., Zheng, Y., Liu, J., Qu, Z., Yan, S., Zhu, Y., Zhang, Q., Chowdhury, M., & Zhang, M. (2024). *Efficient Large Language Models: A Survey*. arXiv. https://doi.org/10.48550/arXiv.2312.03863

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*. arXiv. https://doi.org/10.48550/arXiv.2304.13712

WORKSHOPS / TALLERES

Using UAM Corpustool for linguistic studies of small corpora

Michael O'Donnell

Universidad Autónoma de Madrid

This workshop will provide basic training in the use of UAM Corpustool, a system for manual and automatic annotation of multi-text corpora, search over annotations, and statistical studies over the annotations. The tool is intended for small corpus studies (< 5 million words). It is intended to support research studies into linguistic patterns in self-collected corpora. As such, it is not optimised for searching large corpora. This workshop will explore the use of the tool to analyse political discourse.

In part 1, attendees will be guided through setting up an account on the online version (free), and establishing a new project (adding texts, and specifying annotation layers). In part 2, attendees will use the tool for manual annotation of one text. In part 3, various automatic annotation layers offered by the software will applied. Part 4 will explore the statistical reports that can be generated by the software.

Corpus Sense: A next-generation tool for advanced corpus and discourse analysis

Antonio Moreno Ortiz

Universidad de Málaga

Corpus Sense is a corpus query tool that incorporates advanced functionalities not available in existing applications. It is specially designed for content and discourse analysis, although it also features functionalities commonly found in other corpus tools, combining traditional and AI features to offer users a unique set of tools that is able to easily obtain useful insights from a corpus with minimal effort.

Corpus Sense is a web application designed to work with small to medium-sized corpora (up to 2.5 million tokens). It offers user management capabilities, including corpus upload and sharing (public corpora), and support 22 languages, although no support is given to multilingual corpora.

Unlike other corpus tools, Corpus Sense uses an NLP framework (spaCy) to process and query corpora, and creates word embeddings using Transformers, which are used for several of the content analysis features. The application is, in general, extremely responsive for common operations, such as searching, and offers advanced, NLP-based features, such as graph-based keyword extraction (thus eliminating the need for a reference corpus), named entity recognition and labeling, and literal, pattern-based, and semantic search.

Additionally, state-of-the-art topic modeling is offered by means of a user-friendly interface to BERTopic, producing descriptive, readable topic labels. Another defining, advanced feature, tentatively named "Insights", leverages the power of large language models to produce high-quality descriptions of a variety of aspects of the contents of a corpus. These aspects include rhetorical style and devices, discourse markers, readability, sentiment, emotions, hate speech detection, etc. Unlike most other AI tools, Corpus Sense

does not make use of external LLMs through an API; instead, it runs a freely available multilingual LLM locally. Because of the multilingual capabilities of this LLM, insights can be generated in many languages, regardless of the corpus language.

PARALLEL SESSIONS / SESIONES PARALELAS

Panel 1

Corpus design compilation and types
Diseño, elaboración y tipología de corpus

Corpus oral del léxico del ecuavóley en Quito, Ecuador

Mary Jeanneth Gutiérrez Guarderas

Universidad de Salamanca

Ecuador cuenta con el Corpus del habla del Ecuador (CORPHA), un corpus textual elaborado por la Academia Ecuatoriana de la Lengua (AEL), que documenta la variedad del habla ecuatoriana, a partir del siglo XX, con documentos procedentes de textos literarios, periodísticos y académicos (AEL, s.f.). Sin embargo, existe la necesidad de desarrollar corpus orales que registren la variación social, geográfica y estilística del habla ecuatoriana. Por tal motivo, el objetivo general de este proyecto es la creación de un corpus oral, dialectal y sincrónico del léxico del ecuavóley (adaptación ecuatoriana del voleibol) en Quito, con la finalidad de documentar el funcionamiento de esta variedad lingüística y que sirva como recurso de investigación de la lengua y a otras áreas de estudio (historia, sociología, estudios culturales, etc.).

Después del fútbol, el ecuavóley es el segundo deporte más practicado en el Ecuador. Esta práctica no es solamente un deporte, sino también es un espacio que refleja la identidad ecuatoriana. En el campo de la lingüística, evidencia la variedad dialectal del habla ecuatoriana mediante frases como "para vos mamita", con la que un jugador anticipa al equipo contrario la ganancia con la siguiente jugada; o "esas manos de mantequilla", una frase que los espectadores utilizan para intimidar a los jugadores (Guayga, 2013).

Para alcanzar los objetivos propuestos, es necesario utilizar una metodología científica de índole práctica. De esta manera, en la primera fase del proyecto se realizarán aproximadamente 100 entrevistas sociolingüísticas a jugadores, jueces y espectadores del ecuavóley. Las grabaciones serán exclusivamente orales y se realizarán en los espacios donde se practica este deporte. Para obtener un corpus representativo y que refleje la heterogeneidad social de los informantes se dividirá a Quito en 10 zonas territoriales, de acuerdo a la importancia y representatividad que tienen en este deporte.

Una vez concluida la etapa de las entrevistas lingüísticas, se procederá con la transcripción ortográfica de las grabaciones. Para garantizar la homogeneidad en las transcripciones, se elaborará un Manual de transcripción que establezca criterios y parámetros para la transcripción y el etiquetado. Estas transcripciones y grabaciones constituirán el material de nuestro corpus. Estos serán alojados en una página web que facilitará las búsquedas automáticas, los análisis cuantitativos y la aplicación de métodos estadísticos. Cabe destacar que, para la construcción de este corpus, se tomará como referencia el Corpus Oral y Sonoro del Español Rural (COSER), por compartir características similares al propuesto: ofrece la posibilidad de acceder tanto a las transcripciones como a las grabaciones mediante la consulta simple o avanzada de un lema en la página web del corpus (COSER, s.f.).

Se prevé que el corpus crezca en el futuro e incluya muestras de habla de otras variedades lingüísticas de Ecuador. Por este motivo, es necesario que, una vez concluido

este proyecto, se evalúen y analicen los resultados con el objetivo de destacar las fortalezas y los puntos débiles.

Referencias

Academia Ecuatoriana de la Lengua. (s.f.). *Corpus del habla del Ecuador* (CORPHA). http://www.academiaecuatorianadelalengua.org/corpha-ec-2/

COSER. (s.f.). *Corpus Oral y Sonoro del Español Rural* (COSER). http://www.corpusrural.es

Guaygua Tumbaco, Óscar Enrique. 2013. *Crónicas del ecuavoley casos: El parque El Ejido y La Carolina de Quito, basados en la propuesta de los imaginarios urbanos*. [Tesis de licenciatura, Universidad Politécnica Salesiana, Sede Quito]. Repositorio Institucional Universidad Politécnica Salesiana, Sede Quito. https://dspace.ups.edu.ec/bitstream/123456789/5917/6/UPS-QT03892.pdf.

Creación de corpus multimodales con herramientas de IA en CQPweb

Rosa Illán Castillo

Universidad de Murcia

Esta comunicación presenta un flujo de trabajo integral para la construcción y el análisis de corpus multimodales, abordando desafíos clave relacionados con la precisión en la transcripción, la alineación video, audio y texto y la anotación lingüística. A diferencia de los corpus basados exclusivamente en texto, los corpus multimodales proporcionan información sobre cómo el lenguaje interactúa con las señales acústicas y visuales, permitiendo el estudio del lenguaje abarcando toda su complejidad.

El flujo de trabajo propuesto emplea los últimos avances en inteligencia artificial para mejorar y estos procesos. Whisper (Radford et al., 2022) y whisperX (Bain et al., 2023) garantizan una alta precisión en la transcripción y una alineación precisa a nivel de palabra, sincronizando los datos textuales con eventos visuales en el contenido de video. Esta sincronización permite a los investigadores analizar gestos co-verbales, expresiones faciales y otras señales no verbales en conjunto con estructuras lingüísticas de forma sencilla. Por su parte, spaCy (Honnibal et al., 2020) añade un nivel avanzado de procesamiento del lenguaje natural, ofreciendo anotaciones detalladas sintácticas, morfológicas y léxicas, que se analizan posteriormente mediante las capacidades de gestión y búsqueda en corpus de CQPweb (Hardie, 2012). En conjunto, estas herramientas crean un flujo de trabajo capaz de manejar grandes conjuntos de datos audiovisuales, mejorando algunas de las limitaciones de la investigación llevada a cabo en Uhrig (2018, 2021).

La efectividad del flujo de trabajo se demuestra mediante un estudio de caso con el corpus *NewsScape 2017*, una amplia base de datos audiovisuales compuesta por unas 30.000 horas de video recogidas en los medios de comunicación estadounidenses y desarrollada por el Red Hen Lab (Steen et al., 2018). El estudio de caso destaca cómo el flujo de trabajo permite realizar consultas multimodales, permitiendo a los investigadores explorar patrones lingüísticos en contextos audiovisuales. Por ejemplo, la alineación precisa de los datos textuales con segmentos de vídeo permite identificar instancias en las que se mencionan términos específicos mientras el hablante está visualmente presente, mejorando la calidad y especificidad de los análisis multimodales.

Una de las principales innovaciones de este flujo de trabajo es su extensibilidad. Entre las integraciones planeadas se encuentran herramientas de visión computacional,

análisis de trayectorias gestuales y detección de objetos, que permitirán capturar de manera sistemática los gestos co-verbales y los movimientos corporales en su contexto visual. Estas mejoras profundizarán nuestra comprensión de cómo interactúan el lenguaje y las señales no verbales en diversos contextos comunicativos. Las futuras extensiones también incluyen la incorporación de técnicas avanzadas de procesamiento de lenguaje. Estos desarrollos buscan habilitar búsquedas más matizadas y estudios más completos sobre la interacción entre las modalidades verbal y visual.

Esta metodología representa un avance significativo en la construcción de corpus multimodales, ofreciendo a los investigadores una solución fiable y precisa para analizar el lenguaje en su contexto comunicativo completo. Al integrar datos de audio, texto y video, este trabajo no solo aborda desafíos actuales, sino que también sienta las bases para futuras innovaciones en lingüística, ciencias cognitivas y el análisis computacional de la comunicación humana.

Referencias

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. ArXiv. https://arxiv.org/abs/2303.00747

Hardie, A. (2012). CQPweb – Combining power, flexibility and usability in a corpus analysis tool. International *Journal of Corpus Linguistics*, 17(3), 380–409. https://doi.org/10.1075/ijcl.17.3.04har

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. Zenodo. https://doi.org/10.5281/zenodo.1212303

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. ArXiv, abs/2212.04356. https://arxiv.org/abs/2212.04356

Steen, F. F., Hougaard, A., Joo, J., Olza, I., Pagán Cánovas, C., Pleshakova, A., Ray, S., Uhrig, P., Valenzuela, J., Woźny, J., & Turner, M. (2018). Toward an infrastructure for data-driven multimodal communication research. *Linguistics Vanguard*, 4(1). https://doi.org/10.1515/lingvan-2017-0041

Uhrig, P. (2018). NewsScape and the Distributed Little Red Hen Lab – A digital infrastructure for the large-scale analysis of TV broadcasts. In A. J. Zwierlein, J. Petzold, K. Böhm, & M. Decker (Eds.), *Anglistentag 2017 in Regensburg: Proceedings* (pp. XX–XX). Wissenschaftlicher Verlag Trier.

Uhrig, P. (2021). Large-scale multimodal corpus linguistics – The big data turn. [Postdoctoral dissertation (Habilitation), FAU Erlangen-Nürnberg].

Anglicism and keyword extraction in Spanish humanitarian texts

Loryn Isaacs, Pilar León Araúz

Universidad de Granada

Conceptual variation has been identified by humanitarian actors as an impediment to effective communication and the coordination of disaster response (Khan & Kontinen, 2022; Roepstorff, 2020). While English is the dominant language of communication for the humanitarian world, the challenges that conceptual variation poses are closely tied to the domain's highly interlingual nature. As such, variation in humanitarian discourse needs to be studied both within and across languages. Corpus linguistics offers a variety of methods to identify, track and measure such linguistic phenomena in specialized domains given

large quantities of text data. A principal task in this area of study is anglicism detection and keyword extraction, which provide essential data for identifying a domain's core concepts and usage patterns. To this end, we describe the processing of humanitarian-domain corpora in English and Spanish and the results of a new keyword and anglicism detection method.

In previous work, we developed a family of corpora of over two billion tokens from ReliefWeb, a United Nations database of humanitarian texts, which indicated that English, French and Spanish are the three top languages. The influence of English, by far the largest dataset, in the Spanish and French corpora is evident. In fact, the presence of English content in Spanish corpus documents hampers the isolation and quantification of anglicisms (i.e., unassimilated lexical borrowings: Álvarez-Mellado & Lignos, 2022). There is a small but significant presence of English-only content in the form of bibliographic references and monolingual passages that qualify as noise for studying lexical units which often have low frequencies. In this work we refine and reprocess the corpora to include sentence-level language identification results (employing Stanza, a neural package for natural language processing: Qi et al., 2020). These are used to establish subcorpora further isolating English-only content. Results indicate that filtering said content can have a large effect on anglicism frequencies: for example, roughly half of cases of "accountability" are not authentic uses of the English term in Spanish writing, but rather appearances in sections of English-only content.

We utilize the language-disaggregated subcorpora to then produce a list of anglicism candidates based on the extraction results of term grammars (Blahuš, et al., 2023) and by calculating keyness within the Spanish corpus against two reference corpora: first, the ReliefWeb English corpus, to identify lexical units shared by the English and Spanish corpora, and second, a newly produced corpus of Wikipedia's Spanish encyclopedia content, to rank units by thematic affiliation. As well as describing these results, the methods utilized and current limitations, we situate this work as part of a larger research project on humanitarian discourse and trends in the domain's conceptual variation. We conclude with a discussion of additional tasks to better quantify the appearance of anglicisms in Spanish humanitarian discourse and their place in its terminology.

References

Álvarez-Mellado, E., & Lignos, C. (2022). Detecting Unassimilated Borrowings in Spanish: An Annotated Corpus and Approaches to Modeling. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3868–3888). https://doi.org/10.18653/v1/2022.acl-long.268

Blahuš, M., Jakubíček, M., Cukr, M., Kovář, V., & Suchomel, V. (2023). Development of Evidence-Based Grammars for Terminology Extraction in OneClick Terms. In *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2023 Conference*, (pp. 650–662).

Khan, A.K., and Kontinen T. (2022). Impediments to localization agenda: humanitarian space in the rohingya response in bangladesh. *Journal of International Humanitarian Action* 7(14). https://doi.org/10.1186/s41018-022-00122-1

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 101–108).

Roepstorff, K. (2020). *Localisation and shrinking civic space: Tying up the loose ends.* Centre for Humanitarian Action (blog). https://www.chaberlin.org/wp-

content/uploads/2020/05/2020-05-publication-localisation-shrinking-civic-space-roepstorff-en.pdf

The appearance of the path guides the ſteps of the traveller": Adapting new findings while compiling more recent subcorpora in the Coruña Corpus project

Luis Puente-Castelo, Isabel Moskowich

Universidade da Coruña

The Coruña Corpus is a corpus of eighteenth and nineteenth century English scientific writing, composed of several subcorpora dealing with different scientific disciplines, all sharing the same design and principles of compilation. Up to this point, the subcorpora on Astronomy (Moskowich et al., 2012), Philosophy (Moskowich et al., 2016), History (Moskowich et al., 2019), Life sciences (Lareo et al., 2020), and Chemistry (Moskowich et al., 2022) have been completed and are available online, with several others in different stages of compilation and computerisation.

Our working method involves selecting our samples following a strict set of criteria, and then manually typing and XML-tagging these samples, before three rounds of manual revision and a final standardisation round. Experience shows that this long, thorough process helps consistently avoid the presence of any mistakes in the final versions of each of the subcorpora, thus assuring maximum faithfulness to the original.

This whole process is applied in a staggered manner, so that work is being carried out at once on several subcorpora at several stages: one subcorpus may be undergoing revision while another may be in the process of typing and tagging, and potential samples may be examined and selected for a third.

This staggered approach, which allows for more efficient and much faster work, naturally implies that our methodology has to be open to change: Since the start of the Coruña Corpus project, finding surprises or new elements that had not been considered before while working in a new subcorpus has been a common occurrence. These range from coming across the use of new linguistic elements which make us rethink and perhaps revise our editorial policy, to the implementation of new functions in the purpose-built tool for the Coruña Corpus, the Coruña Corpus Tool (Parapar & Moskowich 2007, Barsaglini-Castro and Valcarce 2020), and their side-effects on our samples.

The objective of this methodological paper is to discuss three of these unforeseen findings, and how the team has faced each of them. The paper will also discuss how opposing needs are balanced, and how the consequences of adding new features or changing policy on subcorpora which are further along the process of compilation, or even on those which have already been published, are addressed.

This also helps display what we think is a particular strength of smaller, highly-specialized, and labour-intensive corpora: identifying (and consequently correcting) problems during the process of compilation becomes much easier as a result of manual compilation.

References

Barsaglini-Castro, A., & Valcarce, D. (2020). The Coruña Corpus Tool: Ten Years On. *Revista de Procesamiento del Lenguaje Natural*, *64*, 13-19.

Lareo, I., Monaco, L.; Esteve-Ramos, M. J., & and Moskowich, I.(comps.) (2020). *Corpus of English Life Sciences Texts*. A Coruña: Universidade da Coruña.

Moskowich, I., Lareo, I., Camiña Rioboó, G., & Crespo, B.(comps.) (2012). *Corpus of English Texts on Astronomy*. Amsterdam: John Benjamins.

Moskowich, I., Lareo, I., Camiña Rioboó, G., & Crespo, B.(comps.) (2016). *Corpus of English Philosophy Texts*. Amsterdam: John Benjamins.

Moskowich, I., Lareo, I., Lojo Sandino, P., & Sánchez-Barreiro, E. (comps.) (2019). *Corpus of History English Texts*. A Coruña: Universidade da Coruña.

Moskowich, I., Puente-Castelo, L., & Monaco, L. (comps.) (2022). *Corpus of English Chemistry Texts*. A Coruña: Universidade da Coruña.

Parapar López, J. & Moskowich, I. (2007). The Coruña Corpus Tool. *Revista de Procesamiento del Lenguaje Natural*, *39*, 289–290.

Panel 2

Discourse, literary analysis and corpora
Discurso, análisis literario y corpus

Unveiling interactive narratives: A sentiment analysis of video games

Anabella Barsaglini-Castro

University of A Coruña

Over the past Ecuador cuentafew decades, the field of modern affective science has witnessed a remarkable growth in interdisciplinary research focused on language and emotion. Linguistic studies, for instance, reveal that almost every component of human language —such as semantics, grammar, discourse, and conversation— communicate emotion (Majid, 2012). It is precisely for this reason that this study will explore the emotional dynamics within the narrative-driven video game Life is Strange: True Colors (2021) through Sentiment Analysis and a corpus linguistics approach. The primary research interest focuses on understanding how dialogues in the game evoke empathy and influence players' moral decision-making, how contextual subliminal messages reinforce the game's emotional tone, and identifying linguistic patterns that construct the narrative and reflect emotional dilemmas. Moreover, the study aims to investigate how linguistic choices in the game shape players' perceptions of characters and events.

Although this is a work in progress in its most early stage, the aim of this research is fourfold: (1) to analyse the emotional impact of the game's dialogue on the player, particularly in critical decision-making scenarios; (2) to examine how subliminal environmental cues (e.g., background text, character actions, and setting descriptions) align with or amplify the emotional tone conveyed in conversations; (3) to identify recurring linguistic patterns or strategies that articulate key emotional dilemmas and foster narrative coherence; and (4) to assess the role of linguistic features in influencing players' attachment to characters and immersion in the narrative. By addressing these aims, the study seeks to contribute to a deeper understanding of the interplay between language, emotion, and player experience in interactive storytelling.

The methodology applied to carry out the study will integrate a discourse analysis approach with tools from corpus linguistics and Natural Language Processing (NLP). First, the game's dialogues and environmental text will be transcribed to construct a comprehensive textual corpus. Using AntConc (Anthony, 2024) as the main corpus analysis software, the study will investigate frequency patterns, concordances, and collocations of emotionally charged words and phrases. Sentiment Analysis will be also applied to categorise and map emotions expressed in dialogues and narrative contexts.

Expected results include the identification of different linguistic patterns that contribute to the game's immersive emotional experience, such as the frequent use of empathy-triggering expressions, persuasive dialogue structures, and contextual alignment between dialogues and environmental cues.

In conclusion, this research underscores the importance of linguistic analysis in understanding narrative-driven video games, particularly those like Life is Strange: True Colors (2021), which prioritise emotional depth and moral complexity. By combining Sentiment Analysis, corpus linguistics, and narrative theory, the study aims to explore how in-game dialogues evoke empathy and shape players' moral choices, how contextual subliminal cues enhance the emotional tone, and how linguistic patterns structure the narrative while reflecting key emotional dilemmas.

References

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc
Deck Nine (2021). Life Is Strange: True Colors [Nintendo Switch]. Square Enix.
Majid, A. (2012). Current emotion research in the language sciences. *Emotion Review*, *4*(4), 432–443. https://doi.org/10.1177/ 175407391244582

A corpus-assisted analysis of 'home' multimodal patterns in migration-themed picture books

María Bîrlea

Universidad de Salamanca

The social attention that migration has received in the last decades is being mirrored in children's literature, with forced migration narratives recognised as an emergent genre (Hope, 2008). Narratives of migration in these books often introduce the topic of homeland, which helps readers develop a sense of empathy towards the migrant characters and their uncertain condition. The present study explores how migration and migrants are represented in a corpus of 60 children's picturebooks published in English. Using methods from Corpus-Assisted Discourse Studies (Gillings, Mautner & Baker, 2023), systemic-functional multimodal discourse analysis (Serafini, 2022), and *collustrations* (McGlashan, 2016) the author(s) analyse linguistic and visual patterns to explore how the ideas of "home" and "identity" are portrayed with regards to migration.

Existing literature define identities as neither fixed nor stable and consider a strong connection between mobility, place and identity (Easthope, 2009). As people acquire new languages and identities throughout their lives, it is not uncommon to find cases of multiple identities. Protagonists in these picturebooks "must learn to inhabit at least two identities, to speak two cultural languages, to translate and negotiate between them" (Hall, 1996:629). For this reason, the degree of integration of the migrant into the host country is also considered in the present study.

We have followed a model of degrees of 'entry' (Budyta-Budzyńska, 2011) that makes a distinction between assimilation, integration, adaptation and separation. Given that it differs depending on how assimilated migrant characters are or wish to be, this has shown to be crucial for analysing migrants' identities and their perceptions of "home". The study goes on to examine how identity is reflected in names and naming, pointing out that some characters consider whether changing their names may boost their assimilation into the host society.

Results show that 40,5% of the corpus describe the adaptation process, 29,7% the separation process, 16,2% the assimilation process and, lastly, 13,5% the integration one. Migrants will either call their homeland their country of origin, their host country or both, depending on the level of their 'entry'. The analysis also reveals that <home> is a place in which the migrant feels protected: they seek a home, refuge. This lemma usually refers to both the <old> home they left and the <new> one they want to settle in. Occasionally, this home is America, which is conceived either as a land full of opportunities in which the characters assimilate very easily or a mundane place which they do not like at the beginning but start adapting to it.

The results obtained through the analysis of frequent linguistic patterns and trends within and across the different subcorpora that make up the corpus thus reveal that identities are represented through names and naming and collustrations of "home". Characters make decisions ranging from rejecting their home countries to accepting the new one completely, and endure an ongoing identity negotiation as they wonder where they fit in a foreign nation.

References

Budyta-Budzyńska, M. (2011). Adaptation, integration, assimilation – an attempt at a theoretical approach. In M. Budyta-Budzyńska (Ed.), *Integration or assimilation: Poles in Iceland*, (pp. 43-64). Warsaw: Wydawnictwo Naukowe Scholar.

Easthope, H. (2009). Fixed Identities in a Mobile World? The Relationship Between Mobility, Place, and Identity. *Identities 15*(1), 61-82.

Gillins, M., Maunter, G., and Baker, P. (2023). *Corpus-Assisted Discourse Studies*. Cambridge: Cambridge University Press.

Hall, S. (1996). Introduction: Who needs "identity"? In Hall, S. and P. du Gay (Eds.), *Questions of Cultural Identity* (pp. 1-17). Sage.

Hope, J. (2008). "One day we had to run": The development of the refugee identity in children's literature and its function in education. *Children's Literature in Education 39*, 295-304.

McGlashan, M. (2016). *The representation of same-sex parents in children's picturebooks: a corpus-assisted multimodal critical discourse analysis*. Doctoral dissertation.

Serafini, F. (2022). *Beyond the visual: An introduction to researching multimodal phenomena*. New York: Teachers College Press.

Applying principal component analysis to news headlines: Dimensions of sensational style in BBC, Guardian and Mail Online headlines

Ruth Breeze

Universidad de Navarra

Newspaper headlines have been discussed from a variety of perspectives, including cognitive psychology, relevance theory and CDA. Experts generally divide headlines in English into two categories, coinciding broadly with serious (broadsheet) and sensational (tabloid) journalism (Conboy 2005). Broadsheet headlines are supposedly sober, providing more information concerning the contents of the news article, while tabloid headlines rely on allusion, hyperbole and emotion to entice readers to read the text (Dor 2003). Since empirical research (Ecker et al. 2014) indicates that hyperbolic or emotional headlines influence reader perceptions and contribute to biased interpretations, it is important to gain a deeper understanding of this phenomenon. However, research on headlines so far has mainly relied on detailed analysis of small datasets, and no systematic corpus study exists.

The present study examines three large datasets of headlines from different media using semantic tagging and Principal Component Analysis. The research question is: how do these headline datasets different in their use of the categories "degree" and "emotion". All 2021 headlines were scraped from BBC news (17464 headlines), the Guardian (43664 headlines) and the Mail Online (46480 headlines). The datasets were cleaned and uploaded to Wmatrix5 for semantic analysis. Using Wmatrix5, we calculated the relative frequency of items tagged for degree (boosters, maximisers, approximators, diminishers, etc.) and emotion (sad, angry, happy, etc.) in each corpus. Principal Component Analysis was used to reduce the dimensionality of the data obtained. Major differences emerged concerning degree. The first Principal Component, which explained 40% of variance, appeared to measure the presence of degree in general (boosting, hedging, etc.) versus absence of degree. Mail Online was high on this component, while BBC was low and Guardian tended towards a middle position. The second Principal Component appeared to measure boosting and caution/hedging (positive) versus maximisers and exclusivisers (negative). Guardian headlines were positive on this Principal Component, while Mail Online headlines were negative. Regarding emotions, the first Principal Component

(20% of variance) distinguished between intense negative emotions “fear” and “shock” (BBC), and the less extreme emotions “worry” and “dislike” (Guardian). The second Principal Component (18%) separated the emotions “happy”, “sad” and “courage”, associated with Mail Online, from the absence of these in the other two media.

These findings shed light on headline style in serious and sensational online newspapers and have relevance for discourse analysis and media literacy. We suggest that PCA might be applied to discursive phenomena in large datasets, enabling researchers to detect patterns that are not immediately obvious, and allowing them to estimate the extent to which these appear to be generalized.

References

Conboy, M. (2005). *Tabloid Britain. Constructing a community through language*. Routledge.

Dor, D. (2003). On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, *35*(5), 695-721.

Ecker, U., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied*, *20*(4), 323-335.

Análisis de la comunicación de y con los influyentes españoles e ingleses en X e Instagram

María Luisa Carrió Pastor

Universitat Politècnica de València

Esta propuesta se centra en el estudio de la comunicación en redes sociales desde un punto de vista pragmático. El estudio consiste en el análisis del discurso digital que realizan y los comentarios que reciben los/las influyentes españoles/as e ingleses/as en X e Instagram. Para ello, se ha tenido en cuenta las características de la comunicación digital, el lenguaje evaluativo, la Teoría de la valoración (Martin y White, 2005; Cavasso y Taboada, 2021) y los estudios sobre el género en el discurso digital (Caldeira et al., 2018; Esposito y Breeze, 2022). Los objetivos de esta investigación son, por un lado, identificar las diferencias entre los posts de los influyentes masculinos y femeninos en X e Instagram, por otro, clasificar y comparar la función de los comentarios que reciben los/as influyentes en las dos redes sociales y, por último, mostrar las diferencias entre las dos lenguas y culturas en los posts y los comentarios. El material para este estudio se ha recopilado de las cuentas de los influyentes en X e Instagram mediante la herramienta ExportComments. Con el fin de conseguir que el corpus sea representativo se eligieron los influyentes con más seguidores en las dos redes sociales, seleccionando los posts que generaron más respuestas. El corpus compuesto de los posts de los/as influyentes y los comentarios que reciben fue seleccionado durante 2023 y 2024 para poder tener una muestra representativa. Se utilizaron como base las diferentes categorías de la Teoría de la valoración para identificar y clasificar el lenguaje evaluativo utilizado por los/as influyentes y sus seguidores/as, proponiendo una clasificación cuyo foco es la función comunicativa que se pretende con los posts y los comentarios en las dos redes sociales. Así mismo, se incluyeron los emoticonos y las imágenes que aparecen tanto en los posts como en los comentarios. Todo ello se analizó de forma semiautomática, primero, buscando las palabras clave con Sketch Engine y, a continuación, se identificaron los comentarios en su contexto, por lo cual, el corpus analizado no pudo ser tan extenso. En los resultados, se identificaron, clasificaron y describieron las diferencias entre hombres y mujeres, entre las dos redes sociales y entre las dos culturas, la española y la británica y cómo se reflejaba en las redes sociales, enfatizando las diferentes funciones comunicativas de las dos lenguas y el interés que persiguen los influyentes.

Finalmente, en las conclusiones se resumieron los hallazgos proponiendo una taxonomía que nos ayude a entender la comunicación digital de y con este tipo de usuarios de las redes sociales.

Referencias

Caldeira, S., De Ridder, S. & Van Bauwel, S. (2018). Exploring the Politics of Gender Representation on Instagram: Self-Representations of Femininity. *Journal of Diversity and Gender Studies*, *5*(1), 23–42.

Cavasso, L. & Taboada, M. (2021). A corpus analysis of online news comments using the Appraisal framework. *Journal of Corpora and Discourse Studies*, *4*(1), 38-51.

Esposito, E. and Breeze, R. (2022). Gender and politics in a digitalised world: Investigating online hostility against UK female MPs. *Discourse & Society*, *33*(3), 303-323

Martin, J. R. & White, P. R. (2005). *The Language of Evaluation: Appraisal in English*. London: Continuum.

Los vídeos testimoniales como género emergente en la comunicación biomédica

Luisa Chierichetti

Università degli Studi di Bergamo

Los videos testimoniales de cirugía estética se estructuran como relatos breves de pacientes que narran sus experiencias después de someterse a procedimientos médicos de este tipo. Su difusión a través de plataformas como YouTube, Instagram y TikTok, donde pueden llegar a audiencias amplias y segmentadas, los destaca como un género emergente dentro de la comunicación biomédica, en el que se combinan elementos de marketing, educación y humanización de la medicina estética. Por un lado, su dimensión persuasiva es innegable y evidente; entre los argumentos destacados se encuentran la posibilidad de acceder a procedimientos electivos no disponibles en el sistema público de salud, el uso de tecnologías innovadoras, la experiencia de un equipo médico altamente especializado y un enfoque en la atención personalizada. Por otro lado, los videos testimoniales transmiten información sanitaria de forma más comprensible y emocionalmente impactante, favoreciendo la confianza y el interés en los procedimientos de cirugía estética. El contexto biosanitario, combinado con el propósito persuasivo, imprime a estos materiales un discurso esencialmente empático y no confrontativo, diseñado para conectar emocionalmente con la audiencia.

La presente propuesta se basa en el subcorpus "Testimonios" del corpus de documentos institucionales (de más de dos millones y medio de palabras) creado en el marco del proyecto DISBIOCOM – Biomedical discourse and communication in multicultural societies, financiado por el Ministerio italiano de Universidades e Investigación. El subcorpus, que gestionamos con SketchEngine, suma 24.578 tokens y se compone de las transcripciones de 45 videos de distintas duraciones, producidos y promovidos por clínicas privadas españolas, las cuales, al ofrecer tratamientos no disponibles en el sistema público, legitiman y fomentan las intervenciones quirúrgicas como un medio para mejorar la calidad de vida y alcanzar objetivos de salud, bienestar y mejora personal.

Las preguntas principales de la investigación se centran en identificar las características de los videos testimoniales como género comunicativo en el ámbito institucional, analizar cómo se expresan y configuran las emociones en estas narrativas, explorar las diferencias discursivas según el género de los pacientes o en casos de cambios de sexo, y examinar las estrategias utilizadas para conectar emocionalmente con el público y promover a las clínicas como actores clave.

La metodología se centra en analizar los videos testimoniales como género comunicativo, identificando sus convenciones discursivas. Se estudian las emociones mediante un análisis léxico-semántico basado en herramientas de corpus y modelos teóricos para clasificar sentimientos positivos y negativos. Además, se examinan las diferencias discursivas según el género de los pacientes, incluyendo las narrativas asociadas con el cambio de sexo. Finalmente, se evalúan los recursos retóricos, visuales y audiovisuales utilizados para generar confianza y conexión emocional.

Los resultados previstos sugieren que los videos testimoniales se configuran como un género híbrido que fusiona relatos autobiográficos con estrategias promocionales, destacando emociones clave que pueden variar según el género y el tipo de paciente. Se plantea además la hipótesis de que la eventual interacción con el médico o la médica, así como el uso de determinados recursos audiovisuales, contribuye significativamente a fortalecer la credibilidad y la conexión emocional con la audiencia.

Referencias

Alba-Juez, L. / Thompson, G. (eds.) (2014). *Evaluation in Context*, Amsterdam/Philadelphia: John Benjamins.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Bamberg, M. (2009). Identity and narration. In Hühn, P., Pier, J., Schmid, W., & Schönert, J. (Eds.), *Handbook of narratology* (pp. 132-143). Berlin/New York: De Gruyter.

Bañón Hernández, A.M. (2018). *Discurso y salud. Análisis de un debate social*. Pamplona: EUNSA.

Bhatia, V. K. (2002). Applied genre analysis: A multi-perspective model. *Ibérica*, *4*, 3-19.

Gillings, M., Mautner, G., & Baker, P. (2023). *Corpus-assisted discourse studies*. Cambridge: Cambridge University Press.

Mackenzie, J.L. / Alba-Juez, L. (Eds.) (2019). *Emotion in Discourse*, Amsterdam/Philadelphia: John Benjamins.

Martin, J.R. / White, P.R.R. (2005). *The Language of Evaluation*, Basingstoke. Hampshire/New York: Palgrave/Macmillan.

Delineating the discursive (de)legitimation strategies outlined by Spanish politicians in their no-confidence motion speeches

Milagros del Saz

Universitat Politècnica de València

The present study explores the discursive (de)legitimation strategies (cf. Reyes, 2011; Rubio-Carbonero & Franco-Guillén, 2022; Van Leeuwen, 2007; 2008; Van Leeuwen & Wodak, 1999) enacted by three Spanish politicians, viz., Pablo Iglesias, Pedro Sánchez, and Santiago Abascal, within the context of a no-confidence motion (NcM) speech against the governing parties in the Spanish Congress in 2017 (Popular Party), 2018 (Popular Party), and 2020 (Partido Socialista Obrero Español). To facilitate the analysis, a corpus analytical approach using comparative keyword analysis was employed to analyze the NcM speeches of the three Spanish candidates with SketchEngine, which constituted a specialized corpus that worked as the 'research corpus.' Then, a qualitative analysis of the concordances where these keywords were used was conducted to access the surrounding text of each word critical to our research in line with previous research carried out by Rivers and Ross (2020) and to unveil the appeals most frequently employed to justify the need to file the motion and provide reasons to evict the incumbent party. The qualitative assessment of the concordances was carried out with the following questions in mind: (a) is the keyword employed for the justification or discrediting of

an action?; and (b) what is the type of (de)legitimation appeal(s) most clearly invoked? In sum, the aim of this piece of research was not to quantify the frequency of occurrence of the (de)legitimation strategies but rather to describe those that were revealing of the style of the politician under examination as resulting from the corpus-assisted comparative keyword search. Findings point to interindividual differences regarding the appeals used. Iglesias heavily relies on altruism to present his group's project as an alternative and on implicit authorization via referencing sources that support his claims to gain the audience's credibility. Sánchez legitimizes his actions by rationalizing his reasons for filing the motion and conveying – via implicit authorization – that the motion is triggered by the need to uphold constitutional principles. Abascal, on his part, relies on the negative association of the out-group (cf. Chilton, 2004) with lexis of a moralizing nature that challenges their credibility and reputation via direct appeals to Sánchez and Iglesias while appealing to emotions and the rationalization of the motion in terms of freedom.

References

Chilton, P. (2004) *Analysing Political Discourse: Theory and Practice*. London: Routledge.
Reyes, A. (2011) Strategies of legitimization in political discourse: From words to actions. *Discourse & Society*, *22*(6), 781–807.
Rivers, D., and Ross, A. (2020) Authority (de)legitimation in the border wall Twitter discourse of President Trump. *Journal of Language and Politics*, *19*(5), 831–856.
Rubio-Carbonero, G., and Franco-Guillén, N. (2022) When in parliamentary debate there is no debate. *Journal of Language and Politics*, *21*(4), 544–566.
Van Leeuwen, T. (2008) *Discourse and Practice: New Tools for Critical Discourse Analysis*. Oxford University Press.
Van Leeuwen, T., and Wodak, R. (1999) Legitimizing immigration control: A discourse-historical analysis. *Discourse Studies*, *10*(1), 83–118.

Paving the road towards eco-tourism in the Po Delta through keyness. A corpus-assisted analysis of English tourist materials from 1960 to 2000

Eleonora Federici, Ilaria Iori

Università degli Studi di Ferrara

Tourism in Po-Delta is one of the main economic drivers to foster sustainable economic growth. The Po Delta is located in Northern Italy, primarily within the regions of Emilia-Romagna and Veneto. It is a wetland area formed by the Po River which covers marshes, and lagoons, a protected area rich in biodiversity, enchanting landscapes, and history. Po Delta Park represents a good example of recent efforts to enhance eco-tourism and preserve the natural area and its biodiversity (Gaglio et al., 2023). Over the last decade, the Po Delta has shifted from an agricultural region to a tourism hotspot, emphasizing changes in marketing approaches, which makes the study of its evolving tourist narratives crucial (Di Giulio et al., 2017). As a tourist destination, it offers a rich variety of landscapes, wildlife, and historical sites, making it an attractive destination for tourists. Nonetheless, the region has been the focus of various conservation efforts, balancing tourism with environmental sustainability. This analysis investigates the promotional strategies used in tourism discourse in the Po Delta from 1960 to 2000 and focuses on the following research questions: (1) What are the key semantic domains for each decade and how do they vary across time? (2) To what extent do the key semantic domains reflect the shift towards eco-tourism? (3) What insights does the analysis offer in terms of promotional strategies? In order to answer these questions, we compiled a corpus of English-

language brochures and magazines published by Italian institutions (the Italian National Tourist Board, the Ferrara province tourism sector, and the Ferrara and Comacchio city tourism offices) on the Po Delta from the 1960s to 2000s. From a methodological perspective, the study employs a corpus-assisted approach (Partington et al., 2013), starting from a key semantic domain analysis with WMatrix (Rayson, 2008), aimed at examining the diachronic variation of narratives in a corpus-driven approach. Key semantic domains can be considered markers of the semantic relations of the "aboutness" of a text (see Bondi, 2010). After having identified the key semantic domains for each decade, the study qualitatively analyses their concordance lines to find phraseological patterns (Sinclair, 2004). Preliminary results show that the key semantic domains reflect a shift towards sustainable tourism. In the 1960s, key semantic domains focused on evaluative language and little attention was paid to semantic domains related to eco-tourism, while in the 1970, the key semantic domains showed a slight emerge of slow experiences promotion. From 1980 until 2000 the key semantic domains seem to suggest increasing awareness to green issues and recreational activities that tourists can do in the Po Delta (e.g., *sailing*, etc.), with more attention being paid to the calm and restoring experience being promoted in the area.

References

Bondi, M. (2010). Perspectives on keywords and keyness: An introduction. In M. Bondi & M. Scott (Eds.), *Studies in Corpus Linguistics* (Vol. 41, pp. 1–18). John Benjamins Publishing Company. https://doi.org/10.1075/scl.41.01bon

Di Giulio, R., Emanueli, L., Lobosco, G., Piaia, E., & Stefani, M. (2017). Selective Retreat Scenarios for the Po River Delta. *The Plan Journal*, *2*(2). https://doi.org/10.15274/tpj.2017.02.02.03

Gaglio, M., Lanzoni, M., Goggi, F., Fano, E. A., & Castaldelli, G. (2023). Integrating payment for ecosystem services in protected areas governance: The case of the Po Delta Park. *Ecosystem Services*, *60*, 101516. https://doi.org/10.1016/j.ecoser.2023.101516

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)* (Vol. 55). John Benjamins Publishing Company. https://doi.org/10.1075/scl.55

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519–549. https://doi.org/10.1075/ijcl.13.4.06ray

Sinclair, J. M. (2004). *Trust the Text: Language, Corpus and Discourse* (R. Carter, Ed.). Routledge.

Interpreting views and reports on violence against women in the Spanish subcorpus of NEWSGEN-VAW: Presence, absence, and contextualization in journalistic corpora

Miguel Fuster-Márquez, José Santaemilia

Universitat de València

The interpretation of textual presences and absences in Corpus Linguistics (CL) and (Critical) Discourse Studies (CDS) has been the subject of extensive debate. Stefanowitsch (2020) posits that in CL, absences can be as informative as presences, whether these absences pertain to linguistic elements, discourse positions, or external factors. Although it may seem intuitive to dismiss what is not found in a corpus as irrelevant, this view is inaccurate. The issue of absences in texts has also been a central focus in Corpus-assisted Discourse Studies (CaDS) (see Partington, 2014; Schröter & Taylor, 2018). O'Halloran (2005) highlights the importance of identifying omissions in texts, suggesting that a focus on absences can address criticisms that CDS often overlook important elements.

A strength of CaDS lies in its use of large, balanced datasets, allowing for reliable generalizations (Baker, 2015). These techniques mitigate cherry-picking concerns and reduce

bias, though not entirely. Data selection in CDS is inherently theory-laden, linking theories, concepts, and empirical indicators systematically (Gilbert, 2008, in Wodak & Meyer, 2016). Self-reflexivity and precise contextualization support robust CaDS. The ability to manage large datasets effectively ensures that the analysis is relevant, and representative of the discourse being scrutinised.

In our critical approach to press analysis, textual presence and absence may have ideological motivations. This study examines the Spanish subcorpus of NEWSGEN-VAW, compiled by researchers at the University of Valencia, which contains over 18 million words. We analyze the entire body of newspaper items (opinion pieces and news stories) produced by *El País*, *El Mundo*, and *ABC* over the last 20 years concerning Gender-Based Violence. Our focus is on the various terms journalists have used to address this issue, such as "violencia machista," "violencia de género," "feminicidio," "violencia doméstica," or "malos tratos," among others. We employ various corpus techniques to explore the ideological positions in these leading Spanish newspapers, facilitated by the annotation system in NEWSGEN-VAW, which allows for a more precise identification of the meaningful presence or absence of these terms in the particular social, legal, and political context of Spain over the last two decades.

Additionally, we consider external factors such as the influential position of the Real Academia Española, earlier work on the subject, and critical feminist approaches (Lazar, 2018; Maruenda-Bataller, 2021; Santaemilia, 2021; Santaemilia & Maruenda-Bataller, 2014; Zurbano-Berenguer, 2012; Fuster-Márquez, 2022; Fuster-Márquez, 2024), as well as recommended ethical approaches to naming practices by journalists (Eastal et al., 2022). These factors are crucial for understanding how terms related to Gender-Based Violence are employed, and the potential biases or ideological underpinnings in their usage. The influence of authoritative bodies like the *Real Academia Española* can shape language practices and affect public perception and policy.

We underscore that the validity of interpretation in CaDS is inherently linked to data quality and methodological coherence, from quantification to the subsequent qualitative approaches typical of Classical Discourse Analysis, where contextual factors in the case of Spanish newspaper articles about Gender-Based Violence play a significant role.

References

Baker, P. (2015). Introduction to Special Issue. *Discourse and Communication*, *9*(2), 143-147.

Easteal, P., Blatchford, A., Holland, K., & Sutherland, G. (2022). Teaching Journalists About Violence Against Women Best Reportage Practices: An Australian Case Study. *Journalism Practice*, *16*(10), 2185-2201.

Fuster-Márquez, M. (2022). Análisis contrastivo de la noticiabilidad en torno a la representación periodística de la violencia de género en la prensa española y estadounidense. En Garofalo, G. (Ed.), *Estudios de género asistidos por corpus: Enfoques multidisciplinarios* (pp. 67-98). Peter Lang.

Fuster-Márquez, M. (2024). *La prensa y la noticia: Un estudio crítico discursivo asistido por corpus*. Comares.

Fuster-Márquez, M. (2025, en prensa). The media presence of the Spanish far-right in speeches about gender-based violence: An approach from corpus linguistics. En Moreno-Serrano, L. M., & Maruenda-Bataller, S. (Eds.), *Discourse Approaches to Gender-Based Violence: Deconstructing Social Inequality Through Linguistic Inquiry*. Mouton de Gruyter.

Lazar, M. M. (2018). Feminist critical discourse analysis. En Flowerdew, J., & Richardson, E. (Eds.), *The Routledge Handbook of Critical Discourse Studies* (pp. 372-387). Routledge.

Maruenda-Bataller, S. (2021). The role of news values in the discursive construction of the female victim in media outlets: A comparative study. En Fuster-Márquez, M., Santaemilia, J., Gregori-Signes, C., & Rodríguez Abruñeiras, P. (Eds.), *Exploring discourse and ideology through corpora* (pp. 141-166). Peter Lang.

O'Halloran, K. (2005). Mystification and social agent absences: A critical discourse analysis using evolutionary psychology. *Journal of Pragmatics*, *37*, 1945-1964.

Partington, A. (2014). Mind the gaps: The role of corpus linguistics in researching absences. *International Journal of Corpus Linguistics*, *19*(1), 118-146.

Santaemilia, J., & Maruenda-Bataller, S. (2014). The linguistic representation of gender violence in (written) media discourse: The term 'woman' in Spanish contemporary newspapers. *Journal of Language Aggression and Conflict*, *2*(2), 249-273.

Santaemilia, J. (2021). News values as evaluation. Main naming practices in Violence Against Women news stories in contemporary Spanish newspapers: El País vs. El Mundo (2005-2010). *Research in Corpus Linguistics*, *9*(2), 90-113.

Schröter, M., & Taylor, C. (Eds.). (2018). *Exploring Silence and Absence in Discourse: Empirical Approaches*. Palgrave Macmillan.

Stefanowitsch, A. (2020). *Corpus Linguistics: A guide to the methodology*. Language Science Press.

Wodak, R., & Meyer, M. (2016). Critical discourse studies: History, agenda, theory and methodology. En Wodak, R., & Meyer, M. (Eds.), *Methods of Critical Discourse Studies* (pp. 1-22). Sage.

Zurbano-Berenguer, (2012). El concepto "violencia de género" en la prensa diaria nacional Española. *Medios de Comunicación, Publicidad y Género*, *7*, 25-44.

Chorriqueta, colchoneta: La construcción del *chemsex* en el discurso de las comunidades de debate social en España

Giovanni Garofalo, Carla Fernández Melendres

Università degli Studi di Bergamo, Universidad de Málaga

El término "chemsex", acrónimo de "chems" (drogas) y "sex" (sexo), hace referencia al policonsumo de sustancias psicoactivas (mefedrona, GHB, metanfetamina, ketamina, etc.) para intensificar la experiencia sexual y está vinculado principalmente con hombres gays, bisexuales y otros hombres que tienen sexo con hombres (GBHSH) (Ministerio de Sanidad, 2020; Salusso et al. 2020). En el seno de esta comunidad, ha ido desarrollándose una jerga al uso de los consumidores habituales de dichas sustancias, que fortalece los lazos de endogrupo (p.ej., el GHB o ácido gammahidroxibutírico, se conoce también como *éxtasis líquido*, *G*, *biberón o chorri*; la ketamina como *queta* y la metanfetamina como *meta* o *tina*). Aunque el uso recreativo y moderado de estas drogas no sea intrínsecamente problemático, representa una auténtica comorbilidad en individuos con VIH y suele conllevar una mayor probabilidad de contraer ITS (EACS 2023). Además, el chemsex puede relacionarse con problemas de salud mental, como la depresión y la ansiedad (Íncera-Fernández et al., 2021; Íncera et al., 2022).

El tratamiento mediático del chemsex tiende a centrarse en la transmisión del VIH, agresiones sexuales e incluso homicidios. Dicho enfoque sensacionalista contribuye a una doble estigmatización de los hombres gays, bisexuales, trans y queer (HGBTQ+), quienes ya son víctimas de prejuicios (Heritage & Baker, 2022) y que, además, se ven señalados como drogadictos.

Esta investigación pretende analizar la respuesta de algunas *comunidades de debate social* (Bañón & Asencio 2023) ante la propagación del chemsex. El concepto de comunidad de debate social se distingue del de 'comunidad de habla' o 'comunidad de práctica' y debe entenderse como "un conjunto de nodos interconectados y estructurados alrededor de una [problemática] o un interés común […] que genera un sentimiento de pertenencia a un mismo colectivo", estableciendo una dinámica reticular entre los miembros (Knoke et al. 2021, en Bañón & Asencio 2023: 171). Se trata, en general de *mesoactores* (Bañón Hernández 2018: 100-103), a saber, colectivos se desempeñan en el ámbito de lo social. Para analizar el discurso de dichos

mesoactores en torno a la adicción al chemsex, hemos compilado y anotado con etiquetas XML un corpus de 23 documentos que contiene alrededor de 50.000 palabras y que recoge las campañas de información de varias asociaciones españolas como RIS, STOP SIDA o CESIDA. Se trata de vídeos presentes en la web y dirigidos al colectivo HGBTQ+, en los que participan profesionales del sector biomédico y de la salud mental, además de figuras influyentes del ambiente LGTBIQ+ español, quienes procuran informar con una actitud empática y cómplice, rehuyendo el juicio moral y apelando a la solidaridad del endogrupo.

Metodológicamente, se indaga la construcción de la identidad de las personas que practican chemsex (Benwell & Stokoe, 2006) recurriendo a un análisis de *keywords* (Baker 2018). A tal efecto, se propone una comparación de las palabras claves extraídas por *Sketch Engine* (Kilgarriff et al., 2014), a través del corpus de referencia Spanish Web 2023, y mediante la aplicación *Corpus Sense* (Moreno-Ortiz, 2024ab), que prescinde del corpus de cotejo y aprovecha los Grandes Modelos de Lenguaje basados en Transformers, gracias a los cuales puede captar la semántica del corpus de estudio.

Los resultados preliminares indican la preeminencia de factores clave (homofobia interiorizada, falta de autoestima, soledad, uso compulsivo de aplicaciones de cita etc.) que configuran el chemsex como problema psicosocial que trasciende la mera dimensión biosanitaria.

Referencias

Bañón, A. & Asencio, A. (2023). Actores y comunidades de debate social. *Lengua y Sociedad. Revista de Lingüística Teórica y Aplicada*, *22*(1), 169-198.

Baker, P. (2018). Keywords. Signposts to objectivity? In Cermáková, A. (ed.) *The Corpus Linguistics Discourse: In Honour of Wolfgang Teubert* (pp. 77-94). Amsterdam: John Benjamins.

Benwell, B., & Stokoe, E. (2006). *Discourse and Identity*. Edinburgh University Press.

EACS (European AIDS Clinical Society) (2024). *Guidelines. Version 12.1. November 2024*. https://eacs.sanfordguide.com [22/12/24].

Heritage, F., & Baker, P. (2022). Crime or culture? Representations of chemsex in the British press and magazines aimed at GBTQ+ men. *Critical Discourse Studies*, *19*(4), 435–453. https://doi.org/10.1080/17405904.2021.1910052 [22/12/24].

Íncera, D., Gámez, M., Ibarguchi, L., García, A., Zaro, I., & Alonso, A. (2022). *Aproximación al Chemsex 2021: Encuesta sobre hábitos sexuales y consumo de drogas en España entre hombres GBHSH*. Apoyo Positivo e Imagina Más.

Íncera-Fernández, D.; Gámez-Guadix, M., & Moreno-Guillén, S. (2021). Mental Health Symptoms Associated with Sexualized Drug Use (Chemsex) among Men Who Have Sex with Men: A Systematic Review. *International Journal of Environmental Research and Public Health*, *18*(24), Article 24. https://doi.org/10.3390/ijerph182413299 [22/12/24].

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9 [22/12/24].

Knoke, D.; Diani, M.; Hollway, J. & Christopoulos, D. (2021). *Multimodal Political Networks*. Cambridge University Press.

Ministerio de Sanidad. (2020). *Abordaje del fenómeno del chemsex. Secretaría del Plan Nacional sobre el Sida*.

Moreno-Ortiz, A. (2024a). *Making Sense of Large Social Media Corpora: Keywords, topics, entities, and sentiment in the Coronavirus Twitter Corpus*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-52719-7 [22/12/24]

Moreno-Ortiz, A. (2024b). Corpus Sense (Version 0.9.7) [Python 3]. Universidad de Málaga. https://corpus-sense.app [22/12/24].

Salusso, D.; Núñez, S.; Cabrini, M.; Rolón, M.J., & Cahn, P. (2020). Chemsex y uso de sustancias durante las relaciones sexuales: resultados de una encuesta realizada en Argentina. *Actualizaciones en sida e infectología*, 28 (103), 40-50.

Lexico-grammatical structure of the STEAL event frame in English. Towards the non-discrete quantitative description of grammatical constructions

Dylan Glynn

Université Paris 8

This study takes the usage-based model (Hopper 1987, Langacker 1987, Schmid 2020) and asks how we can quantifiably account for language structure given the descriptive implications of this model. Following the arguments of Boas (2003) and Glynn (2004), the approach here adopts the premise that grammatical constructions should be identified from a semantic perspective rather than a formal one, first asking how a concept / function is expressed contrary to the more widely adopted approach of beginning with a form and looking for other semantically similar, yet distinct forms. Moreover, following Glynn (2015, 2022) and Dąbrowska (2017) instead of form-meaning pairs, constructions should be understood as combinatory clusters of formal and semantic characteristics of use. In order to test the feasibility of approaching grammatical constructions in these terms, the study examines a set of forms used to express stealing in contemporary English.

Since it is not possible to search corpora for concepts such as STEAL, the first step establishes a list of all the possible expressions used to express this concept. These "keywords" are in turn used to retrieve all STEAL occurrences from the LiveJournal Corpus of English (Speelman & Glynn 2005). The data are manually examined to check for false positives. Only occurrences where the *actus reus* is unquestionably 'taking' without consent and the *mens rea* is one of intention are retained. The relative frequency of each lexically derived STEAL expression is in turn used to calculate and extract proportionally representative sub-samples of each expression resulting in approximately 2000 occurrences. The lexically determined subsamples of instances of STEAL are then submitted to a behavioural profile analysis (Dirven *et al.* 1982, Geeraerts *et al*. 1994, Divjak & Gries 2006). Care is taken to annotate both formal and semantic characteristics independently. The annotation schema is derived from the attribute matrix of the FrameNet entry for STEAL and is supplemented with more fine-grained semantic variables such as valence (semantic prosody), arousal (degree of impact upon the injured party), operationalised with Likert scales and subjected to multiple coding.

This usage-based and lexical approach reveals several lexico-syntactic patterns, including the already established alternation between *rob-steal* described by Goldberg (1995). The ROB-STEAL constructions are instantiated by a wide range of predicates, but only a few are frequent (*cheat*, *nick*, *take*, *steal*, *rob*). The valence patterns reveal 3 lexico-syntactic patterns, two of which further divide into two protentional sub-patterns: (1a) STEAL Goods Cx.; (1b) STEAL *off* Plaintiff Cx.;(2) ROB Plaintiff Cx.; (3a) GO *off with* Goods Cx.; (3b) GO *away with* Goods. A Hierarchical cluster analysis, based on the semantic behavioural features, confirms these patterns and the lexical options associated with each construction. A multiple correspondence analysis is then used to identify the semantic profile of each cluster indicative of its use and the characteristics responsible for the choice of one construction of another. Although a lack of data does not permit confirmatory analysis, the entirely bottom up and quantified approach successfully reveals three to five composite (non-discrete) clusters of form-meaning pairs that can be understood as grammatical constructions.

References

Boas, H. C. (2003). *A Constructional Approach to Resultatives*. CSLI.
Dąbrowska, E. (2017). *Ten Lectures on Grammar in the Mind*. Brill.
Dirven, R., Goossens, L., Putseys, Y. & Vorlat, E. (1982). *The Scene of Linguistic Action and its Perspectivization by* speak, *talk*, *say*, and *tell*. John Benjamins.
Divjak, D. & Gries, St. Th. (2006). Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, *2*(1), 23–60.
Glynn, D. (2004). Constructions at the crossroads: The place of construction grammar between field and frame. *Review of Cognitive Linguistics*, *2*, 197–233. https://doi.org/10.1075/arcl.2.07gly
Glynn, D. (2015). Semasiology and onomasiology: Empirical questions between meaning, naming and context. In J. Daems, E. Zenner, K. Heylen, D. Speelman & H. Cuyckens (Ed.), *Change of Paradigms – New Paradoxes: Recontextualizing Language and Linguistics* (pp. 47-80). De Gruyter Mouton. https://doi.org/10.1515/9783110435597-004
Glynn, D. (2022). Chapter 8. Emergent categories: Quantifying analogically derived similarity in usage. In K. Krawczak, B. Lewandowska-Tomaszczyk & M. Grygiel (Ed.), *Analogy and Contrast in Language: Perspectives from Cognitive Linguistics* (pp. 245-282). John Benjamins. https://doi.org/10.1075/hcp.73.08gly
Geeraerts, D., Grondelaers, S. & Bakema, P. (1994). *The Structure of Lexical Variation: Meaning, Naming, and Context*. De Gruyter Mouton. https://doi.org/10.1515/9783110873061
Goldberg, A. (1995). *Constructions: A Construction Grammar approach to argument structure*. University of Chicago Press.
Hopper, P. (1987). Emergent grammar. *Berkeley Linguistics Society*, *13*, 139–157.
Langacker, R. (1987). *Foundations of Cognitive Grammar*. Vol. 1. *Theoretical prerequisites*. Stanford University Press.
Schmid, H.-J. (2020). *The Dynamics of the Linguistic System. Usage, conventionalization, and entrenchment*. Oxford University Press.
Speelman, D. & Glynn, D. (2005). LiveJournal Corpus of British and American English. University of Leuven.

Spanish resultatives: A collostructional study of Peninsular and American Spanish

Isabel Jiménez Sáez

Universidad de Córdoba, Université Paris 8

This study examines the use of light verbs (LV) in the "pseudo-copulative construction of change" in Spanish (e.g. *se puso nerviosa/o* "she/he became nervous"; *se quedó viuda/o* "she/he became a widower"). Such constructions have received little attention in Spanish since the research tradition does not consider them true resultatives (Demonte and Masullo, 1999; Mateu, 2012; Rodríguez Arrizabalaga, 1999, 2016, among others). Nevertheless, the pairing of LV and resultative phrase is a resource for expressing change of state in Spanish and arguably represents a grammatical construction. Assuming a Cognitive Linguistic model of grammar, LVs are understood as contributing systematically to the semantics of an utterance and are therefore integral to the construal of the event (Brugman, 2001). We focus on the alternation between five LVs *dejar*, *hacer(se)*, *poner(se)*, *quedar(se)*, *volver(se)* and their instantiations of the construction. These verbs all typically translate as 'become' in English and their semantic contribution to the alternation is not systematically transparent to native speakers. The aim, therefore, is to determine how these different light verbs contribute to the representation of the change of state event.

The data are extracted from Sketch Engine's enTenTen subcorpora of American and European Spanish web domains. Despite the use of complex regular expressions, it was not possible to automatically obtain clean concordances of LV + object + resultative phrase constructions. Due to the high number of returns, manual sorting of true positives was not feasible, so estimates of proportions were calculated based on subsamples. A representative sample size was determined via power analysis [p = 0.05, 95% CI] and manually sorted to estimate true positive proportions in the corpora. The estimated proportions of true positives across the different language variety corpora (Argentine, Chilean, Colombian, Mexican, Peninsular and Peruvian Spanish) were: *dejar* (14-25%), *hacer(se)* (32-42%), *poner(se)* (11-42%), *quedar(se)* (23-35%), and *volver(se)* (92-98%). With the possible exception of *poner(se)*, the lack of lexical variation between the regional variants is noteworthy.

To reveal the potential semantic contribution of the LV in the construction, we firstly apply collexeme analysis (Stefanowitsch and Gries, 2003; Gries, 2023). This produces an overall picture of the degree of association between each verb and the construction for each language variety. We do not expect significant variation between varieties at this point. Secondly, we apply covarying collexeme analysis (Gries and Stefanowitsch, 2004b; Stefanowitsch and Gries, 2005) in order to identify patterns of association between the verb and the resultative phrase. The lexical semantics found in the resultative phrase will reveal any systematicity that can be interpreted as an index of the differences in the event construal afforded by the LV. Until the results are entirely calculated, we will not know if there is significant variation between regional varieties. Our findings will contribute to previous research on Spanish LVs of becoming (Bybee and Eddington, 2006; Morimoto and Pavon, 2007; Conde Noguerol, 2013; Gorp, 2017; inter alia) as well as our understanding of the lexico-grammatical encoding of resultativity in Spanish.

References

Brugman, C. (2001). Light verbs and polysemy. *Language Sciences*, *23*, 551-578. https://doi.org/10.1016/S0388-0001(00)00036-X

Bybee, J., & Eddington, D. (2006). A Usage-Based Approach to Spanish Verbs of "Becoming." *Language*, *82*(2), 323–355. http://www.jstor.org/stable/4490159

Conde Noguerol, M. E. *Los verbos de cambio en español* [Doctoral dissertation, Universidade da Coruña]. http://hdl.handle.net/2183/10319

Gries, St. Th., & Stefanowitsch, A. (2004a). Covarying Collexemes in the Into- causative. In M. Achard & S. Kemmer (Eds.), *Language, Culture, and Mind* (pp. 225–236). University of Chicago Press.

Gries, St. Th., & Stefanowitsch, A. (2004b). Covarying Collexemes in the Into-causative. In M. Achard, & S. Kemmer (Eds.), *Language, Culture, and Mind* (pp. 225-236). CSLI.

Gries, St. Th. (2023). Overhauling Collostructional Analysis: Towards More Descriptive Simplicity and More Explanatory Adequacy. *Cognitive Semantics*, *9*(3), 351–386. https://doi.org/10.1163/23526416-bja10056

Gorp, L. V. (2017). *Los verbos pseudocopulativos de cambio en español: estudio semántico-conceptual de hacerse, volverse, ponerse, quedarse*. Iberoamericana/Vervuert.

Morimoto, Y. & Pavón Lucero, M. V. (2007). *Los verbos pseudo-copulativos del español*. Arco Libros.

Rodríguez Arrizabalaga, B. (1999). *La atribución en inglés y español contemporáneos. Contrastes en la expresión del cambio de estado*. Doctoral Dissertation, Universidad de Huelva. https://dialnet.unirioja.es/servlet/autor?codigo=169490

Rodríguez Arrizabalaga, B. (2016). Construcciones resultativas en español. Caracterización sintáctico-semántica. *Philologica Canariensia*, *22*, 55–87. https://doi.org/https://doi.org/10.20420/PhilCan.2016.103

Stefanowitsch, A., & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *2*(8), 209–243. https://doi.org/10.1075/ijcl.8.2.03ste

Stefanowitsch, A., & Gries, St. Th. (2005). Covarying collexemes. Corpus *Linguistics and Linguistic Theory*, *1*(1), 1–43. https://doi.org/10.1515/cllt.2005.1.1.1

Ideology through the lens of syntactic complexity: Czechoslovak and Czech presidential speeches

Miroslav Kubát, Xinying Chen, Michaela Nogolová

University of Ostrava

Analyses of political speeches have a long-standing tradition in stylometry, with a particular focus on presidential addresses. While studies of US presidential speeches dominate the field (e.g. Liu 2012; Lim 2004; Savoy 2010, 2016), similar research has emerged for other nations, e.g. Italy (Tuzzi et al. 2010) or Russia (Kuznetsova 2016). This study examines New Year's presidential speeches from Czechoslovak and Czech presidents. These speeches, delivered annually, provide a unique opportunity for heads of state to address the nation, reflect on the past year, and outline their vision for the future. As a continuous genre spanning nearly a century, these speeches form a valuable corpus for studying both language change and historical trends.

The Czechoslovak context introduces an additional ideological dimension to this corpus. Czechoslovakia was established as a democratic republic in 1918 following the collapse of the Austro-Hungarian monarchy after World War I. After experiencing German Nazi occupation and a brief postwar democratic period, Czechoslovakia transitioned into a communist regime as one of the Soviet satellite states. The Velvet Revolution of 1989 marked the return of democracy, and the subsequent peaceful split in 1993 led to the formation of two independent countries, Czechia and Slovakia. This historical trajectory allows us to analyze the speeches through the lens of two contrasting ideologies –communism and democracy – and observe their impact on the linguistic structures of presidential discourse.

While previous studies have primarily focused on the content, themes, and lexical stylometric features of these speeches such as lexical diversity, average word length, and text activity/descriptivity indices (Čech 2014; David et al. 2014; Jičinský & Marek 2017; Kubát et al. 2021), this study shifts the focus to the syntactic level. By applying methods for analyzing syntactic complexity, we offer a fresh perspective on the corpus. Specifically, we measure four indicators: average sentence length (in words and clauses), average clause length (in words), Mean Dependency Distance (MDD), and Mean Hierarchical Distance (MHD). These measures provide insights into the complexity and structural organization of presidential speeches across different historical and ideological contexts.

The results reveal notable differences in syntactic complexity across the ideological contexts. Democratic speeches tend to have higher sentence length, as observed in the average sentence length in words and clauses. In contrast, communist-era speeches exhibit shorter sentences suggesting a more constrained style. Additionally, longer clauses during the communist period compared to the democratic periods, indicating a shift in the internal organization of sentences. Mean Dependency Distance (MDD) and Mean Hierarchical Distance (MHD) reveal further structural differences: democratic speeches display greater syntactic depth and larger dependency distance, while communist speeches maintain a flatter structure and a smaller dependency distance. In general, democratic speeches demonstrate a higher syntactic complexity by comparison to communist speeches.

References

Čech, R. (2014). Language and ideology: quantitative thematic analysis of new year speeches given by Czechoslovak and Czech presidents (1949-2011). Quality & Quantity, *48*(2), 899–910.

David, J., Čech, R., Davidová Glogarová, J., Radková, L., & Šústková, H. (2013). Slovo a text v historickém kontextu– perspektivy historickosemantické analýzy jazyka [Word and Text in the Historical Context– Perspectives on the Historical-Semantic Analysis of Language]. Brno: Host.

Jičinský, M., & Marek, J. (2017). New year's day speeches of Czech presidents: phonetic analysis and text analysis. In Saeed, K., Homenda, W., & Chaki, R. (Eds.), *Computer Information Systems and Industrial Management* (pp. 110–121). Cham: Springer.

Kubát, M., Mačutek, J., Čech, R. (2021). Communists spoke differently: An analysis of Czechoslovak and Czech annual presidential speeches. *Digital Scholarship in the Humanities*, *36*(1), 138–152.

Kuznetsova, J. (2016). Modern Russian history through the New Year addresses. In Kübler, S., & Dickinson, M. (Eds.), *Proceedings of Computer Linguistics Fest 2016* (pp. 34–38). Bloomington, IN: Indiana University. Retrieved from http://ceur-ws.org/Vol-1607/kuznetsova.pdf (accessed 19 July 2019).

Lim, E. T. (2004). Five trends in presidential rhetoric: an analysis of rhetoric from George Washington to Bill Clinton. *Presidential Studies Quarterly*, *32*(2), 328–348.

Liu, F. (2012). Genre analysis of American presidential inaugural speech. *Theory and Practice in Language Studies*, *2*(11), 2407–2411.

Savoy, J. (2010). Lexical analysis of US political speeches. *Journal of Quantitative Linguistics*, *17*(2), 123–141.

Savoy, J. (2016). Text representation strategies: an example with the State of the Union addresses. *Journal of the Association for Information Science and Technology*, *67*(8), 1858–1870.

Tuzzi, A., Popescu, I.-I., & Altmann, G. (2010). *Quantitative Analysis of Italian Texts*. Lüdenscheid: RAM-Verlag.

Framing online grooming: Media discourses and their implications for public perception and policy in the UK and Spain

Sergio Maruenda-Bataller

IULMA/Universitat de València

The growing prevalence of online grooming (OG) raises urgent questions about how this crime is framed in the media and its subsequent influence on public perceptions and policy responses. Despite the critical role of the media in shaping societal issues, research on OG representations remain scarce, with most studies relying on quantitative content analysis that overlooks the linguistic and discursive dimensions of media portrayals. This study addresses this gap by posing the following research question: how do linguistic choices and narrative structures in the UK/SP press coverage of OG influence public understanding and policy-making regarding this form of gender-based violence (GBV)?

While a few studies (e.g., Cheit, 2003; Cheit et al., 2010) have explored linguistic aspects of OG reporting, there remains a pressing need for a systematic, interdisciplinary approach to examining this discourse. One promising yet underused theoretical framework is news values theory (Bednarek & Caple, 2017), which sheds light on the criteria that drive the selection and framing of news stories. By applying this framework, researchers can uncover the motivations

and biases inherent in OG coverage, providing a critical lens through which media portrayals can be evaluated.

Emerging evidence underscores the significant impact of media representations on public attitudes and policy development. For instance, studies by Kitzinger & Skidmore (1995) and Nair (2019) demonstrate how media narratives influence perceptions of child protection measures and inform legislative frameworks. However, concerns persist regarding the quality and accuracy of OG reporting. Research highlights issues such as child-blaming stereotypes (Saewyc et al., 2013), sensationalist "stranger danger" narratives (Cheit, 2003; Cheit et al., 2010), and the neglect of prevention strategies and victim support services (Döring & Walter, 2020). These representational flaws risk perpetuating harmful attitudes, obstructing the reporting of incidents, and fostering public fear while failing to address the complex realities of OG.

Building on our expertise in discourse analysis, this study applies the Discursive News Values Analysis framework to an ad-hoc corpus of UK/SP press articles (OG_NEWS) on OG. Our research critically examines linguistic choices, naming practices, and narrative structures to uncover patterns of bias, stereotyping, and omission. By framing OG as a form of gender-based violence (GBV), the study highlights how media representations may reinforce or challenge societal attitudes toward this crime.

Our findings reveal significant discursive patterns that shape public perceptions of OG, with implications for the development of more effective prevention and intervention strategies. By identifying specific areas for improvement in media reporting—such as promoting accurate, responsible narratives and incorporating prevention-focused content—this research offers valuable insights for policymakers, educators, and social advocates. Ultimately, this study contributes to the broader goal of fostering a more informed and empathetic societal response to online grooming.

References

Bednarek, M. & H. Caple (2017). *The Discourse of News Values: How News Organizations Create Newsworthiness*. Oxford: Oxford University Press.

Cheit, R. E. (2003). What hysteria? A systematic study of newspaper coverage of accused child molesters. *Child Abuse and Neglect*, *27*(6), 607–623. https://doi.org/10.1016/S0145-2134(03)00108-X.

Cheit, R. E., Shavit, Y., & Reiss-Davis, Z. (2010). Magazine coverage of child sexual abuse, 1992-2004. *Journal of Child Sexual Abuse*, *19*(1), 99–117. https://doi.org/10.1080/10538710903485575.

Döring, N. & R. Walter (2020). Media coverage of child sexual abuse: A framework of issue-specific quality criteria. *Journal of Child Sexual Abuse* 29(4): 393-412.

Kitzinger, J. & Skidmore, P. (1995). Playing safe: Media coverage of child sexual abuse prevention strategies. *Child Abuse Review*, *4*(1), 47–56. https://doi.org/10.1002/car.2380040108.

Nair, P. (2019). Child sexual abuse and media: coverage, representation and advocacy. *Institutionalised Children Explorations and Beyond 6*(1): 38-45. DOI: 10.5958/2349-3011.2019.00005.7.

Saewyc, E. M., Miller, B. B., Rivers, R., Matthews, J., Hilario, C., & Hirakata, P. (2013). Competing discourses about youth sexual exploitation in Canadian news media. *The Canadian Journal of Human Sexuality 22*(2): 95–105. https://doi.org/10.3138/cjhs.2013.2041.

Syntactic complexity across genres in Karel Čapek's writing

Michaela Nogolová, Xinying Chen, Miroslav Kubát

University of Ostrava

This study examines the syntactic complexity of texts spanning multiple genres by the renowned Czech author Karel Čapek. Čapek (1890–1938) was a Czech writer, playwright, and journalist, best known for his science fiction works. The analyzed corpus includes over 700 texts covering diverse genres (novels, short stories, newspaper articles, travel books, poems, scientific studies, personal correspondence, and children's literature) offering a unique opportunity to analyze genre variation within the works of a single author, thus eliminating authorship-related bias common in corpus-based genre studies.

The study focuses on the syntactic aspects of texts, which have traditionally been less common in stylometric analysis compared to lexical features. This disparity is largely due to the historical scarcity of syntactically annotated corpora (treebanks), which has limited researchers' ability to explore sentence structures within texts. However, the accuracy of syntactic annotation has significantly improved in recent years, thanks to advancements in natural language processing (NLP) and the development of robust tools and models. This study applies the Surface Syntactic Universal Dependencies (SUD) framework (Gerdes et al., 2018) to the collected texts and analyzes the syntactic complexity indices derived from the resulting treebanks.

Syntactic complexity in this study is measured using multiple indicators to capture both linear and hierarchical dimensions of sentence structure. These include average sentence length (measured in words and clauses), average clause length (in words), Mean Dependency Distance (MDD), and Mean Hierarchical Distance (MHD). Average sentence length and clause length reflect the level of syntactic elaboration, with longer sentences and clauses often reflecting greater complexity. MDD measures the linear distance between syntactically connected words, capturing how dispersed or tightly connected words are within a sentence, with larger MDD indicating greater complexity (cf. Liu 2008). In contrast, MHD assesses the hierarchical depth of a sentence by calculating the mean vertical distance between nodes in the dependency tree, with higher values indicating deeper layers of subordination and greater syntactic embedding (cf. Liu 2008). By combining these metrics, the study offers a comprehensive evaluation of syntactic complexity, accounting for both surface-level word relationships and the underlying hierarchical structures, enabling a nuanced analysis of genre variation within Čapek's writing.

The results reveal significant variation in syntactic complexity across genres in Karel Čapek's texts. Newspaper articles and travel books exhibit the highest complexity, characterized by longer sentences, multiple clauses, and greater hierarchical depth (MHD). Scientific texts display moderately long sentences with fewer clauses but higher hierarchical depth, reflecting their formal nature. Surprisingly, children's literature demonstrates notable syntactic richness, with sentence lengths and dependency distances (MDD) higher than expected, suggesting a more intricate style than typically seen in this genre. In contrast, poetry and novels feature simpler syntactic structures with shorter sentences and flatter hierarchies. Short stories and correspondence occupy a middle ground, balancing syntactic complexity with readability. These findings highlight Čapek's nuanced adaptation of syntactic strategies to align with the stylistic and communicative demands of each genre.

References

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), pp. 159-191.

Gerdes, K., Kahane, S., & Perrier, G. (2018). SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, 66–74. Association for Computational Linguistics.

La moción de censura en el Congreso de los Diputados de la España democrática: Evolución del léxico (1980-2023)

Andrés Ortega Garrido

Università degli Studi di Bergamo

La moción de censura se presenta como uno de los principales mecanismos de control de la democracia (Vírgala Foruria, 1988; Delgado Ramos, 2019), aun cuando su utilidad u oportunidad pueda ponerse en tela de juicio (Vírgala Foruria, 1987; Simón Yarza, 2015). Desde la perspectiva del análisis del discurso, es uno de los géneros textuales de la política que mejor refleja la contraposición de opiniones y la posible manipulación de la verdad (Rivas López, 2019; Andersen Bølset, 2023). Al igual que en otros discursos políticos, se recurre a falacias y sofismas (Sánchez García 2010) y su enunciación involucra una multiplicidad de destinatarios (Verón, 1987; García Negroni, 2016). La moción de censura ha sido analizada especialmente desde un punto de vista procedimental, dentro del ámbito de los estudios políticos (Sánchez de Dios, 1992; Duce Pérez-Blasco, 2018), pero ha recibido escasa atención desde el campo del análisis del discurso y de la lengua de la política, a pesar de la relevancia que tales alocuciones presentan como elemento de la estrategia y la comunicación políticas.

El mecanismo de la moción de censura, considerado un modo legítimo de desbancar a un gobierno, solamente ha sido puesto en marcha en seis ocasiones en los últimos cuarenta años de democracia española y únicamente tuvo éxito en una ocasión. Las dos primeras mociones de censura datan de 1980 y 1987; tras un lapso de treinta años se produce la siguiente, en 2017, seguida de otras tres en un espacio de tiempo relativamente corto (2018, 2020 y 2023).

En este trabajo, partiendo de una perspectiva lingüística, nos proponemos analizar un corpus formado por los discursos de los candidatos a la presidencia en las seis mociones de censura presentadas hasta la fecha, propuestas por el PSOE (1980), Alianza Popular (1987), Podemos (2017), PSOE (2018) y Vox (2020 y 2023). Teniendo en cuenta el intervalo temporal que separa a las dos primeras mociones de las cuatro últimas y habida cuenta de las transformaciones recientes en las formas de comunicación política (Ríspolo 2023), nos planteamos un análisis de la evolución de tales discursos a través, por un lado, de un estudio cuantitativo de palabras clave y de paquetes léxicos (Biber, Conrad, Cortés 2004; Biber 2005). Por otro lado, llevamos a cabo un análisis cualitativo, en parte sustentado por las concordancias del corpus, que nos ayudará a establecer no solo la evolución de las principales temáticas abordadas y la forma de presentarlas, sino también las modalidades lingüísticas que en cada momento han marcado los discursos. Así, veremos que en las primeras mociones son recurrentes términos más especializados, mientras que en las últimas domina sobre todo al léxico general; igualmente, destaca el papel que desempeñan los vocativos en todos los discursos y la tendencia al insulto en el discurso de 2020, así como el predominio de la cita y el argumento de autoridad en el último de los discursos, el único pronunciado por una persona alejada de la política activa (Ramón Tamames).

Referencias

Andersen Bølset, F. (2023). *Marcos cognitivos, lenguaje político y populismo. Un análisis de los discursos de Santiago Abascal y Pedro Sánchez durante la moción de censura de 2020*. Tesis doctoral. Norges Arktiske Universitet.

Biber, D. (2005). Paquetes léxicos en textos de estudio universitario: variación entre disciplinas académicas. *Revista Signos*, *38*(57), 19-29.

Biber, D., Conrad, S. & Cortés, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371-405.

Delgado Ramos D. (2019). Teoría y práctica de la moción de censura. Notas críticas a propósito de la experiencia reciente. *Revista general de derecho constitucional*, *29*.

Duce Pérez-Blasco, M. C. (2018). La moción de censura. *Corts. Anuario de derecho parlamentario*, *31*, 455-474.

García Negroni, M. M. (2016). Discurso político, contradestinación indirecta y puntos de vista evidenciales. La multidestinación en el discurso político revisitada. *Revista Latinoamericana de Estudios del Discurso*, *16*(1), 37-59.

Ríspolo, F. (2023). La historia, las ideas y los conceptos políticos. Una alternativa teórica para adentrarse en el lenguaje político. *Studia politicae*, *58*, 109-137.

Rivas López, C. (2019). *Análisis moral y político del debate de la moción de censura presentada por Pedro Sánchez*. Trabajo de Fin de Máster inédito. Universidad de Sevilla.

Sánchez de Dios, M. (1992). *La moción de censura (Un estudio comparado)*. Publicaciones del Congreso de los Diputados.

Sánchez García, F. J. (2010). Paralogismos y sofismas del discurso político español. La falacia política en un corpus de debates parlamentarios. *Anuario de Estudios Filológicos*, *33*, 271-290.

Simón Yarza, F. (2015). La moción de censura: ¿constructiva u "obstructiva"? *Revista española de derecho constitucional*, *103*, 87-109.

Verón, E. (1987). La palabra adversativa. Observaciones sobre la enunciación política. En E. Verón, L. Arfuch, M. M. Chirico, E. de Ipola, N. Goldman, M. I. González Bombal & O. Landi (Eds.), *El discurso político. Lenguajes y acontecimientos* (pp. 11-26). Hachette.

Vírgala Foruria, E. (1987). La responsabilidad política del gobierno en la RFA. La moción de censura constructiva y las mociones de reprobación. *Revista española de derecho constitucional*, *21*, 99-135.

Vírgala Foruria, E. (1988). *La moción de censura en la Constitución de 1978 (y en la historia del parlamentarismo español)*. Centro de Estudios Políticos y Constitucionales.

Panel 3

Corpus-based grammatical studies
Estudios gramaticales basados en corpus

La expresión de las relaciones causa-efecto: Estudio de un corpus financiero español

Blanca Carbajo Coronado, Antonio Moreno Sandoval

Universidad Autónoma de Madrid

Este estudio analiza un corpus financiero centrado en relaciones de causa-efecto, compuesto por tres partes: 1) 2.500 fragmentos (111.610 palabras) tomados de un corpus previo de informes financieros en español, seleccionados aleatoriamente por contener relaciones de causalidad (Gozalo, 2004); 2) preguntas asociadas a cada fragmento formuladas libremente por lingüistas para identificar causas o efectos, y 3) las respuestas correspondientes extraídas textualmente. Aunque fue diseñado inicialmente para entrenar modelos de procesamiento del lenguaje natural (PLN), este corpus es también una herramienta valiosa para estudiar la expresión de la causalidad.

El primer objetivo de este trabajo es analizar los marcadores discursivos y verbos causativos utilizados en textos financieros, siguiendo enfoques como el de Sun-Young (2022), y determinar si son específicos del dominio financiero. Asimismo, se busca examinar la subcategorización léxica de los verbos más frecuentes en preguntas (p.ej. *deber*, *permitir*, *provocar*) y respuestas (p.ej. *permitir*, *contar*). Otro objetivo clave es estudiar las estrategias para formular preguntas sobre causas o efectos, identificando estructuras gramaticales, tiempos verbales y colocaciones recurrentes. Para ello, se emplearon la herramienta de análisis de corpus Sketch Engine y spaCy.

Los resultados preliminares indican que el 57% de las preguntas se enfocan en la causa y el 42% en el efecto. Esta distribución podría explicarse por diversos factores: a) un interés de personas no expertas en finanzas por comprender las causas; b) la naturaleza de los textos financieros, donde los efectos suelen tener mayor relevancia que los hechos que los originan; o c) una tendencia natural de las personas a indagar más sobre las causas y no sobre los efectos.

En relación con las preguntas, se llevaron a cabo dos tipos de análisis. Por un lado, se normalizó la representación de las preguntas, formulándolas en presente, con sustantivos en singular y utilizando el pronombre *eso* para reemplazar información relevante (p.ej. *¿A qué se debe eso?*, *¿Qué implicación tiene eso?*). Esta normalización permitió identificar 159 formas diferentes de preguntar por la causa y 168 formas para el efecto, lo que evidencia mayor variedad léxica en las preguntas centradas en el efecto.

Por otro lado, se llevó a cabo un análisis sintáctico de dependencias con spaCy que identificó más de 130 estructuras gramaticales distintas en el total de preguntas. Específicamente, se utilizan 85 estructuras distintas en las preguntas que indagan por la causa y 66 en las que exploran el efecto, lo que demuestra que hay mayor diversidad gramatical en las preguntas sobre causas. Entre estas, la estructura más común es CASE + OBL + ROOT + NSUBJ (p.ej. *¿Por qué sucede eso?*), mientras que en las preguntas sobre el efecto predomina DET + OBJ + ROOT + OBJ (p.ej. *¿Qué consecuencia/efecto tiene eso?*).

En conjunto, se espera que este estudio contribuya al entendimiento de la expresión de la causalidad en el discurso financiero en español.

Referencias

Gozalo Gómez, P. (2004). *La expresión de la causa en castellano*. Madrid: UAM Ediciones.
Sun-Young, O. (2022). Causality in English Academic Writing: A Case of Research Articles in Applied Linguistics and Physical Chemistry. *Korean Journal of English Language and Linguistics*, 22, 40-54.

Using corpus data to test hypotheses about Spanish deictic adverbs: *aquí, acá, allí, ahí, allá*

David Ellingson Eddington

Brigham Young University

Spanish has two deictic adverbs expressing 'here' (*acá*, *aquí*) and three indicating 'there' (*ahí*, *allí*, *allá*). A number of factors have been associated with the use of particular 'there' of 'here' adverbs such as the degree to which the speaker is involved in the event (Maldonado 2013), the type of verb the adverb modifies (stative or non-stative; Di Tullio 2013, Hottenroth 1982, Sánchez Lancis 2001), as well as verbal tense (Real Academia Española 2022, Sedano 1999c, Tolosa 2008). Certain adverbs are also considered more likely candidates for modification by *tan*, *muy*, and *más* (Hottenroth 1982, Sacks 1974, and Hanssen 1913). In addition, some have suggested that particular adverbs indicate a higher degree of precision in describing the space or time of an event (Hottenroth 1982, Miyoshi 1996, Moliner 1997, Nilsson 1984, Ramsey 1956, Real Academia Española 2022, Sacks 1954, Sedano 2000, Tognola 2012, Tolosa 2008).

Most studies of the deictic adverbs are narrow in scope, often considering only one pair of adverbs, or their usage in one geographical area. Many are carried out using small data sets. In contrast, the present study is based on data taken from two large corpora: Corpus del español Web Dialects (Davies 2018b) and News on the web (Davies 2018c). They include data from 20 Spanish-speaking countries. A number of conclusions may be drawn from the corpus data:

1) The idea that *allá* is preferred over *ahí* and *allí* with non-stative verbs of action or movement is not supported.
2) The notion that the allegedly precise adverbs align with precise tenses, and imprecise adverbs with imprecise tenses receives little support.
3) In general, *acá* and *allá* are more often modified with *más*, *muy* and *tan* than their respective counterparts.
4) Very little support was found for the hypothesis that the degree of speaker involvement is related to specific deictic adverbs.
5) In contrast with the other adverbs, *acá* and *allá* do not consistently pattern with vague expressions of time or space.

A corpus analysis of mass shooter manifestos

Jenna Rose Elliot

Aston University

Mass shootings are currently one of the deadliest issues in the United States. Several notorious mass shooters have published their ideologies, motivations, and/or plans in

texts called 'manifestos' prior to their attack, but there is limited research available analyzing these texts. Due to the rise in frequency and severity of mass shootings (*Key Findings…*, n.d.), there is increasing urgency to improve our understanding of these attacks and individuals who carry them out. Previous research indicates mass shooter manifestos could be considered part of an illicit subgenre known as 'targeted violence manifestos' (Kupper et al., 2022; Kupper & Meloy, 2021). Texts of a shared genre will typically contain similar content and grammatical features (Chandler, 1997), but, currently, research on mass shooter manifestos has largely focused on identifying similarities in content and themes (Myketiak, 2016; Pfaffendorf & Davis, 2021; Kupper et al., 2022), and there is very little research on linguistic features of manifestos as a genre. Using corpus analysis, I aim to fill this research gap by answering the following research questions: what are the linguistic features used in mass shooter manifestos, and what similarities are there when comparing them? I will discuss keywords and key multi-word terms, frequently used lemmas, collocates, and pronoun usage and examine instances of shared linguistic features across ten different manifestos written by mass shooters between 2007 and 2022. I will describe similarities in frequently used lemmas and collocates across the manifestos that indicate the presence of shared linguistic features across these texts and, additionally, address differences found when comparing the manifestos. Analyzing the language of these manifestos to identify possible similarities will address the research gap examining linguistic features and, ultimately, indicate if there is evidence that these manifestos are part of the same genre.

References

Chandler, D. (1997). *An introduction to genre theory*.

Key Findings—Comprehensive Mass Shooter Data. (n.d.). The Violence Project. https://www.theviolenceproject.org/key-findings/

Kupper, J., Christensen, T. K., Wing, D., Hurt, M., Schumacher, M., & Meloy, R. (2022). The contagion and copycat effect in transnational far-right terrorism: An analysis of language evidence. *Perspectives on Terrorism*, 16(4), 4–26.

Kupper, J., & Meloy, J. R. (2021). TRAP-18 indicators validated through the forensic linguistic analysis of targeted violence manifestos. *Journal of Threat Assessment and Management*, *8*(4), 174–199. https://doi.org/10.1037/tam0000165

Myketiak, C. (2016). Fragile masculinity: Social inequalities in the narrative frame and discursive construction of a mass shooter's autobiography/manifesto. *Contemporary Social Science*, 11(4), 289–303.

Pfaffendorf, J., & Davis, A. (2021). Masculinity, ritual, and racialized status threat: Examining mass shooter manifestos using structural topic models. *Sociological Inquiry*, 91(2), 287–312.

EMOTIME: Metodología de corpus para el análisis de metáforas espaciotemporales EGO vs. TIME-MOVING

Águeda Salmerón Hortelano, Rosa Illán Castillo

Universidad de Murcia

El proceso de conceptualización de dominios abstractos como el TIEMPO implica establecer conexiones con experiencias sensoriales concretas como el MOVIMIENTO (Lakoff & Johnson, 1980). Esta proyección tiende a estructurarse mediante representaciones metafóricas. Concretamente, se distinguen dos tipos de metáforas: las metáforas *Ego-moving* (EM) y *Time-moving* (TM) (Casasanto & Boroditsky, 2008; Gentner et al, 2002; Illán,

2024; Loerman et al, 2018; Valenzuela & Illán, 2022, etc.). En la metáfora EM, el individuo se concibe como un agente que se mueve a través de un espacio temporal estático (p. ej. *We are approaching winter*) (Boroditsky, 2000). En contraste, en la metáfora TM, el tiempo se percibe como un flujo dinámico de eventos que se desplazan hacia un agente inmóvil (p. ej. *Winter is coming*) (Clark, 1973; Evans, 2004).

El paradigma psicolingüístico nos ha proporcionado una perspectiva empírica y experimental con datos de gran relevancia como la asociación de las metáforas EM con emociones positivas (felicidad), así como de las metáforas TM con emociones negativas como la ansiedad o la depresión (Richmond et al, 2012). Asimismo, se han relacionado con el EM otros matices emocionales de valencia negativa como el enfado (Hauser et al, 2009) o el duelo (Ruscher, 2011). No obstante, no se ha explorado suficientemente su uso en el lenguaje natural y su relación con factores semánticos y emocionales desde una perspectiva lingüística, que otros autores han tratado de suplir con estudios de corpus a pequeña escala (p. ej., McGlone & Pfiester, 2009; Feist & Duffy, 2020; Soriano & Piata, 2022).

El objetivo de este estudio radica en presentar una metodología de extracción de metáforas que pretende huir de las complicaciones de búsqueda causadas por la diversidad léxica de las mismas. Esta búsqueda en el corpus enTenTen21 de Sketch Engine resulta en un corpus especializado, consistiendo en la introducción del evento temporal como sujeto (TM) u objeto (EM) del verbo de manera de movimiento en la herramienta *Concordance*. Los eventos temporales por introducir se determinan según su frecuencia colocacional en *Word Sketch* en base a cada verbo de manera de movimiento extraído de la clasificación propuesta por Illán (2024); y siendo estos eventos posteriormente clasificados semánticamente por su significado temporal mediante una librería de procesamiento de lenguaje natural (NLTK), ejecutada en un código informático generado por ChatGPT.

Este método permite presentar resultados provisionales tanto a nivel global como particular sobre la frecuencia y prevalencia de los verbos de movimiento y sus colocados en cada tipo de metáfora. A su vez, permite la clasificación semántica de los eventos temporales presentes en el corpus, así como la determinación de los sujetos que funcionan como agentes en las metáforas EM frente al evento temporal. Esto apunta a una mayor diversidad léxica en las construcciones de TM. Así, estos resultados facilitarán el establecimiento de tendencias lingüísticas que favorecerán la puesta a prueba de las hipótesis establecidas en futuros experimentos y encuestas.

References

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, *75*(1), 1–28.

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, *106*(2), 579–593.

Clark, H. H. (1973). Space, time, semantics, and the child. En T. E. Moore (Ed.), *Cognitive development and the acquisition of language*, New York: Academic Press.

Evans, V. (2004). The structure of time: Language, meaning and temporal cognition. John Benjamins Publishing.

Feist, M. I., & Duffy, S. E. (2020). On the path of time: Temporal motion in typological perspective. *Language and Cognition*, 12(3), 444-467.

Gentner, D., Imai, M., & Boroditsky, L. (2002). As time goes by: Evidence for two systems in processing space - time metaphors. *Language and Cognitive Processes*, 2002, *27*(5), 537-565.

Glicksohn, J., & Ron-Avni, R. (1997). The relationship between preference for temporal conception and time estimation. *European Journal of Cognitive Psychology*, *9*, 1–15.

Hauser, D. J., Carter, M. S., & Meier, B. P. (2009). Mellow Monday and furious Friday: The approach-related link between anger and time representation. *Cognition and Emotion*, *23*(6), 1166-1180.

Illán, R. (2024). *The semantics of motion verbs within temporal conceptualization in English and Spanish: Understanding spatial dynamic construals of time* [Tesis doctoral, Universidad de Murcia]. Digitum.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Loerman, A., C., & Milfont, T. L. (2018). Time after time: A short-term longitudinal examination of the ego- and time-moving representations. *Journal of Research in Personality 2018*, *74*, 1-5.

McGlone, M. S., & Pfiester, R. A. (2009). Does time fly when you're having fun, or do you? Affect, agency, and time perception. *Journal of Experimental Social Psychology*, *45*(1), 104–111.

Richmond, J., Wilson, J. C., & Zinken, J. (2012). A feeling for the future: How does agency in time metaphors relate to feelings? *European Journal of Social Psychology*, *42*(7), 813-823.

Ruscher, J. B. (2011). Moving forward: The effect of spatiotemporal metaphors on perceptions about grief. *Social Psychology*, *42*, 225–230.

Soriano, C. & Piata, A. (2022). The affect bias in the metaphorical representation of anticipated events: The case of approach. *Metaphor and the social world*, *12*(1), 115-137.

Valenzuela, J., & Illán Castillo, R. (2022). A corpus-based look at time metaphors. In A. Piata, A. Gordejuela & D. Alcaraz Carrión (Eds.), *Time Representations in the Perspective of Human Creativity*, 15-40. Amsterdam: John Benjamins.

Exploring register variability in relativization across Asian Englishes

Cristina Suárez-Gómez

Universitat de les Illes Balears

Recent research on relativization in English highlights ongoing shifts in relative clause structures. Studies of traditional L1 varieties, particularly British and American English, have revealed significant changes across spoken and written registers, as well as formal and informal contexts (Mair & Leech 2005; Leech 2012; Xu & Xiao 2015). Specifically, there is an increasing preference for the invariable relativizer *that* and the zero relativizer over *wh-* forms, especially *which*. Furthermore, constructions featuring preposition stranding are becoming more prevalent, while pied-piping structures are declining. These trends have been attributed to prescriptive norms, including style guide recommendations such as the anti-*which* rule and the "sacred *that*" rule (Leech 2012: 60). They are also linked to processes of colloquialization, whereby written language adopts features of spoken discourse (Leech 2012; Xu & Xiao 2015), and informalization, which refers to the tendency for formal language to incorporate features of informal language (Farrelly & Seoane 2012).

This study investigates whether these preferences are similarly observable in World Englishes, with a focus on Asian varieties. Using data from the *International Corpus of English* (ICE), the analysis examines three representative registers: spontaneous conversations, popular science writing, and academic writing. Preliminary findings suggest a partial alignment with the stylistic shifts observed in L1 varieties. In academic prose, *wh-* markers remain prevalent as indicators of formality, yet *that* and zero relativizers as objects are dominant across text types, reflecting the effect of colloquialization and informalization. Conversely, the preference for pied-piping persists in formal genres, while preposition stranding is more common in informal categories. Within written registers,

academic writing shows stylistic variation between hard and soft sciences, while popular writing exhibits a more uniform style across topics. These findings confirm specific patterns of relativization that differ by genre and register. Expanding this analysis to larger samples and additional varieties will clarify whether the observed shifts in relativization extend from L1 contexts to World Englishes, shedding light on the global evolution of English grammar.

References

Farrelly, M. and Seoane, E. (2012). Democratization. In T. Nevalainen and E. C. Traugott (Eds.), *The Oxford Handbook of the History of English*, (pp. 392–401). Oxford University Press.

ICE = *The International Corpus of English* (n.d.). https://www.ice-corpora.uzh.ch/en.html.

Leech, G. (2012). How grammar has been changing in recent English: Using comparable corpora to track linguistic change. *Foreign Language Teaching Theory and Practice 32*(4), 13–20.

Mair, C and Leech, G. (2006). Current changes in English syntax. In B. Aarts and A. McMahon (Eds), *The Handbook of English Linguistics*. Blackwell.

Xu, X. and Xiao, R. Z. (2015). Recent changes in relative clauses in spoken British English. *English Studies 96*(7), 818-38.

Panel 4

Corpus-based lexicology and lexicography
Lexicología y lexicografía basadas en corpus

The construction of constructions.
A cross-linguistic study on reduplicative genitive superlative constructions

Caterina Chinellato, Pedro Ivorra Ordines

Universidade de Santiago de Compostela / Centro Universitario de la Defensa de Zaragoza

Reduplicative genitive constructions are a kind of reduplicative constructions, namely discontinuous reduplicative constructions (Mattiola and Masini 2022). These constitute polyfunctional constructions of lexical doubling "where other morphological material may appear between the reduplicant and the base" (Velupillai 2012: 101). Traditional approaches considered them as non-canonical instances of reduplication (Stolz 2018), due to their syntactic structures and non-adjacent lexical reduplication. However, few recent studies adopting a constructionist perspective have acknowledged their grammatical value in a variety of syntactic constructions, helping to open up the traditional paradigm of reduplication beyond the word boundary (López Meirama and Mellado Blanco 2018; Masini and Di Donato 2023; Masini and Mattiola 2022; Sommerer 2022). In this same line, the cross-linguistic perspective on discontinuous reduplicative constructions seems to be still under investigated, despite that recent studies showed that it can provide important insights into the description of the linguistic behaviour of these patterns (cf. Dobrovolsk'ij and Mellado Blanco 2021; Schafroth 2024).

Against this background, reduplicative genitive superlative constructions are explored in Italian ([DET N_{1sing} di N_{1pl}]; example 1), English ([the N_{1sing} of N_{2pl}]; example 2), Spanish ([DET N_{1sing} de DET N1pl]; example 3) and French [DET N_{1sing} des N_{1pl}]; example 4) using Corpus Linguistics as a methodological tool (TenTen family from Sketch Engine; cf. Yoon and Gries 2016). Following the basic tenets of Goldbergian Construction Grammar (Goldberg 2019; Hoffmann 2022), this study has the following goals:

- Determine the productivity of these constructions, considering "how many different items occur in the various slots of a construction" (Boas 2013: 247). On top of their type frequencies, type-token ratio and potential productivity are also employed to assess lexical diversity and proneness of innovation, respectively (cf. Baayen 2009).
- Classify the slot fillers into semantic clusters, so as to assess whether there are "pockets of productivity" (Cappelle 2014; cf. Ivorra Ordines forth.). In accordance with Barðdal' (2008) proposal, assess productivity as the result of the inverse correlation between type frequency and semantic coherence.
- Analyse the syntagmatic profile of the constructions, such as context of use, register, macro-syntactic aspects that may be relevant in the holistic description of the constructions (Goldberg 2006: 22), not only at an intralingual level but also cross-linguistically.

1. It. Ma il professore Marino è un Siciliano... **Il paradosso dei paradossi** è che hanno inculcato al Popolo Siciliano il pregiudizio razziale su se stesso. (itTenTen20, 178515826) 'But the professor Marino is a Sicilian… the paradox of paradoxes is that they have instilled the racial prejudice into the Sicilians about themselves. '

2. En. Dominate the arena and prove your worth as a soldier on the battle field. And prepare to be **the champion of champions**... (enTenTen18, 45670443)
3. Fr. Le roi fainéant, le crack des cracks, le cheval du siècle... meilleur trotteur français de tous les temps, Ourasi est aussi le seul à avoir remporté quatre fois le Prix d'Amérique en 1986, 1987, 1988 et 1990. (frTenTen23, 30888707)
   'The do-nothing king, **the best of the bests**, the horse of the century... the greatest French trotter of all time, Ourasi is also the only one to have won the Prix d'Amérique four times in 1986, 1987, 1988, and 1990.'
4. Sp. Yo quería ir a ver a Madonna, la diva de las divas, y no voy a ir porque no tengo enchufe. (esTenTen18, 4620496)
   'I wanted to go see Madonna, **the diva of the divas**, but I'm not going because I don't have any connections.'

References

Baayen, R.H. (2009). "Corpus linguistics in morphology: Morphological productivity". Volume 2: An International Handbook, (ed.) Anke Lüdeling and Merja Kytö, Berlin, New York: De Gruyter Mouton,pp. 899-919.

Barðdal, J. (2008). *Productivity. Evidence from Case and Argument Structure in Icelandic*. Amsterdam: John Benjamins.

Boas, Hans C. (2013) "Cognitive Construction Grammar", in *The Oxford Handbook of Construction Grammar*, (eds.) Hoffmann T., Trausdale G., Oxford: Oxford University Press, pp. 233-254.

Cappelle, B. (2014). "Conventional combinations in pockets of productivity: English resultatives and Dutch ditransitives expressing excess", in *Extending the Scope of Construction Grammar*, (eds.) Boogaart R., Colleman T. and Rutten G., Berlin, Boston: De Gruyter Mouton, pp. 251-282.

Dobrovolsk'ij, D., Mellado Blanco, C. (2021). Von Jahr zu Jahr. Das Pattern [von Xsg zu Xsg] und seine Entsprechungen im Russischen und Spanischen: eine Korpusstudie, in *Aussiger Beiträger* 15, pp. 113-138.

Finkbeiner, R., Freywald, U. (2018). Exact Repetition or total reduplication? Exploring their boundaries in grammar and discourse, in *Exact Repetition in Grammar and Discourse*, (eds.) Finkbeiner R., Freywald U., Boston: de Gruyter (Trends in Linguistics – Studies and Monographs 323), pp. 3-28.

Goldberg, A. (2006) *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Goldberg, A. (2019). *Explain me this. Creativity, competition, and the partial productivity of constructions*. Princeton: Princeton University Press.

Hoffmann, T. (2022). *Construction Grammar: The structure of English*. Cambridge: Cambridge University Press.

Ivorra Ordines, P. (Forth.), *Comparative constructional idioms of ugliness in Spanish, English and French: A contrastive usage-based approach*. Berlin: De Gruyter.

Masini, F., Mattiola, S. (2022). Syntactic discontinuous reduplication with antonymic pairs: a case study from Italian, in *Linguistics*, vol. 60 (1) 2022, pp. 315-345.

Masini, F., Di Donato J. (2023). Non-prototypicality by (discontinuous) reduplication: The N-non-N construction in Italian, in *Zeitschrift für Wortbildung* 7(1), pp. 130-155.

Mattiola, S., Masini, F. (2022). Discontinuous reduplication: a typological sketch, in *STUF - Language Typology and Universals*, vol. 75 (2), pp. 271-316.

López Meirama, B., Mellado Blanco, C. (2018). Las construcciones [*de X a Y*] y [de X a X]: realizaciones idiomáticas y no tan idiomáticas, in *Actas do XIII Congreso Internacional de Lingüística Xeral*, Vigo, pp. 576-583.

Schafroth, E. (2024). *Layer upon layer, mistake after mistake* - a case for learner's dictionaries?, in *Patterns of meaning in lexicography and lexicology*, (eds.) Giacomini, L., Piunno, V., Berlin, Boston: De Gruyter, pp. 159-180.

Sommerer, L. (2022). *Day to day and night after night*: Temporal NPN constructions in Present Day English, in *English Noun Phrases from a Functional Cognitive Perspective: Current Issues*, (eds.) Keizer, E., Sommerer, L., John Benjamins, pp. 363-394.

Stolz, T. (2018). (Non-)Canonical reduplication. in Non-Prototypical Reduplication, (eds.) Urdze A., Berlin, Boston: De Gruyter Mouton, 2018, pp. 201-278.

Yoon, J., Gries, S. (2016). (eds.), Corpus-based Approaches to Construction Grammar. Amsterdam: John Benjamins.

De lo más interesante: Intensificación y creatividad a través de los corpus

Maricel Esteban-Fonollosa

Universitat de València

La comunicación plantea el análisis de la construcción fraseológica [*de lo más* X]. Se trata de una construcción intensificadora propia del registro coloquial. García-Paje (1997) recoge la secuencia *de lo más* entre las estructuras de superlación lograda por medios sintagmáticos. En contexto, la integra como una de las "formas coloquiales de énfasis" en los que la entonación juega un papel importante. Nuestra aproximación la realizamos desde el marco de la Gramática de Construcciones y adoptamos la definición de construcción fraseológica de Mellado Blanco (2020). Según esta autora se trata de construcciones semiesquemáticas formadas por elementos saturados léxicamente y por otros que presentan casillas vacías, las cuales se actualizan en el discurso (Mellado Blanco 2020; Goldberg 2006).

En una primera exploración observamos que la construcción es productiva y el *slot* es prototípicamente ocupado por un adjetivo. En este sentido, encontramos casos en los que, sin embargo, este espacio es tomado por un sustantivo, prototípicamente nombre propio, cuyos atributos principales derivados de él constituyen el valor de la casilla vacía. Los *types* más frecuentes de la construcción son *de lo más interesante*, *de lo más original*, *de lo más importante*, todos ellos dotados de una prosodia positiva que incide en "algo destacado" o "algo diferente" por su variedad o exclusividad. Este hecho puede estar relacionado con el carácter elativo que la construcción adopta posiblemente por la presencia de *más*. Sin embargo, no solo adjetivos de prosodia positiva ocupan el *slot* de la construcción, sino que se encuentran además adjetivos de prosodia negativa o neutra. Paralelamente, en una primera prospección de los hápax, se detectan algunas tendencias entre el elemento que ocupa la casilla vacía: anglicismos (*trendey*, *Fashion*, *Heavy*, *family*, *luxury*); adjetivos terminados en -i (guiri, kuki, friki, Krinki, gili, brilli, Insufi, loquis, pichi); onomatopeyas (*Guau*, *Chachachaaannnnnn*, *puffffff*); nombres propios (*David*, *Raquel*, *Shakespeare*, *MTV*, *Fellini*).

El objetivo de la presentación reside en analizar la relación entre productividad y creatividad de la construcción. Para ello se realizará una descripción de la construcción, con especial atención al análisis de las restricciones y preferencias del *slot*. Asimismo, se analizará su productividad y fijación cognitiva en función de su frecuencia *token* y *type*. Además se realizará una exploración sobre las equivalencias posibles en lengua alemana. Para el estudio nos servimos de una metodología deductivo-inductiva basada en corpus. Los corpus monolingües empleados son el esTenTen23 y el deTenTen20, para el español y el alemán respectivamente. En la búsqueda de equivalencias nos servimos del corpus paralelo alemán-español PaGeS, donde adoptamos el método contrastivo unilateral

(Kątny et. al. 2014), según el cual las estructuras de la lengua origen son el parámetro a partir del cual se buscan equivalencias en la lengua meta.

Referencias

García-Paje, M. (1997). Formas de superlación en español: la repetición, *Verba*, *24*, pp.133-157.

Goldberg, A. E. (2006). *Constructions at work. The Nature of Generalitation in Language*. Oxford: University Press.

Kątny, A., Olszewska, D. & Socka, A. (2014). "Kontrastivität in der Linguistik und ihre Dimensionen". En: *Studia Germanica Gedanensia*, 31, pp. 9–23. En línea: https://czasopisma.bg.ug.edu.pl/index.php/SGG/article/view/1479. (Último acceso 17/12/2024).

Mellado Blanco, C. (2020). Esquemas fraseológicos y construcciones fraseológicas en el continuum léxico-gramática. En Tabares, E. et al. (eds.), *Clases y categorías en la fraseología de la lengua española*. Frankfurt a.M.: Peter Lang, pp. 13-36.

PaGeS Parallel Corpus German-Spanish. < https://www.corpuspages.eu/>

Sketch Engine, https://www.sketchengine.eu/

From everyday to technical senses: Polysemy in the legal-administrative language of English-speaking ombudsmen

Gabriel González Delgado

Universidad de Alicante

Polysemous words, those with multiple senses, are not only a prevalent feature in common language, but also in technical domains. Lexicon in specialized languages is normally formed of a) technical jargon, b) everyday language, and c) everyday words that take a technical meaning in that specialized context (Alcaraz et al., 2013). In legal-administrative language, often characterized by precision and formalism (Alcaraz, 2002), polysemous words are frequently used with meanings that diverge significantly from their dominant senses in everyday communication. This phenomenon may be a source of ambiguity for laypeople and, thus, a challenge for comprehension, as it requires individuals to suppress their dominant, everyday understanding of a word to access the intended subordinate sense used within the administrative domain.

Contrary to homonyms, which have dissimilar meanings, different senses in a polysemous word can be closely related, albeit not similar (Foraker & Murphy, 2012). The processing cost associated with polysemous words arises from sense overlap (i.e., low, moderate, or high overlap) among the different related meanings (Klepousniotou et al., 2008). In particular, as reported by the authors, low overlap between senses increases the switching cost from one sense to another, whereas high-overlapping polysemous words don't depend on context to be accurately processed. These results (Foraker & Murphy, 2012; Klepousniotou et al., 2008) support Frazier & Rayner's partial specification hypothesis (1990), which states that "in a neutral context, a reader activates an underspecified, abstract semantic interpretation of the polysemous word that encompasses or underlies both senses" and would restrict the existence of a core meaning only in high-overlap words, also referred to as metonymic polysemy.

This paper explores the prevalence and characteristics of polysemous words in administrative language, drawing on a corpus specifically compiled from annual reports published by English-speaking ombudsmen, as a sample of this language for specific

purposes. Most frequent polysemous nouns and verbs found in this corpus following a bottom-up approach (top 1,000) are classified considering various aspects: 1) the frequency of use —either dominant or subordinate— of each sense in everyday language, 2) the number of different senses these polysemous words have according to dictionaries, 3) the semantic relations between a word's senses, and 4) the degree of overlap among senses. Subordinate senses of polysemous nouns and verbs in this administrative context generally share moderate or high overlap with their dominant everyday counterpart. Results also show that the most frequently used polysemous words are abstract nouns that retain their dominant sense in the administrative domain. Moreover, it is noteworthy the variability and recurrence of polysemous senses between singular and plural instances of the same lemma. By addressing the challenges posed by polysemy in this technical domain, this study aims to contribute to the broader effort of making institutional discourse more accessible, equitable, and efficient for all users.

References

Alcaraz, E. (2002). *El inglés jurídico: textos y documentos*. Ariel.

Alcaraz, E., Campos, M. A., & Miguélez, C. (2013). *El inglés jurídico norteamericano*. Ariel España.

Foraker, S., & Murphy, G. L. (2012). Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of Memory and Language*, 67(4), 407-425.

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of memory and language*, 29(2), 181-200.

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: the comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6), 1534.

Lemmatisation of Anglo-Saxon nominal compounds

Yosra Hamdoun Bghiyel

Universidad de La Rioja

The lemmatisation of Old English poetic texts represents a fundamental challenge in the study of historical linguistics, lexicography, and corpus linguistics. Compound nouns, in particular, pose unique difficulties due to their complex morphology and variation in attested spellings. Despite significant advancements in digital resources and lexicographical tools, there remains a notable gap in the systematic treatment of compound nouns within historical language corpora. The present study addresses this gap by exploring how a relational database framework can be used to assign unified headwords to compound nouns in *the York-Helsinki Parsed Corpus of Old English Poetry* (YCOEP; Pintzuk and Plug, 2001) under unified headwords, using a relational database framework. This relational database integrates several online reference dictionaries such as *The Dictionary of Old English* (DOE; Healey et al. 2018) and Bosworth-Toller's *An Anglo-Saxon Dictionary* (2014) with several annotated databases and additional contrasting datasets. By correlating these sources, the headwords and attested spellings, the IOED provides a streamlined, comprehensive framework for systematically comparing and selecting the most suitable lemma for each compound noun. The methodology involves four systematic steps to complete the lemmatisation of compound nouns in the YCOEP. First, the YCOEP inventory of attested spellings for compound nouns is compiled into a structured database for classification. Second, these spellings are matched with sources

from the relational database to assign headwords automatically. Third, the assigned headwords are aligned with our database containing the final lemma list for consistency. Finally, the resulting lemmatised inventory is validated against main completed or semicompleted sources of reference: the DOE, VariOE, and Bessinger's *A Short Dictionary of Anglo-Saxon Poetry* (1960) which includes over 5.000 entries, to identify and assess any discrepancies in headword selection. The main conclusions of this research highlight the feasibility of a corpus-based approach to fully lemmatising Old English compound nouns as well as the significant contributions of adopting a relational database framework to the study of this lexical category. Preliminary findings include the creation of 21 new lemmas for compound nouns, and a listing of the lemmatised poetic OE poetic exclusive vocabulary. These contributions not only enhance the lexicographical understanding of Old English compound nouns but also provide a replicable model for similar studies in historical linguistics.

References

Bessinger, J. B. (1960). *A short dictionary of Anglo-Saxon poetry*. University of Toronto Press.
Bosworth, J., & Toller, T. N. (1973). *An Anglo-Saxon dictionary*. Oxford University Press.
Pintzuk, S., & Plug, L. (Comps.). (2001). *The York-Helsinki Parsed Corpus of Old English Poetry*. Retrieved from http://www-users.york.ac.uk/~lang18/pcorpus.html
Healey, A. diPaolo, Wilkin, J. P., & Xiang, X. (2009). *Dictionary of Old English Web Corpus. Dictionary of Old English Project*, Centre for Medieval Studies, University of Toronto.

On the relation between word length and phoneme frequencies

Ján Mačutek, Radek Čech, Michaela Koščová

Mathematical Institute, Slovak Academy of Sciences, Slovakia and Department of Mathematics, Constantine the Philosopher University in Nitra, Slovakia / Department of Czech Language, Faculty of Arts, Masaryk University, Czech Republic / Mathematical Institute, Slovak Academy of Sciences, Slovakia

According to the principle of least effort (Zipf, 1949), people tend to use the least amount of effort to convey their message, seeking thus to maximize the efficiency of communication. Zipf observed this principle in language usage, particularly in the frequency distribution of words (shorter words are used more often). The principle assumes that people will (statistically, with possible exceptions) prefer the simplest, most frequent, and most efficient means of communication, relative to the context and constraints. This tendency is not limited to words, a higher occurrence of shorter units can be observed also in other domains (see e.g. a short overview in Mačutek, 2022, p. 414).

However, the principle of least effort has an impact also on structural properties of language units. The Menzerath-Altmann law (Cramer, 2005), which says that longer words consist, on average, of shorter syllable, is a well-known example (if one uses longer words, words with simpler structure, in this case with shorter syllables, are preferred).

In our presentation, we will show another exemplification of the principle of least effort, namely, that longer words tend to contain phonemes that are easier to pronounce. While the negative geometric distribution serves as a general model for ranked frequencies of phonemes (and also for graphemes) from corpora in many languages (Grzybek et al, 2009; Grzybek & Rusko, 2009), we will show that there are differences among frequencies of particular phonemes in words of different lengths. These differences can be explained by preferring in longer words those phonemes that require less effort to

utter. To put it simply, when one wants to use or has to use a long word (which requires more effort than a short word), one saves effort by using "easier" phonemes (see e.g. Stoel-Gammon, 2010, for one of possible approaches to differentiating between easier and more demanding phonemes). This observation is valid in several languages from different language families.

Our presentation opens another field where the principle of least effort is involved – obviously, length (of words and syllables at least) is an important factor, but also properties of lower-level units of which word consists also play a significant role. This principle thus seems to be one of important "language forces" which shape properties of language units, such as frequency, length, and structure.

References

Cramer, I.M. (2005). Das Menzerathsche Gesetz. In R. Köhler, G. Altmann & R.G. Piotrowski (Eds.), *Quantitative linguistics. An international handbook* (pp. 659–688). De Gruyter.

Grzybek, P., Kelih, E., & Stadlober, E. (2009). Slavic letter frequencies: A common discrete model and regular parameter behavior? In R. Köhler (Ed.), *Issues in quantitative linguistics* (pp. 17–33). RAM-Verlag.

Grzybek, P., & Rusko, M. (2009). Letter, grapheme and (allo-)phone frequencies: the case of Slovak. *Glottotheory*, 2(1), 30–48.

Mačutek, J. (2022). Why do parameter values in the Zipf-Mandelbrot distribution sometimes explode? *Journal of Quantitative Linguistics*, 29(4), 413–424.

Stoel-Gammon, C. (2010). The word complexity measure: Description and application to developmental phonology and disorders. *Clinical Linguistics & Phonetics*, 24(4-5), 271–282.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.

¡Qué congreso ni qué leches!
Análisis de corpus de la construcción fraseológica de RECHAZO [X/ ¡*Qué* X *ni qué* X/Y!]

Carmen Mellado Blanco, Nely M. Iglesias

Universidade de Santiago de Compostela / Universidad de Salamanca

El objetivo principal de nuestra propuesta es describir, a través del análisis de corpus y mediante una metodología inductiva, la construcción binomial exclamativa [X/ ¡Qué X *ni qué* X/Y!] (Olza 2011; Martí Sánchez 2020: 118), que forma parte de una familia de construcciones de rechazo y desacuerdo con la conjunción *ni* (cf. Pérez Salazar 2017; Padilla Herrada 2020). Con respecto a los dos *slots* que contiene, presenta un elemento ecoico X, que remite al discurso anterior y alude a la proposición rechazada por el hablante, y una segunda variable, que puede ser igual a X o distinta (Y). En el primer caso, al darse la repetición del elemento presente en el primer *slot* X (ejemplo 1), se trata de una construcción reduplicativa. En el segundo caso, la construcción licencia, a través de la segunda variable Y, lexemas de naturaleza figurada, carentes de significado referencial, que tienen la función de reforzar la negación, ya implícita en la primera parte del binomio (ejemplos 2a-2b). Además, destaca el uso recurrente del adverbio *nada* (ejemplo 2c). En el segundo patrón, los lexemas actualizados por Y incluyen con frecuencia disfemismos (*hostias*, *cojones*, *mierdas*, *carajo*) (ejemplo 2a), así como expresiones fraseológicas lexicalizadas como *ocho cuartos y niño muerto* (ejemplo 2b).

Como muestra, valgan estos ejemplos extraídos del corpus *esTenTen23* (Sketch Engine):

(1) La periodista ha compartido una fotografía en la que aparece bellísima. "¿Qué filtro ni qué filtro?", escribe para señalar que ella no necesita ningún retoque para estar guapa. (4570390)
(2a) ¡Pero qué contencioso ni qué hostias! Esto son trapos sucios y se lavan en esta casa. (62513868)
(2b) ¿Qué vacío legal ni qué niño muerto? Yo me bajo y me he bajado siempre todo lo que me ha dado la real gana, y soy cliente de Netflix desde el primer día que abrió en España. (1329319)
(2c) Qué libros ni qué próceres ni qué nada. ¿No les parece que hay en el comportamiento de un peatón caraqueño frente a un semáforo bastantes paralelismos con nuestra forma de entender las normas en general, y los trámites burocráticos en particular? (57463590)

Tomando la Gramática de Construcciones como marco teórico (Goldberg 2019), nuestro análisis de corpus, tanto cualitativo como cuantitativo, se centra en las actualizaciones léxicas de las dos aloconstrucciones [X/ ¡*Qué* X *ni qué* X!], así como [X/ ¡*Qué* X *ni qué* Y!], el cual nos permitirá determinar el grado de fijación cognitiva (*entrenchment*) y de productividad de la construcción, además de identificar los posibles patrones de su potencial creativo y los constructos lexicalizados o en vías de lexicalización (Mellado Blanco 2020). Finalmente, a nivel discursivo, se describirán las principales funciones de esta construcción dialogal (Fuentes Rodríguez 2023).

Referencias

esTenTen23, The Sketch Engine. http://www.sketchengine.co.uk
Fuentes Rodríguez, Catalina (2023): "Construcciones exclamativas de rechazo". Spanish in Context 20/1: 178 – 207. DOI: https://doi.org/10.1075/sic.21026.rod.
Goldberg, Adele (2019): *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*. Princeton: Princeton University Press.
Martí Sánchez, Manuel (2020): "Construcciones fraseológicas y frasemas gramaticales con "ni" incoordinado". *Romanica Olomucensia* 32/1: 111-126, doi: 10.5507/ro.2020.006.
Mellado Blanco, Carmen (2020): "(No) me importa un comino y sus variantes diatópicas. Estudio de corpus desde la gramática de construcciones". *ELUA. Estudios de Lingüística. Universidad de Alicante* 7: 89-111. DOI: 10.14198/ELUA2020.ANEXO7.06.
Pérez Salazar, Carmela (2017): "*Ni por lumbre*: modelo fraseológico para la negación y el rechazo en la historia del español". En Carmen Mellado Blanco / Katrin Berty / Inés Olza (eds.): *Discurso repetido y fraseología textual (español y español-alemán)*. Madrid/Frankfurt a. M.: Iberoamericana/Vervuert, 269-298.
Olza Moreno, Inés (2011): "*¡Qué fraseología ni qué narices!* Fraseologismos somáticos del español y expresión del rechazo metapragmático". En Antonio Pamies Bertrán, Juan de Dios Luque Durán / Patricia Fernández Martín (eds.): *Paremiología y herencia cultural*. Granada: Granada Lingvística/Método Editores, 181-191.
Padilla-Herrada, María Soledad (2020): "Expresiones de rechazo introducidas por «ni» + constituyente no oracional". *Rilce. Revista De Filología Hispánica* 36(3): 1165-1192. https://doi.org/10.15581/008.36.3.1165-92

From intuition to computation: A comparative study of semantic categorization methods

Daniela Pettersson-Traba, Iván Tamaredo Meira

Universidad Complutense de Madrid

The purpose of this paper is to compare two approaches to semantic classification which share the goal of grouping semantically similar words into more general categories but achieve this in different ways. The referent-based approach to semantic analysis (Geeraerts et al., 1994) focuses on the properties of referents: words that denote concepts which share properties are classified under the same category, while words that denote unrelated concepts are assigned to different categories. Since the properties of referents used to analyze words are "manually and preliminarily collected on the basis of real-world familiarity" with the referents (Geeraerts et al., 2023, p. 35), the referent-based approach is a top-down method of semantic classification. As such, this method can also be introspective, at least in some of its applications, as in many cases the categories are selected by the researcher beforehand on the basis of his/her own linguistic intuitions (e.g., Gries & Otani, 2010; Liu, 2010).

The referent-based approach is, however, not the only method of semantic categorization. Over the last decades, many studies have been published that implement the distributional approach set out by Firth. Following Firth's famous quote "[y]ou shall know a word by the company it keeps" (1957, p. 11), words are semantically related if they share collocates. Since the relevant collocates emerge directly from the data, the distributional approach can be characterized as a bottom-up, data-driven, and quantitative method of semantic categorization, particularly since scholars have recently resorted to techniques developed in the field of computational linguistics, mainly vector space modelling, to analyze words in a distributional manner (De Pascale, 2019; Geeraerts et al., 2023).

Even though both the referent-based and the distributional approaches figure prominently in the literature as methods of semantic categorization, little attention has been devoted to whether they result in (in)compatible classifications. The present paper aims to fill this gap. To this purpose, adjectives from the domain of SMELL are used as a case study and both referent-based and distributional categorization methods are applied to classify their collocates. Drawing on *COHA* (Davies, 2010), the noun collocates of the adjectives are semantically categorized as follows. First, a referent-based classification is implemented by resorting to the semantic taxonomies available in the USAS (Archer et al., 2002) and the *HTOED*. Second, a classification based on type-based vector space modelling is applied following the steps in De Pascale et al. (2018). Finally, the results of the two methods are compared to answer the following questions: Do the referent-based and distributional methods result in similar groupings of nouns? What is the optimal number of categories according to each method? How well do the resulting categorizations distinguish between adjectives from the domain of smell? The present paper, therefore, sheds light on the adequacy of existing approaches to semantic classification by comparing the results of the two main prominent methods when applied to the same dataset.

References

Archer, D., Wilson, A., & Rayson, P. (2002). *Introduction to the USAS category system*. [Unpublished manuscript]. University Centre for Computer Corpus Research on Language. https://ucrel.lancs.ac.uk/usas/usas_guide.pdf

COHA = Davies, M. (2010). *The Corpus of Historical American English (COHA): 400 Million Words*, 1810-2009. https://corpus.byu.edu/coha/

De Pascale, S. (2019). *Token-based vector space models as semantic control in lexical lectometry*. [Unpublished doctoral dissertation]. KU Leuven.

De Pascale, S., Marzo, S., & Speelman, D. (2018). Cultural models in contact: Revealing attitudes toward regional varieties of Italian with Vector Space Models. In E. Zenner, A. Backus, & E. Winter-Froemel (Eds.), *Cognitive Contact Linguistics: Placing Usage, Meaning and Mind at the Core of Contact-Induced Variation and Change* (pp. 213–250). De Gruyter Mouton. https://doi.org/10.1515/9783110619430

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In J. R. Firth (Ed.), *Studies in Linguistic Analysis*, (pp. 1–32). Blackwell.

Geeraerts, D., Grondelaers, S., & Bakema, P. (1994). *The Structure of Lexical Variation: Meaning, Naming, and Context*. De Gruyter Mouton. https://doi.org/10.1515/9783110873061

Geeraerts, D., Speelman, D., Heylen, K., Montes, M., De Pascale, S., Franco, K., & Lang, M. (2023). *Lexical Variation and Change: A Distributional Semantic Approach*. Oxford University Press. https://doi.org/10.1093/oso/9780198890676.001.0001

Gries, S. Th., & Otani, N. Behavioral profiles: a corpus-based perspective on synonymy and antonymy. *ICAME Journal* 34, 121–150.

HTOED = Historical Thesaurus of the Oxford English Dictionary. http://www.oed.com/thesaurus

Liu, D. (2010) Is it a *chief*, *main*, *major*, *primary*, or *principal concern*? A corpus-based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics* 15(1), 56–87. https://doi.org/10.1075/ijcl.15.1.03liu

Recursos terminológicos sensibles al género: Corpus e inteligencia artificial en la salud de la mujer

Chelo Vargas-Sierra, Antonio Moreno Sandoval

Universidad de Alicante / Universidad Autónoma de Madrid

La terminología biomédica sobre la salud femenina ha sido tradicionalmente un reflejo de enfoques estigmatizantes y marcados por sesgos de género, lo que ha dificultado una representación de las cuestiones específicas de salud femenina de manera inclusiva y precisa. Además, la carencia de recursos especializados y accesibles en entornos multilingües limita el acceso a información terminológica fiable. Este trabajo analiza los avances de un proyecto orientado al desarrollo de recursos terminológicos digitales multilingües centrados en la salud femenina, con un enfoque basado en corpus y sensibilidad de género. Los principales objetivos incluyen neutralizar sesgos, promover representaciones positivas de la mujer y mejorar la visibilización de temas clave de salud femenina mediante el diseño de fichas terminológicas específicas, la revisión de definiciones y la selección de contextos no estigmatizantes. El proyecto también busca ofrecer herramientas eficaces para su aplicación en entornos biomédicos y académicos. La metodología adoptada combina enfoques tradicionales y tecnológicos, siendo el diseño, compilación y explotación de corpus especializados un pilar central. Los corpus han permitido identificar y analizar términos relevantes, mientras que los modelos de inteligencia artificial (IA) se han incorporado para optimizar la labor terminográfica. Estas tecnologías han facilitado la creación de recursos que integran información terminológica enriquecida con sensibilidad de género. Los resultados muestran que los recursos creados son efectivos para abordar las problemáticas identificadas, mejorando significativamente

la representación terminológica de la salud femenina. En particular, el banco terminológico generado no solo contribuye a la visibilización de los temas de salud explorados, sino que también sienta las bases para futuras aplicaciones en otros dominios especializados, destacando el potencial de combinar corpus, IA y sensibilidad de género en la terminología biomédica.

Referencias

Ädel, A. (2020). Corpus compilation. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. XX-XX). Springer. https://doi.org/10.1007/978-3-030-46216-1_1

Ancochea, J., Izquierdo, J. L., & Soriano, J. B. (2021). Evidence of gender differences in the diagnosis and management of Coronavirus Disease 2019 patients: An analysis of electronic health records using natural language processing and machine learning. *Women's Health*, 30(3), 393–404. https://doi.org/10.1089/jwh.2020.8721

Baškot, B., & Barišić, M. (2022). Web scraping analysis: Gender differences in local online media mentions. *PaKSoM* 2022, 381–387.

Ruiz Cantero, M. T. (Coord.). (2019). *Perspectiva de género en medicina*. Fundación Dr. Antoni Esteve. Retrieved from https://www.esteve.org/libros/perspectiva-de-genero-en-medicina/?doing_wp_cron=1670839661.3348269462585449218750

Valls-Llobet, C. (2021). *Mujeres invisibles para la medicina: Desvelando nuestra salud.* Capitán Swing.

Vargas-Sierra, C. (2024). Un proyecto de terminología sensible al género en temas de salud de la mujer: Conceptos e innovaciones. In *Perspectivas para la visibilización del género* (pp. 453–465). Peter Lang.

Vargas-Sierra, C., & Moreno-Sandoval, A. (Eds.). (2024). *Women and metaphors: Terminology, lexicon and representations of women's health in biomedical discourse. Cultura, Lenguaje y Representación (CLR), 34.*

Panel 5

Corpora, contrastive studies and translation
Corpus, estudios contrastivos y traducción

A corpus-based study of meaning in banking complaints on social media: A contrastive approach in English and Spanish from systemic functional linguistics

Yolanda Blázquez-López

Universidad Politécnica de Madrid

Speakers provide clues that help others infer their intentions, which is essential for effective communication. The study of intentions modelling is increasingly significant in banking, where managing customer complaints and understanding user needs are crucial for customer satisfaction and reputation. While computational models have advanced classifying intentions, challenges remain in capturing context-specific meanings. To this end, we conducted a qualitative study, involving a contrastive analysis of banking complaints in English and Spanish on X (formerly Twitter). We aimed to understand how individuals express their banking complaints in both languages, focusing on the explicit and implicit meanings behind their intentions and the comparison of the lexico-grammatical patterns that characterise them. Thus, we began our research by manually compiling two corpora from the corporate customer support profiles of twelve retail banks on X. The first corpus included 1,400 tweets in Spanish from six banks operating in Spain (i.e. Banco Santander, BBVA, Caixabank, Banco Sabadell, ING, and Openbank), while the second contained 1,500 tweets in English from six banks operating in the UK (i.e. Barclays, Santander UK, HSBC, Lloyds Bank, Monzo and Starling Bank). Each corpus featured both the customer's post and the bank's response. We then classified the tweets using a deep taxonomy of intentions we developed using our domain knowledge. With this taxonomy, we conducted our contrastive analysis central to this study. Therefore, we started by formulating four research questions based on our collection of tweets within each intention: (Q1) What are the similarities and differences in how customers encode their beliefs about banks through complaints in both languages? (Q2) How do the participants' roles in each context differ by language based on linguistic evidence? (Q3) Can we identify consistent patterns of conversational structures in both languages that suggest the two corpora share the same genre? (Q4) What are the main lexico-grammatical differences between both languages, based on customers' intentions? To address these questions, we adopted a methodology based on the Systemic Functional Linguistics (SFL) approach developed by Halliday (1994), which enabled us to contrast both corpora by examining the two primary strands of language involved in social interactions, i.e. external and internal. In the external strand, we examined differences and similarities in contextual features like ideology (i.e. consumers' beliefs about banking), genre (i.e. conventional conversational structures), and register configuration (i.e. field, tenor, mode). In the internal strand, we compared the different strategies customers used to construct three types of meaning in both languages (i.e. ideational, interpersonal, and textual) from their respective lexico-grammar patterns (i.e. transitivity, mood, and theme). We can conclude that our study contributes to a better understanding of customer complaints on social media in English and Spanish, achieving two main outcomes: (i) deep insights into how individuals express complaints in both languages, highlighting explicit and implicit meanings and comparing their lexico-grammatical and contextual patterns, and (ii) a new approach based on SFL to

improve annotation of training corpora used by computational intention modelling systems in multilingual contexts. Furthermore, our study can be replicated in other languages and domains.

References

Bateman, J., McDonald, D., Hiippala, T., Couto-Vale, D., & Costetchi, E. (2019). Systemic Functional Linguistics and Computation: New Directions, New Challenges. In G. Thompson, W. L. Bowcher, L. Fontaine, & D. Schönthal (Eds.), *The Cambridge Handbook of Systemic Functional Linguistics* (pp. 561–586). Cambridge University Press. https://doi.org/10.1017/9781316337936.024.

Eggins, S. (2004). *An Introduction to Systemic Functional Linguistics* (2nd ed.). Bloomsbury Publishing.

Halliday, M. A. K. (1994). *Introduction to Functional Grammar* (2nd ed.). Edward Arnold.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2013). *Halliday's Introduction to Functional Grammar* (4th ed.). Routledge.

Huang, Q., Xia, X., Lo, D., & Murphy, G. C. (2020). Automating intention mining. *IEEE Transactions on Software Engineering*, 46(10), 1098-1119. https://doi.org/10.1109/TSE.2018.2876340.

Lirola, M. M. (2007). *Aspectos Esenciales de la Gramática Sistémica Funcional*. Universidad de Alicante.

Martin, J. R., & Rose, D. (2007). *Working with Discourse: Meaning Beyond the Clause* (2nd ed.). Bloomsbury Publishing.

Németh T., E. (2020). Linguistic and Contextual Clues of Intentions and Perspectives in Human Communication. In L. Hunyadi, I. Szekrényes (Eds.), *The Temporal Structure of Multimodal Communication. Intelligent Systems Reference Library* (pp. 3–21). Springer. https://doi.org/10.1007/978-3-030-22895-8_1.

Schleppegrell, M. J., & Oteíza, T. (2023). Systemic Functional Linguistics: Exploring meaning in language. In M. Handford & J. P. Gee (Eds.), *The Routledge Handbook of Discourse Analysis* (pp. 156-169). Routledge.

Stauss, B., & Seidel, W. (2019). *Effective Complaint Management: The Business Case for Customer Satisfaction*. Springer. https://doi.org/10.1007/978-3-319-98705-7.

Zhang, Y., Singh, S., Sengupta, S., Shalyminov, I., Su, H., Song, H., & Mansour, S. (2024). Can Your Model Tell a Negation from an Implicature? Unravelling Challenges with Intent Encoders. *Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.48550/arXiv.2403.04314.

Emotions in 18th and 19th century correspondence: From Irish English to Portuguese and Spanish through CORIECOR and post scriptum (1750-1833)

Jesús Candelario Menacho

Universidad de Extremadura

Since the 1980s, scholars have increasingly focused on the language of letters in a wide variety of ways, ranging from historical and sociolinguistic to even economic perspectives. However, despite the substantial scholarly attention given to many aspects of the epistolary genre, the emotional content on letters has not received special treatment in terms of quantitative analysis. So, the primary aim of this paper is to address this gap by contributing an in-depth empirical analysis of emotions in the epistolary genre, not only within the Irish context but also across the Portuguese and Spanish languages.

The data found in the Corpus of Irish English Correspondence (CORIECOR) facilitates the study of emotions that would otherwise be unattainable. This corpus enables "not only internal linguistic research but also historical sociolinguistic work" (McCafferty & Amador-Moreno 2012: 268). The results will be compared with those obtained from a Portuguese and Spanish letters corpus entitled Post Scriptum, which consists of private letters written in Portugal and Spain during the Early Modern period (CLUL 2014). The time span for this study will be 1750-1833 to ensure compatibility with CORIECOR. These epistles deal with a broad range of subjects, making them rich in detail. The comparison of the two corpora will shed some light on how emotions are linguistically constructed in these different cultural contexts.

Therefore, this study proposes some key working hypotheses. First, it is hypothesized that the letters will cover a wide range of emotional topics, including homesickness, separation, migration, health, and work, among many others. In addition, it can be expected to find money as one of the most prominent themes in this study, given that migration often occurred due to economic reasons and is interconnected with the topics mentioned before. These hypotheses aim to address some of the following questions: What emotional content characterizes letters in Ireland, Portugal, and Spain? What role do emotions play in these letters? Do any of these topics determine the ways emotions are expressed?

The quantitative data will undergo statistical testing to validate the results, which will also be analyzed qualitatively. Linguistic Inquiry and Word Count (LIWC-22) will be used to "analyze others' language and understand their thoughts, feelings, personality, and the ways they connect with others" (Pennebaker et al. 2022). When necessary, additional statistical tools such as Antconc (Anthony 2023) or Lancsbox (Brezina, Weill-Tessier & McEnery 2021) will be applied to gain a fuller insight of the linguistic phenomena. These tools will help explore the interaction of different variables and assess their significance.

References

Anthony, L. (2023). *Antconc* (4.2.4). Tokio: Waseda University.

Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). *#LancsBox v. 6.x*. [software package].

CLUL (Ed.). (2014). *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. [13-12-2024]. URL: http://ps.clul.ul.pt.

McCafferty, K., & Amador-Moreno, C. P. (2012). A Corpus of Irish English Correspondence (CORIECOR): A tool for studying the history and evolution of Irish English. In B. Migge & M. Ní Chiosáin (Eds.), *New perspectives on Irish English* (pp. 265–287). Amsterdam: John Benjamins.

Pennebaker, J. W., Boyd, R. L., Booth, R. J., Ashokkumar, A., & Francis, M. E. (2022). *Linguistic Inquiry and Word Count: LIWC-22*. Pennebaker Conglomerates. https://www.liwc.app

Repeat or diversify? A multi-factorial study of English-to-Polish translation of reporting verbs in literary novels

Łukasz Grabowski

University of Opole, Poland

In the last decade or so, corpus linguistics helped translation scholars enhance methodological rigour, contributing to reproducibility and replicability of research findings. It has also opened up and rejuvenated research areas, such as stylistics,

translation universals, translationese or translatorese, to name but a few. Examining repeated patterns of language use is a cornerstone of much corpus linguistic research, also oriented at translation, where repetition – manifested on several linguistic levels (morphological, syntactic, lexical etc.) – plays a very important role, notably in literary texts.

Capitalizing on a preliminary research (Mastropierro & Grabowski, 2024), this study focuses on the identification of the predictors of repetition or lexical variety in the translation of reporting verbs from English into Polish. Using a sample of 20 literary novels, we fit multiple negative binomial regression with mixed effects to assess the effect that selected predictor variables (i.e. frequency of a source-text verb, number of its translation equivalents in lexical databases, its number of senses, type of verb, length in characters, date of translation, and translators) have on the response variable: the number of different target language verb types a source text reporting verb is translated into Polish. In short, if the number of types is high then it means that translators opted for lexical variety (i.e. used various TT reporting verbs as translation equivalents of a ST verb). Reporting verbs were retrieved using CQL queries from InterCorp v.15 (Čermák & Rosen 2012; Čermák 2019), a large multilingual parallel corpus which includes, among others, English novels and their translations into Polish.

The overall model fit per the lowest AIC and BIC values obtained through backward elimination reveals that semantic category of a ST reporting verb, its frequency and length as well as the translator as a random effect have the largest individual contributions to explaining the proportion of variation in the response variable in the Polish translations (marginal r-squared = 0.76, conditional r-squared = 0.79). More precisely, the model allows us to explain almost 80% of the variation in the response variable, that is, the number of different verb types a ST verb is translated into. We also identified semantic types of English ST verbs whose Polish equivalents tend to be consistently repeated in translation or whose Polish equivalents are varied. The low level of variance (0.03) in the random effect means that the impact of individual translators is relatively similar. In other words, there is some variability between the translators, but it is relatively small, and no single translator significantly influenced overall results. We also found that Polish translators prefer a variety of reporting verb equivalents of English neutral reporting verbs (e.g. said, told), which is in sync with the dominant stylistic conventions in Polish dictating that repetition be avoided in texts. The presentation concludes with a presentation of limitations of the study and an outline of avenues for future research, where the methodology could be easily adopted to other language pairs, including English-to-Spanish or Spanish-to-English translation.

References

Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, *17*(3), 411–27.

Čermák, P. (2019). "InterCorp: A parallel corpus of 40 languages". In I. Doval & M. Sanchez-Nieto (eds), *Parallel Corpora for Contrastive and Translation Studies: New resources and applications* (pp. 93-101). Amsterdam: John Benjamins.

Mastropierro, L. & Grabowski, Ł. (2024). Repeated reporting verbs in English novels and their Italian and Polish translations: a preliminary multifactorial study. In M. Kajzer-Wietrzny (Ed.), Corpus-based Translation and Interpreting Studies. Special issue of *Across Languages and Cultures. A Multidisciplinary Journal for Translation and Interpreting Studies*, *25* (2), 310–330.

Exploring the semantics of phraseology in food and drink promotional texts

Leticia Moreno-Pérez

Universidad de Valladolid - ACTRES

The relationship between language and food, known as *semiofoodscape* (Järlehed and Moriarty, 2018), is the perspective through which food semiotic landscapes can be examined. Semiofoodscapes are dependent on "the pragmatic components involved" (López-Arroyo and Sanz-Valdivieso, 2024, p. 84), as the different product types, actors and inscriptional genres may have certain linguistic features in common. This is particularly relevant when it comes to the subjectivity displayed in the process of description and evaluation: after facing the challenge of finding the proper words to describe perceptions comes the challenge of meaning reaching others the same way, as "the individual nature of tasting means that descriptors may be perceived differently among various tasters" (López-Arroyo and Sanz-Valdivieso, 2024, p. 84). This variation is one of the reasons why the semantics of food and drink is so challenging and has been explored from different perspectives and approaches. However, the process actors handle meaning in this sensory process is still a question under study.

This paper intends to describe the semantic characteristics of one of the key linguistic features in specialized domains: multiword expressions. These constructs, which are based on semantic relationships, pose further challenges to linguists, mainly related to their conceptualization, classification, and scope (Granger and Paquot, 2008; Gray and Biber, 2015, among others), but are essential in professional communication. Their description will be carried out using the data present in the CLANES English/Spanish comparable corpus, a multi-layer corpus on promotional texts including different subgenres for wine, cheese, tea, dried meats and biscuits compiled by the ACTRES Research Group with more than 1.8 million words (Sanjurjo-González, forthcoming).

First, we will explain and justify the tagging process of MWE-related data since it was tailor made due to the semantic complexity of the domain. This was followed by a thorough standardization process, necessary to ensure systematicity and the consistency validation carried out to prepare data. Next, we will focus on the analysis of the semantic patterns found in the different subgenres of the corpus, whose differences and similarities will be described in detail, contrasting both languages.

Findings show that the different products described in the subgenres under study offer unique semiofoodscape elements, and overarching trends and innovations in the use of semiotic resources across different food categories have been identified. The study confirms that the semantics of food and drink MWEs in English and Spanish is mostly transparent, as meanings mostly come from the sum of the individual parts of the expressions; also, although no prototypical semantic patterns were found in any of the two languages, it was observed that Spanish and English handle descriptions and categorizations differently, as the focus of the former shows a more specialized point of view, while the latter aims at a broader audience. These aspects seem to be essential for successful international professional communication.

References

Granger, S. & Paquot, M. (2008). Disentangling the Phraseological Web. In S. Granger & F. Meunier (Eds.), *Phraseology. An Interdisciplinary Perspective* (pp. 27-49). John Benjamins.

Gray, B. & Biber, D. (2015). Phraseology. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 125-145). Cambridge University Press.

Järlehed, J. & Moriarty, M. (2018). Culture and class in a glass: Scaling the Semiosfoodscape. *Language & Communication*, 62, 26-38.
López-Arroyo, B. & Sanz-Valdivieso, L. (2024). A Hint of the Summery Goodness of Green Grass: A Look at English Descriptors in Tasting Notes. *Revista de Lingüística y Lenguas Aplicadas*, 19, 84-103.
Sanjurjo-González, H. (Forthcoming). CLANES: A Multilayer English-Spanish Comparable Corpus. In R. Rabadán & E. Ramón (Eds.), *Cross-linguistic Mediated Communication: Hybrid Text Production English-Spanish*. Peter Lang.

Meaning and persuasion through adjectives: A corpus-based comparison of original and translated food product descriptions

Isabel Pizarro Sánchez

Universidad de Valladolid

Adjectives play an important role in commercial discourse. They serve as persuasive devices that shape meaning and enhance communication by constructing a positive and appealing representation of products. This research investigates the use of adjectives in English-language descriptions of bakery, pastry, and biscuit products —a sector that generates over €2.5 billion in exports in Spain— comparing original texts with translations from Spanish.

Focusing on the 'description' move within the online product description genre, the study addresses the following questions: What are the differences and similarities in the frequency, variety, morphology, and semantic typology of adjectives used in original English texts and their Spanish-to-English translations? To what extent do the English translations reflect the persuasive strategies typical of Spanish or adopt strategies specific to English, and what are the implications for translation?

Using a mixed methods research approach, the study analyses a corpus of 300 texts (100 original English texts, 100 Spanish texts, and 100 English translations) to explore the frequency, variety, morphology, and semantic typology of adjectives. The quantitative phase involved examining the corpus with Sketch Engine to identify adjectives, measure their frequencies (both types and tokens), and explore their semantic profiles through the concordance and word sketch tools. The qualitative phase complements this analysis by interpreting the concordance lines and semantic patterns in context, providing a deeper understanding of the adjective's role in constructing persuasive and culturally specific messages. The semantic classification was adapted from Dixon (1982) and Edo Marzá (2011), focusing on how adjectives convey meaning and persuasion in commercial contexts.

The findings aim to determine whether translations follow the persuasive strategies of the source language (Spanish) or adapt to the conventions of the target language (English). According to preliminary results, the original English texts and the translations differ in the frequency and variety of adjectives. Semantically, both sets of texts rely heavily on evaluative adjectives. However, the translations occasionally introduce subtle shifts in meaning due to linguistic transfer. These findings highlight the difficulties of maintaining persuasiveness in translation and suggest the need for strategies that balance fidelity to the source text with the linguistic norms of the target language.

This study contributes to the broader discussion of meaning in corpus linguistics by demonstrating how corpus-based methods can reveal nuanced patterns in the use of linguistic resources across original and translated texts. The limitations of the study include the size of the corpus and the focus on a single product category, which may affect

the generalisation of the findings. Future studies may include more products in the corpus and explore the role of other linguistic features in constructing meaning and persuasion.

References

Dixon, R. M. W. (1982). *Where have all the adjectives gone? And other essays in semantics and syntax*. De Gruyter Mouton.

Edo Marzá, N. (2011). *The specialised lexical component of the language of tourism: A corpus-based study of secondary term formation in English and Spanish* (Doctoral dissertation, Universitat Jaume I)

Comparing the expression of quality assurance in English and Spanish online cheese descriptions

Noelia Ramón, Belén Labrador

Universidad de León

Industries that want to promote their products abroad must write their texts in English, currently the global language used in trade. Companies engaging in international business need to rely on accurate and idiomatic texts to convince potential customers to buy their products (Bhatia & Bremner, 2012; Sing, 2017). This paper will focus on one type of promotional text from the food and drink industry: online cheese descriptions. Promotional texts in this domain include aspects, such as the product presentation, its elaboration process, and technical characteristics. One of these aspects is the expression of quality assurance to increase the persuasive strength of promotional texts. In the case of cheeses, this quality assurance is represented by objective evidence such as awards, medals and certifications obtained by the cheese, as well as comments by experts in the field. A good knowledge of how to operate the expression of quality assurance in English will greatly benefit non-native professionals involved in promoting their products online (Izquierdo & Pérez-Blanco, 2020). In online promotional texts in the cheese industry references to quality standards are essential, so Spanish-speaking professionals in the field will need to write these comments in English. The aim of this paper is to identify the various lexical and phraseological resources used by manufacturers and retailers in English and Spanish to express quality assurance, and compare the differences in use.

The analysis makes use of an English-Spanish comparable corpus of Online Cheese Descriptions (135,213 words in English vs. 145,254 words in Spanish). A preliminary study on a sample of this corpus has investigated the rhetorical structure of this subgenre and has shown that the move of quality assurance is given greater importance in English than in Spanish: both in number of words (10% vs. 1%) and in number of texts containing this move (50% vs. 10%). This finding has prompted further research on the expression of quality assurance in the whole corpus. The corpus contains online texts extracted from different types of British and Spanish websites dealing with cheeses: 600 text in English and 400 texts in Spanish. The corpus was compiled, tagged and explored using software specifically designed for our purposes. The semantic tagging was carried out following the annotation scheme called USAS (UCREL Semantic Analysis System) (Rayson et al., 2004). For this contrastive study, the instances with the semantic labels of A5.1 Evaluation: good/bad, A5.4 Evaluation: authenticity and S7.3 Competition are analysed and compared. All the concordance lines are investigated from a lexical and phraseological perspective to extract the main linguistic patterns in each language used to express objective quality features of cheeses. These resources contribute to expressing a positive

evaluation of the cheeses to be sold (Hunston & Sinclair, 2000), thus strengthening the persuasive function of this type of text. The results can be used to enhance second-language writing for marketing dairy products in English, thus supporting international professionals in multilingual contexts.

References

Bhatia, V. K. & Bremner, S. (2012). English for business communication. *Language Teaching*, *45*(4), 410-445.

Hunston, S. & Sinclair, J. (2000). A local grammar of evaluation. In S. Hunston & G. Thompson (Eds), *Evaluation in Text: Authorial Stance and the Construction of Discourse* (pp. 74-101). OUP.

Izquierdo, M. & Pérez Blanco, M. (2020). A multi-level contrastive analysis of promotional strategies in specialised discourse. *English for Specific Purposes*, *58*, 43-57.

Rayson, P., Archer, D., Piao, S. & McEnery, T. (2004). The UCREL Semantic Analysis System. In *Proceedings of the Workshop Beyond Named Entity Recognition, Semantic Labelling NLP Tasks (LREC 2004)* (pp. 7-12). European Language Resources Association.

Sing, C. S. (2017). English as a lingua franca in international business contexts: Pedagogical implications for the teaching of English for Specific Business Purposes. In F. Rainer & G. Mautner (Eds.), *Business communication: Linguistic approaches* (pp. 319-356). De Gruyter.

Evolución de los corpus en traducción: ¿Vamos hacia la popularización de su uso gracias a la IA?

Patricia Rodríguez-Inés

Universitat Autònoma de Barcelona

El objetivo principal de esta comunicación es explorar las posibilidades de la Inteligencia Artificial Generativa en la explotación de corpus, sobre todo para la traducción, y comparar su ejecución con la de herramientas de análisis de corpus "clásicas" (p. ej. SketchEngine o AntConc).

En las décadas de antes y después del 2000 (Rodríguez-Inés 2008), el análisis de corpus se realizaba con herramientas independientes y especializadas, como WordSmith Tools (Scott 1996), diseñadas específicamente para el análisis lingüístico detallado, o se interrogaban los pocos corpus online existentes a través de la web. Sin embargo, el panorama del trabajo con corpus ha evolucionado considerablemente. Desde hace tiempo algunas funciones de análisis de corpus se integran en otros tipos de software, como las herramientas de traducción asistida por ordenador (TAO) o gestores terminológicos, lo que amplía el acceso a los corpus y su utilidad en contextos profesionales. Además, la aparición de nuevas herramientas en línea para el análisis de corpus (p. ej. Voyant Tools) también favorece su uso sin que sea necesario un aprendizaje profundo de un software en particular.

Con la popularización de la inteligencia artificial generativa (IAG), ahora surgen nuevas preguntas sobre su potencial en el uso de corpus. ¿Podrían aplicaciones como ChatGPT (OpenAI 2023) llegar a rivalizar con herramientas especializadas como SketchEngine (Kilgarriff et al. 2004) o AntConc (Anthony 2012) para el análisis de corpus? Además, ¿podrían las tecnologías de IAG ayudar a popularizar los análisis basados en corpus en los Estudios de Traducción haciendo que estos métodos sean más fáciles de usar y se integren en flujos de trabajo de investigación más amplios? Esta presentación

explorará estas cuestiones a través de la realización de pruebas comparativas, examinando las capacidades y limitaciones de diferentes herramientas y tecnologías en el contexto de los estudios de traducción. Se contrastarán los resultados obtenidos con SketchEngine/AntConc y una aplicación de IAG en cuanto a recuentos de casos de una palabra o expresión en un corpus, obtención de listas de palabras por orden alfabético y de frecuencia, incluida la búsqueda de hapax legomena, así como extracciones de concordancias, colocaciones y algunos cálculos estadísticos, complementadas con interpretaciones de los resultados.

Las conclusiones pretenden aportar ideas sobre una futura dirección del uso de corpus y el papel potencial de la IAG en este ámbito.

References

Anthony, L. (2012). *AntConc* (3.3.2) [Software de ordenador]. Tokio, Japón: Universidad de Waseda. Disponible en http://www.antlab.sci.waseda.ac.jp/

Kilgarriff, A., Rychlý, P., Smrž, P. and Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*: 105-116.

OpenAI. (2023). *ChatGPT* (versión del 15 de julio) [Modelo de lenguaje de gran tamaño]. https://chat.openai.com/chat

Rodríguez-Inés, P. (2008). *Uso de corpus electrónicos en la formación de traductores (inglés-español-inglés)*. Tesis doctoral - Universitat Autònoma de Barcelona. Departament de Traducció i d'Interpretació. https://ddd.uab.cat/record/129868

Scott, M. (1996). *WordSmith Tools*. Oxford: Oxford University Press.

"Let alone" and "por no decir": A contrastive study of a complementary alternation discourse construction in English and Spanish

Pilar Ron Vaz

Universidad de Huelva

Complementary alternation discourse constructions are constructions that involve the juxtaposition of two elements, X and Y, contrasted within an entailment scale, where Y represents a stronger element and is considered to be more (or less) likely to occur. Instances of these constructions include *X if not Y; X let alone Y; X never mind Y*, among others (Iza Erviti, 2015). Within these, the *X let alone Y* construction is the one that has garnered more attention. Previous studies (Fillmore et al., 1988; Sawada, 2003; Capelle et al., 2015) indicate that explaining this construction is complex. Firstly, it is not limited to negative environments, as illustrated in (1):

(1) *The two Manchester clubs will no doubt compete for the league, let alone top four, but they should be wary of the Stamford Bridge side in London SW6.*

Secondly, *X let alone Y* can present Y as an additional element to be interpreted (and highlighted), as in (2), or contrast two elements with one more or less likely to occur than the other, as in (3):

(2) *The books that we choose to keep -let alone read- can say a lot about who we are and how we see ourselves.*

(3) *I suspect that W will be seen by historians as near Lincoln in stature and foresight. His policies were ones that we had to take, and your furthering the leftist myths shows a stark inability on your part to seek, let alone accept, truth.*

This paper aims to contribute to advance the research on the *X let alone Y* construction by (a) conduction a corpus analysis, and (b) by contrasting it to the Spanish construction *X por no decir Y*, which exhibits a similar usage distinction. This construction may present the Y element as an additional element to be interpreted (and highlighted), as in (4), or the two elements are presented in contrast with one being considered more or less likely to occur than the other, as in (5):

(4) *Por cierto, los hornos convencionales son bastante más caros que los microondas y gastan muchísima más energía, por no decir que tardan mucho más en calentar algo.*
(5) *El repudio que siente Tywin hacia su nieto es cada vez más evidente (por no decir que es del todo evidente).*

This study analyzes the blog and web sections of the Corpus of Contemporary American English (Davies, 2008) and compares it to the European Spanish data form the Corpus del Español (Davies, 2016). The study focuses on two primary objectives: (a) characterizing the relationships between the X and Y elements; and (b) the analyzing the discursive functions of these constructions. The findings reveal that these two factors are interrelated and that the nature of the relationship and whether an actual entailment scale is presented affects the interpretation and use of the construction. Moreover, the study demonstrates that, despite similarities, the English and Spanish constructions have distinct uses that render them not fully equivalent.

References

Cappelle, B., Dugas, E. & Tobin V. (2015). An afterthought on let alone. *Journal of Pragmatics*, *80*, 70-85.

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA): One billion words, 1990-2019*. https://www.english-corpora.org/coca/

Davies, M. (2016-). *Corpus del Español: Web/Dialects*. http://www.corpusdelespanol.org/web-dial

Fillmore, C. J., Kay, P. & O'Connor, C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, *64*, 501–538.

Iza Erviti, A. (2015). Complementary alternation discourse constructions in English: A preliminary study. *International Journal of English Studies*, *15*, 71-96. https://doi.org/10.6018/ijes/2015/1/194941

Sawada, O. (2003). Rethinking the let alone construction: what are its construction specific characteristics? *Journal of the Pan-Pacific Association of Applied Linguistics*, *7*, 135-151.

Identificación de equivalencias culturales mediante anotación semántica de estructuras argumentales

Sara Rupérez-León, Beatriz Sánchez-Cárdenas

Universidad de Valladolid, Universidad de Granada

Es bien sabido que las diferencias culturales emergen en el proceso traductológico, ya que la lengua refleja la manera en que un grupo lingüístico percibe e interpreta la realidad

(L'Homme 2020). Basándonos en los postulados teóricos de la Terminología basada en Marcos (Faber 2012, 2015, 2022) y de la Terminología Cultural (Diki-Kidiri 2000, 2022), sostenemos que la cultura influye en la estructura conceptual y lingüística no sólo de la lengua general, sino también de los ámbitos de especialidad. Los términos y sus estructuras cognitivas también están marcados culturalmente.

Esta contribución explora cómo determinar la equivalencia interlingüística (francés-español) en el ámbito de la silvicultura sostenible, aún por definir de manera exhaustiva en los recursos terminológicos dedicados a las Ciencias Ambientales, como EcoLexicon, GEMET o DicoEnviro, así como en recursos terminológicos generales como IATE o Termium Plus.

Nos centramos en el evento del ciclo forestal sostenible, que se define como el conjunto de procesos que se llevan a cabo en los bosques en los que intervienen agentes para su protección o para la obtención sostenible de recursos forestales. Engloba, por lo tanto, agentes y procesos que están marcados culturalmente.

Analizamos el corpus comparable FORESCOR (3 802 115 tokens), compilado a partir de 394 textos jurídicos y técnicos del ámbito forestal. Realizamos, en primer lugar, un vaciado terminológico en Sketch Engine (Kilgarriff et al., 2014) y seleccionamos manualmente los términos simples y compuestos referentes a las entidades y los procesos del evento analizado.

En segundo lugar, identificamos estructuras léxico-gramaticales partiendo de las concordancias y los esquemas distribucionales en los que intervienen los términos. Para ello, llevamos a cabo diferentes estrategias, como el estudio de las colocaciones verbales en forma de triples en las que participan los términos (sustantivo | verbo | sustantivo; *technicien* | *effectuer* | *martelage*) (Sánchez-Cárdenas y Ramisch 2019). Observamos que los sustantivos deverbales, al heredar la capacidad argumental de los verbos, pueden establecer relaciones argumentales sin necesidad de un verbo, de manera que se vinculan a través de otros elementos gramaticales, como preposiciones o conjunciones. A partir de este tipo de concordancias, en tercer lugar, realizamos la anotación semántica manual de las estructuras argumentales, a partir de la cual se generalizan estructuras recurrentes. Utilizamos para ello categorías conceptuales, como entidad y proceso (Gil-Berrozpe y Cabezas-García, 2023), y roles semánticos, como Agente o Resultado (Rojas-García, 2022), los cuales etiquetamos semánticamente con la herramienta catma (Meister, 2023). A partir de los patrones emerge la estructura de los marcos semánticos en cada lengua. Su comparación permite identificar equivalencias interlingüísticas (p. ej., *abatteuse* [fr] – *procesadora* [es]) y, en última instancia, aquellos casos de anisomorfismo cultural (p. ej., *cloisonnement* [fr] – *rodal* [es]). Este enfoque pone de relieve la importancia de la cultura en la construcción de marcos semánticos y en la identificación de equivalencias interlingüísticas.

Referencias

Biber, D. (2012). Corpus-based and corpus-driven analyses of language variation and use. En B. Heine, y H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (pp. 193-197). Oxford: Oxford University Press.

Diki-Kidiri, M. (2000). Un enfoque cultural de la terminología. *Terminologies nouvelles*, *21*, 27-31.

Diki-Kidiri, M. (2022). Cultural Terminology: An introduction to theory and method. En P. Faber y M. C. L'Homme (Eds.), *Theoretical perspectives on Terminology. Explaining terms, concepts and specialized knowledge* (pp. 197-216). John Benjamins Publishing Company.

Faber, P., y Cabezas-García, M. (2019). Specialized Knowledge Representation: From Terms to Frames. *Research in Language*, *17*(2), 197–211. https://doi.org/10.2478/rela-2019-0012

Faber, P. (2009). The cognitive shift in terminology and specialized translation. *MonTI. Monografías de Traducción e Interpretación*, *1*, 107-134.

Faber, P. (2010). Terminología, traducción especializada y adquisición de conocimiento. En Alarcón Navío (Ed.), *La traducción en contextos especializados: propuestas didácticas*. Granada: Atrio.

Faber, P. (2012). *A cognitive linguistics view of terminology and specialized language*. Berlín: De Gruyter.

Faber, P. (2015). Frames as a framework for terminology. En H. J. Kockaert y F. Steurs (Eds.), *Handbook of terminology* (Vol. 1, pp. 14–33). John Benjamins Publishing Company.

Faber, P. (2022). Frame-based terminology. En P. Faber y M. C. L'Homme (Eds.), *Theoretical perspectives on terminology: Explaining terms, concepts and specialized knowledge* (pp. 353–376). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.23.int

Faber, P., y L'Homme, M. C. (2022). *Theoretical perspectives on Terminology. Explaining terms, concepts and specialized knowledge* (pp. 1-12). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.23.int

Gil-Berrozpe, J. C. y Cabezas-García, M. (2023). A terminological template for describing and representing hyponymic information. En C. Vargas Sierra y J. A. Sánchez Fajardo (Eds.), *Temas actuales de traducción especializada, docencia, transcreación y terminología* (pp. 555-589). Tirant lo Blanch.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovár, V., Michelfeit, J., Rychlý, P., y Suchomel, V. (2014). *The Sketch Engine: Ten years on. Lexicography*, *1*(1), 7-36.

L'Homme, M. C. (2020). *Lexical Semantics for Terminology: An Introduction*. John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.20

León-Araúz, P., Reimerink, A., Cabezas-García, M., y Faber, P. (2024). Ideological Knowledge Representation: Framing Climate Change in EcoLexicon. En *LREC-COLING 2024*, 8617-8626.

Meister, J. C. (2023). From TACT to CATMA or A mindful approach to text annotation and analysis. En J. Nyhan, G. Rockwell, S. Sinclair, y A. Ortolja-Baird (Eds.), *On Making in the Digital Humanities: Essays on the Scholarship of Digital Humanities Development in Honour of John Bradley*. UCL Press. https://www.uclpress.co.uk/products/211148

Rojas-García, J. (2022). Semantic Representation of Context for Description of Named Rivers in a Terminological Knowledge Base. *Frontiers in Psychology*, *13*. doi: 10.3389/fpsyg.2022.847024

Sánchez-Cárdenas, B. (2024). Extracting semantic frames from specialized corpora for lexicographic purposes. *Círculo de Lingüística Aplicado a la Comunicación*, *99*, 163-177. https://dx.doi.org/10.5209/clac.90626

Sánchez-Cárdenas, B., y Ramisch, C. (2019). Eliciting specialized frames from corpora using argument-structure extraction techniques. Terminology. *International Journal of Theoretical and Applied Issues in Specialized Communicatio*n, *25*(1), 1-31. https://doi.org/10.1075/term.00026.san

Panel 6

Linguistic variation and change through corpora
Variación y cambio lingüístico basado en corpus

Economy and informality: On the complexity of no(t)-fragments

Laura Abalo Dieste

Universidade de Vigo

This paper investigates the use of *no(t)*-fragments in contemporary British English, focusing on the tension between the communicative pressures of efficiency and expressiveness. Drawing on Fernández-Pena (2021, p. 136), fragments are defined as "stand-alone constructions which, despite their reduced, non-canonical, fragmentary structure, are still semantically, discursively and pragmatically equivalent to a complete clause construction". As formally defective yet propositionally felicitous constructions, fragments illustrate both syntactic reduction and pragmatic richness. Building on Goldberg (2019)'s account of communicative pressures on the productivity of constructions, this study situates *no(t)*-fragments as a lens through which to explore two competing forces: the drive for syntactic economy or densification (Biber & Gray, 2012, 2015) and the sociolinguistic trend of increasing colloquialisation (Mair, 1997).

Adopting the framework of Construction Grammar, fragments emerge as partial realisations of full constructions, cognitively triggering their complete version, until the reduced construction become entrenched and independently stored in the long-term memory (Bauer & Hoffmann, 2020). Consequently, phrasal, short and conventionalized complements are more likely to occur in sufficiently frequent fragments than clausal, longer and more expressive complements (Brinton, 2014; Hugou, 2017). Such tendencies align with the principle of linguistic economy, particularly favoured in registers where fragments have been attested, such as headlines (Quirk et al., 1985, p. 845) and note-taking (Janda, 1985).

Despite the economical features, the informal distribution and expressive versatility of *no(t)*-fragments may render them particularly receptive to colloquialisation trends, thus accommodating both syntactically simpler and clausal complementation patterns. Colloquialisation, characterised by a shift towards a more speech-like style, has been traced to the seventeenth century and is increasingly evident in contemporary formal written English (Rodríguez-Puente & Obaya-Cueli, 2022). Thus, the complexity of, specifically, *no*-constructions (Cappelle & Depraetere, 2016), as in (1), and *not*-fragments (Cappelle, 2021, p. 69), exemplified in (2), has yet to be systematically explored, particularly regarding their susceptibility to the competing influences of colloquialisation and linguistic economy.

(1) I'm not quite sure what for but cos I burped / *no need to be excused for that* it's no problem / thank you anything bigger erm expensive it's about ninety (BNC2014: Sp1m1f121)

(2) to obtain central funding for this small group of children. / *Not a comfortable option* / Many will see this paper as a fundamental attack on (BNC1994: AcaMed16)

By querying the two releases of the *British National Corpus* —the BNC1994 (BNC Consortium, 2007) and the BNC2014 (Brezina et al., 2021; Love et al., 2017)—, this paper analyses the length and syntactic complementation of *no(t)*-fragments across written and spoken registers via the software #LancsBox X (v. 5.0.3) (Brezina & Platt, 2024). Specifically, this paper addresses three research questions: (i) Do syntactically less complex *no(t)*-fragments increase in frequency from 1994 to 2014? (ii) Do *no(t)*-fragments favour phrasal over clausal

complementation? (iii) Which specific non-canonical syntactic patterns are characteristic of *no(t)*-fragments? By tracing the diachronic development and register-based distribution of *no(t)*-fragments, the findings reveal insights into the interplay of communicative pressures — efficiency and expressiveness— and sociolinguistic trends in the use of reduced constructions in contemporary British English.

References

Bauer, E.-M., & Hoffmann, T. (2020). *Turns out* is not ellipsis? A usage-based construction grammar view on reduced constructions. *Acta Linguistica Hafniensia*, *52*(2), 240–259. https://doi.org/10.1080/03740463.2020.1777036

Biber, D., & Gray, B. (2012). The competing demands of popularization vs. economy: Written language in the age of mass literacy. In T. Nevalainen & E. C. Traugott (Eds.), *The Oxford Handbook of the History of English* (pp. 314–328). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199922765.013.0028

Biber, D., & Gray, B. (2015). *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511920776

BNC Consortium. (2007). *British National Corpus: XML edition*. Oxford Text Archive. http://www.natcorp.ox.ac.uk/cpr.xml?ID=reference

Brezina, V., & Platt, W. (2024). *#LancsBox X* [Software v.5.0.3]. Lancaster University. https://lancsbox.lancs.ac.uk/

Brinton, L. J. (2014). The Extremes of Insubordination: Exclamatory *as if! Journal of English Linguistics*, *42*(2), 93–113. https://doi.org/10.1177/0075424214521425

Cappelle, B. (2021). *Not*-fragments and negative expansion. *Constructions and Frames*, *13*(1), 55–81. https://doi.org/10.1075/cf.00047.cap

Cappelle, B., & Depraetere, I. (2016). Modal meaning in Construction Grammar. *Constructions and Frames*, *8*(1), 1–6. https://doi.org/10.1075/cf.8.1.01cap

Fernández-Pena, Y. (2021). Towards an empirical characterisation and a corpus-driven taxonomy of fragments in written contemporary English. *Revista Electrónica de Lingüística Aplicada*, *20*(1), 136–154.

Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Hugou, V. (2017). The WHX construction (what the hell…?) and intensity. A corpus-based study. *Lexis. Journal in English Lexicology*, *10*, 1–30. https://doi.org/10.4000/lexis.1103

Janda, R. D. (1985). Note-taking English as a simplified register. *Discourse Processes*, *8*(4), 437–454. https://doi.org/10.1080/01638538509544626

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, *22*(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Mair, C. (1997). The spread of the going-to-future in written English: A corpus-based investigation into language change in progress. In R. Hickey & S. Puppel (Eds.), *Language History and Linguistic Modelling. A Festschrift for Jacek Fisiak on His 60th Birthday* (pp. 1537–1543). De Gruyter.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive grammar of the English language*. Longman.

Rodríguez-Puente, P., & Obaya-Cueli, M. (2022). Phrasal verbs in Early Modern English spoken language: A colloquialization conspiracy? *English Language and Linguistics*, 1–25. https://doi.org/10.1017/S1360674322000065

From saying it literally to literally breathtaking: Some diachronic notes on the adverb *literally*

Zeltia Blanco Suárez

Universidade de Santiago de Compostela

Although *-ly* adverbs have been the focus of much scholarly debate both from a diachronic and a synchronic perspective (cf., among others, Peters, 1994; González-Álvarez, 1996; Tagliamonte & Ito, 2002; Nevalainen, 2008; Lewis, 2020; Aijmer, 2023), some other forms have received considerably less attention. This is in fact the case of *literally*, despite its increasing frequency in recent decades (cf., however, Claridge, 2011; Aijmer, 2023).

Literally is classified variously in the reference grammars: as an emphasiser and a style disjunct of respect (Quirk et al., 1985: 583, 615-616, 619), as a metalinguistic adjunct (Huddleston & Pullum, 2002: 775), and as a stance adverbial (Biber et al., 2021: 758). Other scholars actually refer to the intensifying or degree function of *literally*, including Bolinger (1972), Goatly (1997), Israel (2002), Claridge (2011), Calhoun (2015), and Aijmer (2023). To date, however, a comprehensive diachronic corpus-based study of the history of *literally* has not yet been undertaken. The present paper, therefore, seeks to address this gap by tracing its diachrony up to present-day English (PDE).

According to the OED, *literally* is first recorded in the 15th century as a manner adverb, with the meaning 'in a literal, exact, or actual sense', as shown in example (1):

(1) *c*1429 *Mirour Mans Saluacioune* (1986) l. 553 *Litteraly haf ȝe herde this dreme and what it ment*. (*OED* s.v. *literally*, adv. I.1.a.)

The first intensifying uses of *literally*, meaning 'virtually, as good as' and 'completely, utterly, absolutely', date to the 18th century, as in example (2).

(2) 1769 F. Brooke *Hist. Emily Montague* IV. ccxvii. 83 *He is a fortunate man to be introduced to such a party of fine women at his arrival; it is literally to feed among the lilies.* (*OED* s.v. *literally*, adv. I.1.c.)

Our preliminary data seem to suggest that, similarly to other intensifiers, *literally* has developed along a grammaticalisation cline (Traugott, 1982, 2010), gradually allowing a wider range of collocates and syntactic contexts, hence its additional use as a parenthetical in both the left and right periphery, as in examples (3)-(4), respectively:

(3) *Literally, he groped along, feeling the fronts of the houses.* (CLMET3.0)

(4) *They ought not to be doing, literally*. (BNC1994)

In order to account for its diachronic development, we have considered the different syntactic and pragmatic functions of *literally*, along with its different collocates and positions at phrasal and clausal level. To this end, evidence has been drawn from the *Early English Books Online Corpus* (EEBOCorp 1.0), the *Corpus of Late Modern English Texts, version 3.0* (CLMET3.0), and the *British National Corpus 1994 and 2014* (BNC1994 and BNC2014).

Corpora

BNC1994. *The British National Corpus*. www.natcorp.ox.ac.uk.
BNC2014. *The British National Corpus 2014*. http://corpora.lancs.ac.uk/bnc2014/

EEBOCorp 1.0. *Early English Books Online Corpus 1.0*, compiled by P. Petré. (2013). Available at https://lirias.kuleuven.be/handle/123456789/416330/.

CLMET3.0. De Smet, H., Diller, H.J., & Tyrkkö, J. (2013). *The Corpus of Late Modern English Texts*, version 3.0. Leuven: K.U. Leuven. https://perswww.kuleuven.be/~u0044428/clmet3_0.htm

References

Aijmer, K. 2023. Looking at grammaticalization from the perspective of short-time changes in real time: A comparative corpus-based study of *literally*. In De Smet, H., Petré, P., & Szmrecsanyi, B. (Eds.), *Context, intent and variation in grammaticalization* (pp.19-46). Berlin: Mouton de Gruyter.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2021). *Grammar of spoken and written English*. Amsterdam & Philadelphia: John Benjamins.

Bolinger, D. (1972). *Degree words*. The Hague: Mouton.

Calhoun, K. (2015). "It is the worst of our time": Youth language, language attitudes, and arguments about *literally*. *Texas Linguistics Forum*, *58*, 1-10.

Claridge, C. (2011). *Hyperbole in English: A corpus-based study of exaggeration*. Cambridge: Cambridge University Press.

Goatly, A. (1997). *The language of metaphors*. London & New York: Routledge.

González-Álvarez, D. (1996). Epistemic disjuncts in Early Modern English. *International Journal of Corpus Linguistics*, *1*(2), 219-256.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Israel, M. (2002). *Literally* speaking. *Journal of Pragmatics*, *34*, 423-432.

Lewis, D. (2020). Speaker stance and evaluative *-ly* adverbs in the Modern English period. *Language Sciences*, *82*, 1-13.

Nevalainen, T. (2008). Social variation in intensifier use: Constraint on *–ly* adverbialization in the past? *English Language and Linguistics*, *12*(2), 289-315.

OED (*Oxford English Dictionary*). 3rd edn. in progress: *OED Online*, March 2000-, ed. Michael Profitt. www.oed.com

Peters, H. (1994). Degree adverbs in Early Modern English. In Kastovsky, D. (Ed.), *Studies in Early Modern English* (pp. 269-288). Berlin: Mouton de Gruyter.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A comprehensive grammar of the English language*. London & New York: Longman.

Tagliamonte, S., & Ito, R. (2002). Think *really* different: Continuity and specialization in the English dual form adverbs. *Journal of Sociolinguistics*, *6*(2), 236-266.

Traugott, E. C. (1982). From propositional to textual and expressive meanings: Some semantic-pragmatic aspects of grammaticalization. In Lehmann, W. P., & Malkiel, Y. (Eds.), *Perspectives on historical linguistics* (pp. 245-271). Amsterdam: John Benjamins.

Traugott, E. C., & Dasher, R. (2002). *Regularity in semantic change*. Cambridge: Cambridge University Press.

'You spelled "spelt" wrong': On the configuration of the regular and the irregular forms of the past in English

Javier Calle Martín, Marta Pacheco-Franco

Universidad de Málaga

The expression of the past tense in English has been plagued with oddities throughout its history. One clear example is the morphological duality that exists in verbs like *dream*, *burn* and *spell*, among others. Their preterite and participial forms may feature two morphemes, either the regular *-ed* or the irregular *-t*. For most of these verbs, the regular and irregular variants already co-existed in the sixteenth century, and they seem to have survived down to Present-day English unaffected by the overall standardisation of morphology and spelling of the Early Modern period. In Present-day English, some have tried to explain this duplicity in functional terms, arguing that *-ed* and *-t* represent differences in phonology and in grammar. However, it seems that preference for one form over the other is determined on the grounds of diatopic variation, with American English favouring the regularised variants *dreamed*, *burned* and *spelled*, while *dreamt*, *burnt* and *spelt* are dominant in British and Commonwealth English. Regardless of how the variants are used, the co-existence of these competing forms is interesting in its own right and, as such, has received some scholarly attention (Anderwald, 2014; Balle-Mascaró & Suárez-Gómez, 2015; Levin, 2009). Nevertheless, there are gaps in the literature that would shed some much-needed light on the phenomenon from both diachronic and synchronic perspectives.

For that purpose, the study has been conceived under a dual scope. First, the spelling variants are considered in terms of diatopic variation. The analysis centres on the historical development of British and American English in the periods 1600-1999 and 1750-1999, respectively, and then moves on to usage in Present-day English, where other varieties of World Englishes are also considered. Then, the study adopts a functional approach for the purpose of illustrating the potential specialisation of the spelling variants on the grounds of part of speech. As a corpus-based study, the data were drawn from *A Representative Corpus of Historical English Registers* (*ARCHER*) for the diachronic component of the study and from the *Corpus of News on the Web* (*NOW*) for insight into Present-day English. The results suggest that there was an overall shift towards the regularisation of the morphological paradigm throughout the history of British and American English, though the trend changed in twentieth-century Britain and the *-t* forms gained ground once again. In any event, most varieties of World Englishes seem to be leaning towards the regularisation of the paradigm, including British English, where the expansion of the irregular forms seems to have been reverted once more. It is worth noting that the data do not point towards the possible specialisation of any of the variants in terms of grammatical function.

References

A Representative Corpus of Historical English Registers (*ARCHER*) 3.2. 1990-2013. Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. Current member universities are Northern Arizona, Southern California, Freiburg, Heidelberg, Helsinki, Uppsala, Michigan, Manchester, Lancaster, Bamberg, Zurich, Trier, Santiago de Compostela and Leicester. https:// www.projects.alc.manchester.ac.uk/archer/.

Anderwald, L. (2014). *Burned*, *dwelled*, *dreamed*: The Evolution of a Morphological Americanism and the Role of Prescriptive Grammar Writing. *American Speech*, *89*(4), 408-440.

Balle-Mascaró, B. & Suárez-Gómez, C. (2015). Morphological Variation of Verbs in Native Varieties of English. In C. Suárez-Gómez & E. Seoane (Eds.), *Englishes Today: Multiple Varieties, Multiple Perspectives* (pp. 33-49). Cambridge Scholars Publishing.

Davies, M. (2016-). *Corpus of News on the Web (NOW)*. Available online at https://www.english-corpora.org/now/.

Levin, M. (2009). The Formation of the Preterite and the Past Participle. In G. Rohdenburg & J. Schlüter (Eds.), *One Language, Two Grammars? Differences between British and American English* (pp. 60-85). Cambridge University Press.

El análisis de la variación terminológica: Correlación de variables cuantitativas desde un corpus de la lengua del ferrocarril

Andrea Fernández Vivanco

Universidad de Salamanca

La variación terminológica o variación denominativa hace referencia a la convivencia de varias denominaciones para un mismo concepto en los textos especializados (Freixa, 2022). Esta variación se produce en un entorno poliédrico (Cabré, 2023) en el que se funden el contexto lingüístico y las condiciones de producción, de circulación y de recepción del discurso especializado. Las propuestas que pretenden clasificar las causas de la variación terminológica (Delavigne, 2017; Freixa, 2022) incluyen variables ligadas a la producción textual (causas dialectales, discursivas e interlingüísticas), variables relacionadas con la recepción textual (causas funcionales) y variables preliminares. En estas propuestas académicas surge, además, la dimensión cognitiva como causa de variación o bien como constante que impregna las demás causas (León Arnauz, 2017).

La utilización de corpus textuales ha contribuido al desarrollo teórico de la terminología como disciplina (Cabré, 1999; Temmermann, 2000) y a la elaboración de productos terminográficos. Los corpus se erigen como herramienta para acceder a datos lingüísticos auténticos que reflejan el uso real de los términos en contextos especializados. También nuestro estudio se sirve del corpus para la extracción de términos, el análisis de sus patrones de uso y la identificación de variantes terminológicas.

La presente comunicación pretende, en primer lugar, exponer brevemente la metodología de creación y etiquetado de *Ferrolex*, un corpus en español de textos de la industria ferroviaria. A continuación, exponemos una taxonomía de variables de variación terminológica que se construye sobre estudios realizados dentro de la propia disciplina de la terminología y se nutre de propuestas de disciplinas afines como la semántica (Földes, 2023) y la lexicología. Este corpus se ha utilizado para realizar una extracción terminológica, a partir de la cual se han identificado aquellos términos que presentan variación denominativa. Sobre este conjunto de datos nos planteamos la hipótesis de que existe una correlación entre la dimensión cognitiva de la variación y (a) las causas diastráticas, (b) las causas funcionales y (c) las causas sociolingüísticas. Para confirmar o refutar esta hipótesis de partida nos servimos de los datos cuantitativos obtenidos a partir del etiquetado de textos previo a la compilación del corpus.

Finalmente, los resultados de esta investigación se comparan con un estudio anterior que analizaba el mismo fenómeno a partir de entradas lexicográficas en recursos especializados del ferrocarril. La comparación del estudio de la terminología *in vivo* e *in vitro* pone de relieve el papel del corpus en la identificación de fenómenos lingüísticos relacionados con la coocurrencia, la semántica contextual y la evolución diacrónica de la terminología, entre otros.

References

Cabré, T. (2023). *Terminology: Cognition, Language and Communication*. Amsterdam: John Benjamins.

Delavigne, V. (2017). Term Usage and Socioterminological Variation. The Impact of Social and Local Issues on the Movement of Terms. In P. Drouin, A. Francoeur, J. Humbley, & A. Picton (Eds.), *Multiple Perspectives on Terminological Variation* (pp. 31-56). Amsterdam: John Benjamins.

Földes, C. (2023). Lexikalische Variation im Fokus: Ein Problemaufriss. *Studia Germanistica*, 5-22.

Freixa, J. (2022). Causes of terminological variation. In P. Faber, & M.-C. L'Homme (Eds.), *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge* (pp. 399-420). Amsterdam: John Benjamins.

León Arauz, P. (2017). Term and Concept Variation in Specialized Knowledge Dynamics. In P. F. Drouin, J. Humbley, & A. Picton (Eds.), *Multiple Perspectives on Terminological Variation* (pp. 213-258). Amsterdam: John Benjamins.

Pérez Hernández, M. C. (2002). Terminografía basada en corpus : aspectos fundamentales de la gestión terminológica. *Estudios de lingüística del español*.

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive Approach*. Amsterdam/Philadelphia: John Benjamins.

Winter, T. (2019). Terminologie im nicht-translatorischen Kontext. In P. Drewer, & D. Pulitano (Eds.), *Terminologie: Epochen – Schwerpunkte – Umsetzungen* (pp. 95-104). Berlin: Springer.

Exploring PRICE words in the dialect of nineteenth-century Lancashire through a corpus-based framework

Nadia Hamade Almeida

Universidad Camilo José Cela

The Lancashire dialect has been prominently featured in various literary works, including *The Late Lancashire Witches* (1634), *That Lass O' Lowrie's* (1877), and *Hard Times* (1854). Regional literature is commonly categorized into two distinct types: dialect literature and literary dialect. Dialect literature refers to works that are entirely composed in a non-standard variety of a language, which restricts their accessibility to readers who possess the ability to read and understand the dialect represented. Alternatively, literary-dialect texts are predominantly written in Standard English or a prestigious variety, except for the dialogue, which reflects a specific regional dialect. Nevertheless, literary-dialect authors were not dialect specialists, and as such, their representation of dialects lacked the precision and depth that would be expected from a more linguistically rigorous approach.

One of the most salient characteristics of this type of representation is the presence of semi-phonetic spellings, which relate to non-standard or deviant orthographical conventions based upon the Standard English orthography. For instance, the use of <ee> and <oi> to suggest the monophthong [iː] and the diphthong [ɔɪ], respectively.

Literary-dialect texts are considered valuable resources for linguists in dialect studies (Ruano-García 2007: 111; Beal 2011: 204). This is due to the potential of a detailed analysis of non-standard spellings to yield phonological insights into a regional variety at a specific point in time. As a result, this study utilizes such texts to explore the Lancashire vernacular. Since a comprehensive examination of the Lancashire dialect exceeds the scope of this paper, the focus is specifically placed on the sounds and spellings associated with the PRICE lexical set, according to Wells' (1982a: 155) classification of words containing the diphthong [aɪ]. This study will be

categorized into two different groups, depending on the presence or absence of the <gh> digraph in the standard orthography of PRICE words.

This paper outlines and explains the different dialectal pronunciations and the possible coexistence of sounds within the same lexical group, considering historical and sociolinguistic factors. To achieve this, a corpus consisting of nineteen literary-dialect works by five different authors was compiled and analyzed manually. Since the dialect is primarily represented in the characters' speech, the study focuses primarily on these dialogues. The various non-standard spellings related to the PRICE lexical set were treated as primary sources and linked to their corresponding phonetic representations in the Lancashire dialect. In this context, García-Bermejo Giner (1999: 252) argues that comparing standard and non-standard orthography is invaluable when conducting a phonological analysis through literary-dialect texts.

The findings reveal the presence of three distinct pronunciations within the PRICE lexical set. One of these pronunciations represents a traditional dialect form, which might exhibit a regressive tendency, while the other two correspond to modern variants, derived from the RP diphthong [aɪ]. According to Wells (1982), the presence of traditional sounds and RP variants within the same lexical set is explained by the process of relexification, which involves the replacement of older sounds with more modern ones.

References

García-Bermejo Giner, F. (1999). Methods for the linguistic analysis of early modern English literary dialects. In P. Alonso (Ed.), *Teaching and research in English and linguistics* (pp. 249-266). Celarayn.

Ruano-García, J. (2007). Thou'rt a strange fille: A possible source for 'y-tensing' in seventeenth-century Lancashire dialect? *Sederi*, *17*, 109-127.

Wells, J. C. (1982). *Accents of English*. Cambridge University Press.

Lemmatisation of Anglo-Saxon poetic adjectives

Yosra Hamdoun Bghiyel

Universidad de La Rioja

The lemmatisation of Old English poetic texts represents a fundamental challenge in the study of historical linguistics, lexicography, and corpus linguistics. Adjectives, in particular, pose unique difficulties due to their morphological complexity, variation in attested spellings, and their syntactic and semantic roles within Old English poetry. Despite significant advancements in digital resources and lexicographical tools, there remains a notable gap in the systematic treatment of adjectives within historical language corpora, particularly in capturing their linguistic variation and diachronic change. This study addresses this gap by exploring how a relational database framework can be used to assign unified headwords to adjectives in the York-Helsinki Parsed Corpus of Old English Poetry (YCOEP; Pintzuk and Plug, 2001). The relational database integrates several online reference dictionaries such as The Dictionary of Old English (DOE; Healey et al., 2018) and Bosworth-Toller's An Anglo-Saxon Dictionary (1973) with annotated databases and additional contrasting datasets. By correlating these sources, the headwords and attested spellings, the IOED provides a streamlined, comprehensive framework for systematically comparing and selecting the most suitable lemma for each adjective. The methodology involves four systematic steps to account for linguistic variation in the lemmatisation of adjectives in the YCOEP. First, the YCOEP inventory of attested spellings for adjectives is compiled into a structured database for classification. Second, these spellings are matched with sources from the relational database to assign headwords automatically. Third,

the assigned headwords are aligned with our database containing the final lemma list for consistency. Finally, the resulting lemmatised inventory is validated against main completed or semi-completed sources of reference: the DOE, VariOE, and Bessinger's A Short Dictionary of Anglo-Saxon Poetry (1960), which includes over 5,000 entries, to identify and assess any discrepancies in headword selection. The main conclusions of this research highlight the feasibility of a corpus-based approach to fully lemmatising Old English adjectives while addressing their linguistic variation across texts and periods. Preliminary findings include the creation of 15 new lemmas for adjectives and a listing of lemmatised poetic Old English vocabulary exclusive to this corpus. These contributions provide a replicable model for capturing variation and diachronic change in historical language corpora.

References

Bessinger, J. B. (1960). *A short dictionary of Anglo-Saxon poetry*. University of Toronto Press.

Bosworth, J., & Toller, T. N. (1973). *An Anglo-Saxon dictionary*. Oxford University Press.

Pintzuk, S., & Plug, L. (Comps.). (2001). *The York-Helsinki Parsed Corpus of Old English Poetry*. Retrieved from http://www-users.york.ac.uk/~lang18/pcorpus.html

Healey, A. diPaolo, Wilkin, J. P., & Xiang, X. (2018). *Dictionary of Old English Web Corpus. Dictionary of Old English Project*, Centre for Medieval Studies, University of Toronto.

De la patologización a la despatologización: transformaciones semánticas en la terminología clínica sobre diversidad sexual y de género

Emma Machado de Souza

Universidad de Salamanca

La evolución de los términos relacionados con la diversidad sexual y de género en el ámbito clínico refleja importantes transformaciones sociales e ideológicas. Conceptos como "inversión sexual", "homosexualidad egodistónica" o "transexualismo", presentes en manuales y documentos institucionales desde mediados del siglo XX, han sido gradualmente sustituidos o resignificados, lo que evidencia un tránsito desde un discurso patologizante hacia uno más inclusivo.

Este estudio se centra en analizar estas transformaciones semánticas a través de un corpus representativo de textos clínicos y académicos publicados en España entre 1939 y 2023, abordando cómo los términos clave han cambiado en su uso, asociaciones y contexto discursivo. La investigación busca responder a dos preguntas fundamentales: ¿cómo ha evolucionado el significado de estos términos a lo largo del tiempo en textos clínicos españoles? Y ¿qué relaciones semánticas reflejan las transiciones ideológicas subyacentes en cada periodo histórico?

Para responder a estas cuestiones, se ha construido un corpus compuesto por manuales universitarios de referencia en psiquiatría y psicología, informes de asociaciones profesionales y documentos normativos relevantes. El análisis se realiza en cuatro periodos históricos: franquismo (1939 – 1975), transición democrática (1975 – 1985), democracia prematrimonio igualitario (1986 – 2004) y democracia postmatrimonio igualitario (2005 – 2023). Se hace uso de herramientas como *AntConc*, *Voyant Tools* y *Gephi* para analizar la frecuencia, las coocurrencias léxicas y las redes semánticas asociadas a los términos clave. La metodología combina enfoques cuantitativos y cualitativos. Por un lado, se identifican patrones de frecuencia y coocurrencia para observar cómo los términos se agrupan en distintos contextos. Por otro lado, se complementa mediante el uso del análisis crítico del discurso para interpretar cómo los

cambios léxicos reflejan las dinámicas culturales e ideológicas de cada periodo (Cameron y Kulick, 2003; Motschenbacher, 2020).

Los resultados esperables indican que los términos patologizantes, como "inversión sexual" o "transexualismo", aparecen frecuentemente vinculados a nociones de enfermedad, desviación y tratamiento en los periodos iniciales. En contraste, términos como "diversidad sexual y de género" emergen en contextos recientes asociados a derechos, reconocimiento y políticas inclusivas (Sánchez, 2020; Gasch-Gallen et al., 2021). Las redes semánticas generadas muestran transiciones progresivas en las asociaciones léxicas, entre las que destacan hitos específicos como la eliminación de la homosexualidad del DSM-II (Drescher, 2015) y la creciente influencia de directrices internacionales, como las establecidas en la CIE-11 de la OMS. Estas transformaciones evidencias cambios terminológicos, así como transformaciones en los marcos discursivos que sustentan las prácticas clínicas y académicas.

El análisis de estas dinámicas aporta una contribución al estudio del significado en corpus especializados, lo que demuestra cómo el lenguaje refleja y reconfigura las concepciones sociales sobre la diversidad sexual y de género. Además, evidencia el papel de la lingüística de corpus como herramienta para rastrear la evolución de los discursos institucionales en relación con los cambios culturales e ideológicos (Rivas, 2020). Este enfoque interdisciplinar combina lingüística, semántica y análisis crítico con el fin de ofrecer una perspectiva integral sobre el impacto de las transformaciones terminológicas en el ámbito clínico español.

Referencias

Cameron, D., y Kulick, D. (2003). *Language and sexuality*. Cambridge University Press

Drescher, J. (2015) Out of DSM: Depathologizing Homosexuality. *Behavioral Sciences*, *5*(4), 565-75. doi: 10.3390/bs5040565.

Gasch-Gallén, A., Gregori-Flor, N., Hurtado-García, I., Suess-Schwend, A., y Ruiz-Cantero, M. (2021). Diversidad afectivo-sexual, corporal y de género más allá del binarismo en la formación en ciencias de la salud. *Gaceta Sanitaria*, *35*(4), 383-388. https://dx.doi.org/10.1016/j.gaceta.2019.12.003.

Motschenbacher, H. (2020). *Queer linguistics: A critical introduction.* Edinburgh University Press.

Rivas, A. (2022). Diversidad sexual y de género y discriminación. Tratamiento internacional. *Revista Latinoamericana de Derecho Social*, *1*(34e). https://doi.org/10.22201/iij.24487899e.2022.34e.16820 .

Sánchez, T. (2020). Sexo y género: una mirada interdisciplinar desde la psicología y la clínica. *Revista de la Asociación Española de Neuropsiquiatría*, *40*(138), 87-114.https://dx.doi.org/10.4321/s0211-573520200020006.

Zimman, L. (2019). Transgender language reform. *Journal of Language and Discrimination*, *3*(1), 84-105. https://doi.org/10.1558/jld.33139

La ratio de categoría gramatical en español. Rasgos estilísticos y sociolingüísticos

Inmaculada Martínez Martínez

Universidad de Cantabria

La Ratio de Categoría Gramatical (RCG) se estudia desde diversas perspectivas: el estilo de una obra literaria o autor, los cambios históricos o el género de la obra, entre otras. En cambio, en el campo de la sociolingüística, no hemos encontrado apenas bibliografía sobre la relación entre la RCG y factores extralingüísticos como el sexo, la edad o los antecedentes educativos, probablemente debido al hecho de que la RCG no muestre mucha variación sociolingüística. Si este es el caso, la constancia sociolingüística de la RCG indicará la esencia de la lengua.

En esta comunicación, tras una visión general de las investigaciones anteriores sobre la RCG, revisaremos el Diccionario de frecuencias de vocabulario español (Juilland Chang-Rodríguez: 1964) con el objetivo de observar variaciones de RCG encontradas en diversas tipologías (obras teatrales, novelas, ensayos, prensa, textos científicos). Confirmamos que la RCG de palabras de contenido varía considerablemente entre dichas tipologías. En concreto, en obras teatrales hay relativamente más verbos y menos sustantivos y adjetivos con diferencias destacadas. En ensayos, prensa y textos científicos, la proporción se invierte mientras las novelas exhiben un estado intermedio entre ambos.

Seguidamente, presentaremos los aspectos característicos del habla observados en las entrevistas realizadas en Santander con 54 hablantes clasificados por sexo, edad y nivel educativo y correspondientes al proyecto PRESEEA (Proyecto para el Estudio Sociolingüístico del Español de España y América). Cabe señalar aquí una menor variabilidad que la encontrada en el Diccionario de frecuencias. La RCG constante observada en las entrevistas no puede ser resultado del azar, puesto que se trata de los hábitos lingüísticos que cada hablante ha adquirido de forma independiente en diferentes entornos lingüísticos. Examinaremos las causas de la constancia de la RCG en el lenguaje hablado de acuerdo con el "Sistema" y "Norma" del lenguaje (Coseriu: 1973) y la "Ley de los grandes números" establecida por la estadística teórica (Corbalán y Sanz: 2011).

Dentro de este estándar de uso, encontramos ciertos sesgos hacia los sustantivos en hombres, hacia los verbos en mujeres, hacia los adverbios en personas mayores y hacia los sustantivos y adjetivos en personas con un alto nivel educativo. Estos sesgos que constituyen la norma lingüística son fiables, puesto que el coeficiente de variación de la RCG se volvió lo suficientemente pequeño debido a la cantidad expandida del número de materiales.

Referencias

Corbalán, Fernando y Gerardo Sanz. 2011. *La conquista del azar. La teoría de probabilidades.* Navarra. RBA Coleccionables.

Coseriu, Eugenio. 1973. *Teoría del lenguaje y lingüística general. Cinco estudios.* Madrid: Gredos.

Juilland, A. and E. Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton.

Involvement in male and female late Modern English scientific discourse: Coruña Corpus

Leida María Mónaco

Universidade da Coruña

Previous research on the language of women and men suggests a number of generalisations, specifically, that women use language to prioritise community relationships and cooperation, while men use it to assert dominance and control (Romaine 1996: 117). Likewise, it has been suggested that female discourse tends to be less detached, more personal, tentative and narrative than male, whereas the latter is more informational and contains more persuasive strategies (see Lakoff 1990; Biber & Burges 2000; Argamon et al. 2003). Using Biber's (1989) multidimensional analysis, Geisler's (2003) study of nineteenth-century letters seems to confirm that male and female language presents differences along several dimensions of variation, the former showing more abstractness and elaboration, and the latter more involvement, becoming more persuasive with time. By contrast, a comparative microscopic analysis of scientific and non-scientific female nineteenth-century texts (Crespo 2019) demonstrates that female historians in the 1800s tended to write in a more objective and impersonal way than their non-scientist colleagues, reflecting the general trend of scientific

writing – which was predominantly male – during that period (Halliday 1988; Taavitsainen 1994; Atkinson 1999).

The aim of this study is to look at variation in scientific texts belonging to the Coruña Corpus of English Scientific Writing (Moskowich & Crespo 2007; Crespo & Moskowich 2020) written by nineteenth-century male and female historians and life scientists (including biologists, zoologists, geologists and botanists) in order to spot differences related to subregister (i.e. sex of the author and scientific discipline). The analysis of history texts is currently underway, whereas the data for life sciences texts are extracted from a previous multidimensional analysis of register variation (Monaco 2017), scrutinised this time through the lens of the sex-of-the-author variable. Preliminary findings reveal that there is no variation between male and female nineteenth-century life sciences texts with respect to Dimension 1 (Involved vs Informational Style), and we expect to see a similar picture in the history subcorpus, which, if it were the case, would confirm Crespo's (2019) findings that female scientific writing was evolving in a similar way as male scientific writing in the 1800s. On the other hand, a(nother) previous microscopic study by Crespo & Moskowich (2015) of involvement features in nineteenth-century life sciences and history texts – which focused on those written by female authors only – suggests that the latter have a lower degree of involvement than the former, which is likewise expected to be reflected in terms of dimension scores in the present study. The genre variable (i.e. whether the texts analysed are treatises, articles, or lectures, among other possible genres) will also be taken into consideration in order to offer a more complete picture of variation.

References

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse*, *23*(3). https://doi.org/10.1515/text.2003.014

Atkinson, D. (1999). Scientific discourse in sociohistorical context: The Philosophical Transactions of the Royal Society of London, 1675–1975. Mahwah, NJ: Erlbaum

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511621024.

Biber, D., & Burges, J. (2000). Historical change in the language use of women and men. *Journal of English Linguistics*, *28*(1) https://doi.org/10.1177/00754240022004857

Crespo, B. (2019). "How intimate was the tone of female history writing in the Modern period? Evidence from the Corpus of History English Texts". In Moskowich, Isabel; Begoña Crespo, Luis Puente-Castelo & Leida Maria Monaco (Eds.) *Writing history in Late Modern English: Explorations of the Coruña Corpus* (pp. 186–213). Amsterdam: John Benjamins.

Crespo, B., & Moskowich, I. (2015). Involved in writing Science: Nineteenth-Century Women in the Coruña Corpus. International Journal of Language and Linguistics, *2*(5), 76–88. http://ijllnet.com/journals/Vol_2_No_5_November_2015/8.pdf

Crespo, B., & Moskowich, I. (2020). Astronomy, Philosophy, Life Sciences and History Texts: Setting the scene for the study of modern scientific writing. English Studies, *101*(6), 665–684. https://doi.org/10.1080/0013838x.2020.1798635

Geisler, C. (2003). Gender-Based variation in Nineteenth-Century English letter writing. In P. Leistyna and C. F. Meyer (Eds.) *Corpus Analysis* (pp. 87–106). Brills. https://doi.org/10.1163/9789004334410_007

Halliday, M. A. K. (1988). On the language of physical science. In M. Ghadessy (Ed.), *Registers of written English: Situational factors and linguistic features* (pp. 162-178). London: Pinter.

Lakoff, R. T. (1990). *Talking Power: The politics of language in our lives*. New York: Basic Books.

Monaco, L. M. (2017). A multidimensional analysis of late Modern English scientific texts from the Coruña Corpus. Unpublished PhD Dissertation. Available at:

https://ruc.udc.es/dspace/bitstream/handle/2183/19322/Monaco_LeidaMaria_TD_2017.pdf?sequence=2

Moskowich, I. & Crespo, B. (2007). Presenting the Coruña Corpus: A Collection of Samples for the Historical Study of English Scientific Writing. In J. Pérez Guerra et al. (Eds.), *'Of Varying Language and Opposing Creed': New Insights into Late Modern English* (pp. 341-357). Bern: Peter Lang. 341–357.

Romaine, S. (1996). *Language in society: An introduction to sociolinguistics*. Oxford: Oxford University Press.

Taavitsainen, I. (1994). On the evolution of scientific writing between 1375 and 1675: repertoire of emotive features. In F. Fernández et al. (Eds.), *Papers from the 7th International Conference on English Historical Linguistics* (pp. 329-342), Valencia, Sept. 1992. Amsterdam/Philadelphia: John Benjamins.

Retrospective verbs across time

Raquel Pereira Romasanta

Universidade de Santiago de Compostela

The clausal verb complementation profiles of the retrospective verbs (*remember*, *regret*, and *forget*) have drawn the attention from researchers. This set of verbs allows catenative complementation with either infinitive complements (indicating prospective actions) or *-ing* complements (indicating reprospective actions). To this group, the verbs *recall* and *recollect* can be added, as they are exclusively used with retrospective meanings. Moreover, these verbs exhibit non-categorical or probabilistic variability between finite *that*- and zero-clauses and nonfinite *-ing* clauses. This alternation has been the primary focus of diachronic research, particularly concerning *remember* and *regret* (Cuyckens et al., 2014; Fanego, 1996; Heyvaert & Cuyckens, 2010). Research generally indicates an increased use of *-ing* clauses complementing these verbs over time, often at the expense of finite *that*-/zero-clauses. Additionally, Fanego (1996, p. 74) observes that *remember* is "the only retrospective predicate that occurs in both the *-ing* and infinitival constructions prior to the second half of the eighteenth century" and suggests that it may have set a pattern for the other retrospective verbs. However, this claim, as well as the complementation profiles of *forget*, *recall*, and *recollect* remain underexplored.

At a synchronic level, the verbs *remember* and *regret* have been studied individually across various English varieties (e.g., García-Castro, 2018, 2019, 2020 with *remember*; Author, 2019, 2021, 2022, 2023 with *regret*). Research on *forget* is currently underway (Pavón, University of Vigo). Overall, studies suggest that nativized varieties of English tend to use *-ing* clauses less frequently, often attributed to cognitive processes associated with second language acquisition, such as simplification. Despite this, the retrospective verbs *recall* and *recollect* remain largely neglected in the literature.

The present study aims to consolidate existing research and investigate the complementation patterns of all five retrospective verbs (*remember*, *forget*, *regret*, *recall*, and *recollect*), examining both diachronic and synchronic dimensions. Using data from CEECS (*Corpus of Early English Correspondence Sampler*, Nevalainen et al., 1998), CLMET 3.1 (*Corpus of Late Modern English texts*; De Smet et al., 2015), and GloWbE (*Corpus of Global-Web Based English*; Davies, 2015), the diachronic analysis traces the historical evolution of these verbs, from Old English to Present-day English, highlighting a gradual increase *-ing* complementation, particularly for *remember* and *regret* since the 20th century. The synchronic analysis explores four varieties of English—British, Bangladeshi, Indian, and Sri Lankan—revealing significant regional differentiation. In nativized varieties, there is a marked reduction in the use of nonfinite *-ing* clauses and notable variability across verb, with *recall* and *recollect* favoring finite

constructions. Furthermore, the study identifies innovative patterns emerging in nativized varieties, including the increasing use of prepositional phrases as complementation structures.

References

Author. (2019).
Author. (2021).
Author. (2022).
Author. (2023).
Cuyckens, H., D'hoedt, F., & Szmrecsanyi, B. (2014). Variability in verb complementation in Late Modern English: Finite vs. non-finite patterns. In M. Hundt (Ed.), *Late Modern English syntax* (pp. 182-203). Cambridge University Press. https://doi.org/10.1017/CBO9781139507226.014.
Davies, M. (2015). Introducing the 1.9 billion word Global Web-Based English Corpus (GloWbE). *The 21st Century Text 5*.
De Smet, H., Flach, S., Tyrkkö, J., & Diller, H.-J. (2015). *The Corpus of Late Modern English (CLMET), version 3.1: Improved tokenization and linguistic annotation*. KU Leuven, FU Berlin, U Tampere, RU Bochum. Available from https://perswww.kuleuven.be/~u0044428/clmet3_1.htm.
Fanego, T. (1996). On the historical development of English retrospective verbs. *Neuphilologische Mitteilungen*, *97*, 71-79.
García-Castro, L. (2018). *The complementation profile of remember in post-colonial Englishes* (Unpublished doctoral dissertation). Universidade de Vigo.
García-Castro, L. (2019). Synchronic variability in the complementation profile of remember: finite vs non-finite clauses in Indian and British English. *Miscelánea: A Journal of English and American Studies, 59*, 137-164. https://doi.org/10.26754/ojs_misc/mj.20196343.
García-Castro, L. (2020). Finite and non-finite complement clauses in postcolonial Englishes. *World Englishes*, *39*(3), 411-426. https://doi.org/10.1111/weng.12481.
Heyvaert, L., and Cuyckens, H. (2010). Finite and gerundive complementation in Modern and Present-day English: Semantics, variation and change. In M. E. Winters, H. Tissari, & K. Allan (Eds.), *Historical cognitive linguistics* (pp. 132-159). Mouton de Gruyter. https://doi.org/10.1515/9783110226447.132
Nevalainen, T., Raumolin-Brunberg, H., Keränen, J., Nevala, M., Nurmi, A., & Palander-Collin, M. (1998). *Corpus of Early English Correspondence Sampler (CEECS)*. Department of Modern Languages, University of Helsinki. Available from https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2461.

Más allá de la vaguedad: Los apéndices generalizadores en el corpus ESLORA

Paula Rodríguez Abruñeiras

Universidade de Santiago de Compostela

Este estudio examina el uso de apéndices generalizadores (AGs) (Gille y Håggkvist 2006; *general extenders* en su terminología anglosajona; Overstreet 1999) en el español oral de Galicia, prestando especial atención a la fórmula 'y tal'. A partir del corpus ESLORA, que recoge conversaciones y entrevistas informales, se identifican numerosas fórmulas que actúan como AGs, destacando por su frecuencia 'y tal', pero también otras como 'y todo', 'o algo' y 'ni nada'. El trabajo pretende dar respuesta a las siguientes preguntas de investigación:

PI1. ¿Cómo influyen factores sociolingüísticos, como la edad y el sexo de los hablantes, en el uso de los AGs en general y de 'y tal' en particular?

PI2. ¿Qué diferencias de uso existen entre los géneros discursivos analizados (conversaciones vs. entrevistas)?

PI3. ¿Qué funciones desempeña 'y tal' y cómo influyen estas en la estructuración del discurso?

Los datos de corpus muestran que el uso de AGs es común en todas las franjas de edad (resultados que concuerdan con los de Borreguero Zuloaga 2022), aunque los jóvenes tienden a utilizarlos con mayor frecuencia. Sin embargo, se detecta un patrón inverso para 'y tal', más común en hablantes mayores de 54 años. Esto sugiere que 'y tal' es una fórmula muy consolidada en el español gallego y con una larga tradición en esta variedad. Desde una perspectiva de género, el estudio revela que los hombres usan AGs, y en particular 'y tal', más frecuentemente que las mujeres. Esta diferencia se podría atribuir a los estereotipos masculinos de economía lingüística y priorización de la información (Cheshire 2005; Overstreet y Yule 2021). El análisis también muestra que, si bien no existen diferencias significativas en el uso de AGs en general en los dos tipos de discurso analizados, 'y tal' sí es considerablemente más frecuente en entrevistas que en conversaciones (resultados que concuerdan con Montañez Mesas 2008), lo que parece deberse a su función recurrente de cierre discursivo, facilitando el cambio de turno entre emisor e interlocutor o la delimitación de fragmentos temáticos. Esta función es particularmente útil en las dinámicas de las entrevistas, donde la gestión clara de los turnos de habla es esencial.

Si nos centramos de forma más detallada en 'y tal', las funciones de este AG se clasifican en dos grandes categorías (Borreguero Zuloaga 2022): funciones interactivas (imprecisión, aproximación) y funciones metadiscursivas (cierre de listados, delimitación de ejemplos, marca de discurso citado, señalización de lugares de relevancia transicional). Estas funciones permiten al hablante organizar el discurso y facilitar la comprensión del interlocutor. La multifuncionalidad de 'y tal' pone de relieve su valor pragmático, demostrando que, aunque es semánticamente vago, puede aportar una gran precisión en la interacción comunicativa. Por lo tanto, este estudio destaca la importancia de los AGs en la construcción del significado y en la gestión de las dinámicas sociales en el habla. Los resultados sugieren que las partículas de vaguedad no son meramente elementos de relleno, sino herramientas esenciales para una comunicación efectiva.

Referencias

Borreguero Zuloaga, M. (2022). General extenders in Spanish interactions: Frequent forms, pragmatic functions y todo eso. *Anuari de Filologia Estudis de Lingüística*, *12*, 155-187. https://doi.org/10.1344/AFEL2022.12.8.

Cheshire, J. (2005). Syntactic variation and beyond: Gender and social class variation in the use of discourse-new markers. *Journal of Sociolinguistics*, *9*, 479-508. https://doi.org/10.1111/j.1360-6441.2005.00303.x .

ESLORA: *Corpus para el estudio del español oral*, versión 2.2 de noviembre de 2023, ISSN: 2444-1430. http://eslora.usc.es .

Gille, J. y Häggkvist, C. (2006). Los niveles del diálogo y los apéndices conversacionales. En J. Falk, J. Gille y F. Wachtmeister Bermúdez (Eds.), *Discurso, interacción e identidad. Homenaje a Lars Fant* (pp. 65-80). Stockholm University.

Montañez Mesas, M.P. (2008). La partícula *y tal* en el español hablado de Valencia. *ELUA*, *22*, 193-212. https://doi.org/10.14198/ELUA2008.22.10.

Overstreet, M. (1999). *Whales, Candlelight, and Stuff like That. General Extenders in English Discourse*. Oxford University Press.

Overstreet, M. y Yule, G. (2021). *General Extenders. The Forms and Functions of a New Linguistic Category*. Cambridge University Press.

The effect of plain language on lexical aspects of UK legal decisions

Paula Rodríguez Puente

Universidad de Oviedo

This paper explores the development of the language of British legal decisions after the emergence in the 1970s of the Plain Language Movement. In the UK the Plain English Campaign was launched in 1979 and has had a notable effect on the language of English legal documents, where there has been a significant reduction and simplification of certain linguistic features, such as nominalizations, the passive voice and the modal *shall*, as well as changes directed to the development of a less abstract, more interpersonal style (Williams 2007, 2013, 2022; Rodríguez-Puente 2019, 2020, 2024).

The data come from *The Corpus of Contemporary English Legal Decisions, 1950-2021* (CoCELD; Rodríguez-Puente & Hernández-Coalla, 2022), a specialized corpus of legal decisions which contains case law produced within the UK and the Commonwealth. The period covered by CoCELD (1950-2021) makes this corpus particularly suitable to explore recent developments in the language of judicial decisions which may have been triggered since the emergence of the Plain Language Movement. More precisely, the focus is on the evolution of the use of certain aspects of the lexicon signalled as "to be avoided" in several legal drafting manuals (e.g. Butt & Castle, 2006; Rose, 2017), namely 1) compound adverbs (e.g. *herewith*, *thereby*), 2) vague referential words (e.g. *the said*, *the same*, *aforementioned*), 3) Latin words and expressions (e.g. *bona fide*, *inter alia*), and 4) doublets (e.g. *fit and proper*, *unless and until*). Preliminary results indicate that, despite concrete proposals to write in a type of language which can be readily understood by those who are affected by the law and suggestions for amendments in legal drafting manuals directed to achieve this goal, there are still numerous barriers to improving the transparency and effectiveness of legal writing.

References

Butt, P. & Castle, R. (2006). *Modern legal drafting: A guide to using clearer language*. Cambridge University Press. https://doi.org/10.1017/CBO9781107282148 .

Rodríguez-Puente, P. (2019). Interpersonality in legal written discourse. A diachronic analysis of personal pronouns in law reports, 1535 to present. In T. Fanego and P. Rodríguez-Puente (Eds.), *Corpus-based research on variation in English legal discourse* (pp. 172-199). John Benjamins. https://doi.org/10.1075/scl.91.08rod.

Rodríguez-Puente, P. (2020). Historical legal discourse: British law reports. In E. Friginal & J.A. Hardy (Eds.), *The Routledge handbook of corpus approaches to discourse analysis* (pp. 499-517). Routledge. https://doi.org/10.4324/9780429259982 .

Rodríguez-Puente, P. (2024). Is legal discourse really 'outside the ravages of time'? A diachronic analysis of nominalizations in British judicial decisions. In L. Caon, M. S. Gordon & T. Porck (Eds.), *Unlocking the history of English. Pragmatics, prescriptivism and text types* (pp. 102-129). John Benjamins. https://doi.org/10.1075/cilt.364.05rod.

Rodríguez-Puente, P. & Hernández-Coalla, D. (2022). *The Corpus of Contemporary English Legal Decisions, 1950-2021 (CoCELD)*. University of Oviedo.

Rose, R. (2017). *Commonwealth legislative drafting manual*. Commonwealth Secretariat. https://doi.org/10.14217/9781848599635-en .

Williams, C. (2007) *Tradition and Change in Legal English. Verbal Constructions in Prescriptive Texts* (2nd. edn.), Peter Lang. https://doi.org/10.3726/978-3-0351-0317-5.

Williams, C. (2013). Changes in the verb phrase legislative language in English. In B. Aarts, J. Close, G. Leech & S. Wallis (Eds.), *The verb phrase in English: Investigating recent language*

change with corpora (pp. 353–371). Cambridge University Press. https://doi.org/10.1017/CBO9781139060998.015
Williams, C. (2022) *The impact of plain language on legal English in the United Kingdom*. Routledge. https://doi.org/10.4324/9781003025009

On the distribution of try to + INF and try and + INF in some varieties of English

Jesús Romero Barranco

Universidad de Málaga

Infinitival complementation in present-day English is, in most cases, introduced by *to*. A few verbs, however, present different patterns for the introduction of these clauses and, consequently, three different groups of verbs could be distinguished in these environments: 1) verbs such as *want*, which selects for a *to*-clause (e.g. *They want to sign the agreement*); 2) verbs such as *dare* or *help*, which select either for a *to*-clause or a bare infinitive (e.g. *Could you help me sign this form? / The new organisation will help to manage with the new contracts*); and 3) a verb like *try*, which allows for the subordinator *to* along with *and* as markers of infinitival subordination (e.g. *The prisoners were trying to escape / The judge will try and find out what happened that night*).

The exceptional use of *and* as an infinitive marker in these structures has been considered by some authors as exclusive of the verb *try* (Biber et al. 1999, 738; Tottie 2012, 201), while other scholars have identified some similar structures such as *remember and* (Ross 2013, 121–22; 2014, 211). The distribution of *try and* is described in standard grammars such as Quirk et al. (1985: 978-979), Biber et al. (1999: 738-739), and Huddleston and Pullum (2002: 1302); and its use in L1 varieties of English has been discussed in Lind (1983), Kjellmer (2000), Rohdenburg (2003), Hommenberg and Tottie (2007) and Brook and Tagliamonte (2016). From a historical perspective, the variants *try to* and *try and* appeared almost simultaneously in the late 1500s and have developed differently in the dialects of English (see Tottie 2012; Ross 2013, 2014). The distribution of these two variants in World Englishes is, as far as I have been able to investigate, yet unexplored, and their analysis, together with some conditioning factors at play, could shed some new light on the phenomenon. The present paper, therefore, attempts to answer the following research questions: 1) What is the distribution of the variants *try to* and *try and* in a set of varieties of English (i.e. Indian English, Hong Kong English, Philippine English and Singapore English)?; 2) Are there other verbs introducing infinitive clauses with *and* in the English varieties studied? and 3) What factors (formality, dialect variation, phonology, etc.) may have influenced the choice of one of the variants in a given context? The source of evidence comes from the Indian, Hong Kong, Philippine and Singaporean components of the *International Corpus of English*.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, U.K.: Pearson Education.
Brook, M., & Tagliamonte, S. A. (2026). Why Does Canadian English Use *Try to* but British English Use *Try and*? Let's Try and/to Figure It Out. *American Speech* 91.3: 301–326.
Hommerberg, C., & Tottie, G. (2007). Try to or try and? Verb Complementation in British and American English. *ICAME Journal 31*, 45–64.
Huddleston, R, & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Kjellmer, G. (2000). Auxiliary Marginalities: The Case of try. In J. M. Kirk (Ed.), *Corpora Galore: Analyses and Techniques in Describing English; Papers from the Nineteenth International*

Conference on English Language Research on Computerised Corpora (ICAME 1998) (pp. 115–24). Amsterdam: Rodopi.
Lind, Å. (1983). The Variant Forms try and/try to. *English Studies 64.6*, 550–63.
Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
Rohdenburg, G. (2003). Cognitive Complexity and horror aequi as Factors Determining the Use of Interrogative Clause Linkers in English. In G. Rohdenburg and B. Mondorf (Eds.), *Determinants of Grammatical Variation* (pp. 205–49). Berlin: Mouton de Gruyter.
Ross, D. (2013). Dialectal Variation and Diachronic Development of try-Complementation. In D. Ross and A. Kimball (Eds.), *Proceedings of the Fifth Meeting of the Illinois Language and Linguistics Society*, *Studies in the Linguistic Sciences: Illinois Working Papers*, *38*, 108–47.
Ross, Daniel. (2014). The Importance of Exhaustive Description in Measuring Linguistic Complexity: The Case of English *try* and Pseudocoordination." In F. J. Newmeyer and L. B. Preston (Eds.), *Measuring Grammatical Complexity* (pp. 202–16). Oxford: Oxford University Press.
Tottie, G. (2012). On the History of *try* with Verbal Complements. In S. Chevalier and T. Honegger (Eds.), *Words, Words, Words: Philology and Beyond; Festschrift for Andreas Fischer on the Occasion of his 65th Birthday* (pp. 199–214). Tübingen: Francke.

Compiling a corpus to trace the evolution of metalinguistic variation in Early Modern English texts: The MetaLing Project

Daniel Russo

University of Insubria

This paper is part of the MetaLing project (Andreani & Russo 2023), which seeks to compile a corpus of Early Modern English texts that describe language and linguistic phenomena, focusing on metalinguistic terminology between 1500 and 1700. The project integrates historical linguistics, corpus linguistics, and digital humanities to explore the evolution of linguistic metalanguage and its socio-historical implications.

This study employs a hybrid methodology combining close textual analysis with computational tools (Koester 2010; Sangiacomo et al. 2022), facilitating the extraction of a representative corpus from texts spanning diverse genres. By enriching the MetaLing database with metalinguistic terms derived from these sources, we aim to trace the trajectories of linguistic standardisation, language contact, and conceptual shifts in the Early Modern English-speaking world. Our study highlights the dynamic nature of language contact, linguistic standardisation, and the evolution of vernacular languages within the broader context of European linguistic history. By examining the terminological implications of linguistic variety, we aim to provide scholars with valuable insights into the historical underpinnings of language teaching and the role of metalinguistic awareness in educational contexts. This research enhances our understanding of linguistic metalanguage and sheds light on the socio-cultural and historical factors that influenced language teaching practices over the centuries. Our findings underscore the dynamic evolution of linguistic metalanguage within the broader framework of European intellectual history. They offer critical insights into how Early Modern scholars conceptualised language and its variations.

We focus on the creation of this corpus using R.C. Alston's *A Bibliography of the English Language from the Invention of Printing to the Year 1800*, specifically volume 9 (*English Dialects, Scottish Dialects: Cant and Vulgar English*, 1971). This volume is foundational for uncovering lexical diversity, regional variation, and sociolectal dynamics, offering a gateway to understanding the interplay between linguistic variation and emerging metalinguistic concepts.

Alston's bibliography provides a wide-ranging repository of sources, which we analyse to identify and contextualise the terminological shifts and developments that informed Early Modern linguistic discourse.

This interdisciplinary approach embodies the synergy of traditional and digital methods in corpus creation, shedding light on the interconnections between linguistic variation, historical contexts, and evolving computational paradigms.

References

Alston, R. C. (1971). *A bibliography of the English language from the invention of printing to the year 1800: A systematic record of writings on English, and on other languages on English, based on the collections of the principal libraries of the world* (Vol. 9: *English dialects, Scottish dialects: Cant and vulgar English*). Leeds: Scolar Press.

Andreani, A. & Russo, D. (2023). Building a corpus of the metalanguage of English linguistics 1500-1700: Methodological issues. *Linguistica e Filologia*, *43*, 151–174.

Koester, A. (2010). Building small specialised corpora. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66–79). London: Routledge.

Sangiacomo, A., Tanasescu, R., Donker, S., & Hogenbirk, H. (2022). Mapping the evolution of early modern natural philosophy: Corpus collection and authority acknowledgement. *Annals of Science*, *79*(1), 1–39.

Female literary representation of grammatical features in the Devonshire dialect (1850-1950): A corpus study

Almudena Santalla Rodríguez

Universidad de Salamanca

This paper explores the unconventional morphosyntactic features depicted in the works of thirty-one women using Devonshire dialect within the span of one century – from mid-19th- to mid-20th-century in an attempt to ascertain whether these features comply to the characteristics classified by well-known linguists such as Richard Wakelin or Ossi Ihalainen, among others. It departs from the premise, as Sumner Ives states, that the grammatical forms used "do not appear in the textbooks – except as awful warnings" (Ives, 1971[1950]:137) and that they include "archaisms which are no longer part of the standard, such as double negatives, and faulty congruence, such as singular verb with a plural subject" (Blake, 1981:15). To do so, I undertake both a quantitative and qualitative analysis of the most prominent grammatical features. For the quantitative analysis, Antconc has served as the main tool and the starting point to classify the main features in terms of frequency, whereas for the qualitative analysis, the data obtained have been compared to the evidence provided by 19th-century non-literary sources. In order to trace the chronological evolution, all the works within the corpus have been classified accordingly. Thus, the date of publication sets the timeline, but also the division between dialect literature and literary dialect has been taken into account (Shorrocks, 1996), as DL texts were written for a more reduced audience, but its main goal was to represent the dialect as vividly as possible, whereas LD works contained less dialectal traits but were intended for a wider audience, and sometimes the features depicted were more literary devices than real representations of the dialect. The first results apparently show two relevant aspects to take into consideration. First, that some of the features included by linguists had already or almost disappeared during this period, as is the case of *ich* for 'I' or its proclitic form *cham* for 'I am,' which are not present in the corpus. At the same time, there are some others which are recurrent in the literary corpus, as for instance the use of for to + infinitive in purpose clauses or the use of

the suffix <-like> in adverbs, but which are not included in the non-literary representations of the dialect. On the other hand, the analysis shows the progressive disappearance of these non-standard features, which attests the influence of the 1870 Education Act and the rising mobility of population due to the improvement of the means of transport. Thus, this paper seeks to add to this field of research by looking at these literary representations of Devonshire dialect and, by the comparison with non-literary studies, obtain more information about the grammatical characteristics of the dialect and the paved way to standardisation which was progressively advancing in that period under study.

References

Primary sources:

Bray, A.E. (1871). *Hartland Forest*. Longmans, Green & Co.
Chanter, G. (1896) *The Witch of Withyford*. MacMillan & Co. Ltd.
Chanter, G.(1901). *The Rainbow Garden*. R. Brimley Johnson.
Chase, B. (1915). *Through a Dartmoor Window*. Longmans, Green and Co.
Clarke, Mrs. H. (1898). *A Lad of Devon*. Thomas Nelson and Sons.
Coleridge, C.R. (1893) *Waynflete*. A. D. Innes & Co.
Corelli, M. (1896). *The Mighty Atom*. Methuen & Co. Ltd.
Crook, M. (?) *A Bit o'Binder String*. In Marten, C. Flibberts and Skriddicks (Eds.) (pp. 28-29). Peninsula Press.
Crook, M. (?) *Oh t'be a Blackburd*. In Marten, C. Flibberts and Skriddicks (Eds.) (p. 9). Peninsula Press.
Dalzell, E. (1895). 'Anner, a West Country Tragedy. *Cassell's Family Magazine*, 328-338.
Dart, E. (1922). *Sareel*. Boni and Liveright.
Hartier, M. and Hartier, O. (1896). *Home Brewed*. Simpkin Marshall Hamilton Kent & Co.
Hawker, B. (1898). *Overlooked*. Wells Gardner, Darton & Co.
Hewett, S. (1892). *The Peasant Speech of Devon*. Elliot Stock.
Kelly, M.E. (1929). *The Pageant of Bradstone*. J.H. Lawrence.
Malet, L. (1919). *Deadham Hard: A Romance*. Dodd, Mead & Co.
O'Neill, H.C. (1892). *Devonshire Idyls*. Gibbings & Co. Ltd.
Palmer, M. R. (1837). *A Dialogue in the Devonshire Dialect*. Richard Taylor.
Pasture, Mrs. H. de la (1907). *Peter's Mother*. E. P. Dutton & Co.
Peard, F.M. (1880). *Mother Molly*. George Bell & Sons.
Pedler, M. (1920). *The Hermit of the Far End*. George H. Doran Co.
Pennell, M.E. (1909-10). A Devonshire Singer. *The Devon and Exeter Gazette*.
Rita (1916). *The Iron Star. A Romance of Dartmoor*. G. P. Putnam Sons.
Sharland, E. C. (1885). *Ways And Means In A Devonshire Village*. Society for Promoting Christian Knowledge.
Shrimp, S. (1876?). *Jottings and Dottings in the Devonshire Dialect*. R. Richards.
Spender, Mrs. J.K. (1888) *Her Brother's Keeper*. Spencer Blackett.
St Aubyn, A. (1898). *A Fair Impostor. A Story of Exmoor*. F.V. White & Co.
Volo Non Valeo (1876). *The Old House of the Downs, by Volo non Valeo*. John Hodges.
Watson, H.H. (1906). *Andrew Goodfellow*. MacMillan & Co. Ltd.
Whitby, B. (1892). *The Awakening of Mary Fenwick*. D. Appleton and Company.
Willcocks, M.P. (1905). *Widdicombe*. John Lane the Bodley Head.
Zack (1896). *Life is Life*. Charles Scribner's Sons.

Secondary sources:

Beal, J.C. (2004). *English in Modern Times*. Taylor & Francis.

Beal, J.C. (2017). Nineteenth-century Dialect Literature and the Enregisterment of Urban Vernacular. In Jane Hodson (Ed.), *Dialect and Literature in the Long Nineteenth Century*. Routledge.

Blake, N. (1981). *Non-Standard Language in English Literature*. Westview Press.

Edney, S. (2011). Recent Studies in Victorian English Literary Dialect and its Linguistic Connections. *Literature Compass*, *8*, 660-674. https://doi.org/10.1111/j.17414113.2011.00831.x

Ekwall, E. (1980) [1975]. *A History of Modern English Sounds and Morphology*. Blackwell.

Ferguson, L. S. (1998). Drawing Fictional Lines: Dialect and Narrative in the Victorian Novel. In *Race, Gender, Religion, and Other Dangerous Things*, *32*(1), 1-17.

García-Bermejo Giner, M.F. (1999). Methods for the Linguistic Analysis of Early Modern English Literary Dialects. In P. Alonso et al. (Eds.) *Teaching and Research in English Language and Linguistics* (pp. 249-266). Celarayn.

García-Bermejo Giner, M.F. (2008). Towards a History of English Literary Dialects and Dialect Literature in the 18th and 19th Centuries: The Salamanca Corpus. In B. Heselwood and C. Upton (Eds.), *Proceedings of Methods XIII: Papers from the Thirteenth International Conference on Methods in Dialectology* (pp. 31-41). Peter Lang.

Görlach, M. (1998). *Annotated Bibliography of Nineteenth-Century Grammars of English*. John Benjamins Publishing.

Görlach, M. (1999a). *English in 19th Century England*. Cambridge University Press.

Görlach, M. (1999b). Attitudes towards British English Dialects in the 19th Century. *Leeds Studies in English*, *30*, 139-164.

Hodson, J. (2016) "Talking like a Servant: What nineteenth century novels can tell us about the social history of the language. *Journal of Historical Linguistics*, *2*(1), 27-46. doi 10.1515/jhsl-2016-0002.

Hodson, J. (Ed.) (2017). *Dialect and Literature in the Long Nineteenth Century*. Routledge.

Ihalainen, O. (1994). The dialects of England since 1776. In R. Burchfield (Ed.) *The English Language V: English in Britain and overseas: origins and Development* (pp. 197-274). Cambridge University Press.

Ives, S. (1971[1950]). A Theory of Literary Dialect Dialect. In J. Williamson and V. Burke (Eds.), *A Various Language* (pp. 145–177). Holt, Rinehart & Winston.

Kerswill, P. (2018). Dialect Formation and Dialect Change in the Industrial Revolution: British Vernacular English in the Nineteenth Century. In L. Wright (Ed.), *Southern English Varieties Then and Now* (pp. 8-38). De Gruyter.

Klemola, J. (2008). The Historical Geographical Distribution of Periphrastic Do in Southern Dialects. In L. Wright (Ed.), *Southern English Varieties Then and Now* (pp. 21-74). De Gruyter Mouton.

Matthews, W. (1939). South Western Dialect in the Early Modern Period. *Neophilologus*, *24*, 193-209.

Melchers, G. (2010). "Southern English in writing," in R. Hickey (Ed.) Varieties of English in Writing, pp. 81-98. John Benjamins.

Page, N. (1973). Speech in the English Novel. Longman.

Percy, C. (2020). "British Women's Roles in the Standardization and Study of English. In Ayres-Bennet Sanson (Ed.) *Women in the History of Linguistics*. Oxford University Press.

Picone, M. (2014). Literary Dialect and the Linguistic Reconstruction of Nineteenth Century Louisiana. *American Speech*, *89*(2), 143-169.

Priestley, J. (1772). The Rudiments of English Grammar: adapted to the Use of Schools. J. and F. Rivington.

Ruano-García, J. (forthcoming). The Language of Dialect Writing. R. Hickey, M. Kÿto, and E. Smitterberg (Eds.) *The New Cambridge History of the English Language. Vol. II Documentation, Data Sources and Modelling*. Cambridge UP.

Shorrocks, G. (1996). Non-Standard Dialect Literature and Popular Culture. J. Klemola, M. Kytö and M. Rissanen (Eds.), *Speech Past and Present. Studies in English Dialectology in Memory of Ossi Ihalainen* (pp. 385–411). Peter Lang.
Wakelin, M. F. (1986). *The Southwest of England*. John Benjamins Publishing Company.
Wright, J. (1898-1905). *English Dialect Dictionary*. 6 vols. Henry Frowde.

Distributional semantics meets sociolinguistics: A study of language attitudes among Spanish students

Mario Serrano Losada, Iván Tamaredo

Universidad Complutense de Madrid

The study of language attitudes has been a thriving subfield within sociolinguistic research since the 1960s (see, e.g., the seminal work by Lambert et al., 1960). Our attitudes toward different languages and language varieties are pervasive in our everyday lives, as we infer and make assumptions about our interlocutors on the basis of their accent or other features. Such cognitive shortcuts are often based on bias and prejudice (see Kristiansen, 2003), so that language attitudes may easily work against us, triggering discrimination (see, among others, Sancho-Pascual, 2020). The aim of this study is to uncover the attitudes of Spanish university students (n=69) towards ten stimuli representing different language varieties. The following features of the stimuli are considered in the analysis: language (English vs. Spanish), gender (male vs. female), nativeness (native vs. non-native), and, in the case of the native Spanish varieties, regional variation (Madrid vs. Murcia).

To measure language attitudes, we employed a two-step experimental design following De Pascale et al. (2018). First, participants completed a Free Association Task in which they listened to the ten stimuli and, as quickly and spontaneously as possible, provided the first three adjectives (in Spanish) that came to mind to describe each speaker. The second step involved classifying the adjectives provided by the participants into more general semantic categories by means of bag-of-words vector space modelling (Divjak & Fieller, 2014; Glynn, 2014; Peirsman & Geeraerts, 2009). The use of vector space models for semantic categorization is a computational implementation of the principles of distributional semantics, summarized in Firth's famous quote "You shall know a word by the company it keeps" (1957, p. 11). First, words co-occurring with the adjectives in an L5-R5 context window were retrieved from the *CORPES XXI* corpus. Then, the adjectives were grouped based on the similarity of their collocational profiles, clustering together those adjectives that co-occurred with the same words. Finally, the resulting groups of adjectives were interpreted and assigned a semantic category on the basis of their meanings.

The results show that participants' attitudes vary according to the factors examined, since they systematically portrayed, for instance, male and female speakers using adjectives belonging to different semantic categories: overall, while male speakers were associated with adjectives describing aspects of social interaction, female speakers were characterized using adjectives connected to appearance and kindness. However, the factors also interact, modulating their effects: both male and female non-native speakers of English are depicted using adjectives denoting insecurity. All in all, the results shed light on the language attitudes of Spanish university students, revealing a complex network of attitudes towards language variation.

References

CORPES XXI = Real Academia Española. *Corpus del Español del Siglo XXI* (online). http://www.rae.es (accessed July 2024)

De Pascale, S., Marzo, S., & Speelman, D. (2018). Cultural models in contact: Revealing attitudes toward regional varieties of Italian with Vector Space Models. In E. Zenner, A. Backus, & E. Winter-Froemel (Eds.), *Cognitive Contact Linguistics: Placing Usage, Meaning and Mind at the Core of Contact-Induced Variation and Change* (pp. 213–250). De Gruyter Mouton. https://doi.org/10.1515/9783110619430

Divjak, D., & Fieller, N. (2014). Cluster Analysis. Finding structure in linguistic data. In D. Glynn & J. Robinson (Eds.), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, (Vol. 43, pp. 405–441). John Benjamins Publishing Company. https://doi.org/10.1075/hcp.43.16div

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In J. R. Firth (Ed.), *Studies in Linguistic Analysis*, (pp. 1–32). Blackwell.

Glynn, D. (2014). Correspondence analysis: Exploring data and identifying patterns. In D. Glynn & J. Robinson (Eds.), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, (Vol. 43, pp. 443–485). John Benjamins Publishing Company. https://doi.org/10.1075/hcp.43.17gly

Kristiansen, G. (2003). How to do things with allophones: Linguistic stereotypes as cognitive reference points in social cognition. In R. Dirven, R. Frank, & M. Pütz (Eds.), *Cognitive Models in Language and Thought: Ideologies, Metaphors, and Meanings*, (pp. 69–120). De Gruyter Mouton. https://doi.org/10.1515/9783110892901.69

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology*, *60*(1), 44–51. https://doi.org/10.1037/h0044430

Peirsman, Y., & Geeraerts, D. 2009. Predicting strong associations on the basis of corpus data. In A. Lascarides, C. Gardent, & J. Nivre (Eds.), *Proceedings of the 12th conference of the European Chapter of the ACL (EACL 2009)* (pp. 648–656). Association for Computational Linguistics. https://aclanthology.org/E09-1074

Sancho-Pascual, M. (2020). Integrating the immigrant population into the city of Madrid (Spain): preliminary data about the sociolinguistic attitudes of the host community. *Journal of Multilingual and Multicultural Development*, *41*(1), 72–84. https://doi.org/10.1080/01434632.2019.1621880

Grammaticalization and phonetic reduction in spoken English: Insights from the sort / kind / type of X construction

David Tizón Couto

Universidade de Vigo

This study investigates the grammaticalization, phonetic reduction, and variation of the sort/kind/type of X (SKT) construction in spoken American English, offering a new perspective on how cognitive and communicative factors influence linguistic change. Previous research has documented the SKT construction's evolution from a binomial noun phrase (a kind of tree) to a qualifying adverbial and pragmatic marker (I kind of like this), highlighting shifts in function and grammatical status (cf. Aijmer, 1984; Brems & Davidse, 2010; Denison, 2011; Reichelt, 2021). In line with grammaticalization theory, increased 'ancillariness' (reduced discursive prominence; Boye & Harder, 2012) suggests that more grammaticalized forms are accompanied by prosodic backgrounding and phonetic reduction. Desemanticization, decategorialization and

phonetic reduction have been frequently discussed in connection with the SKT construction. Phonetic reduction leads to variant forms represented as kinda and sorta; regarding prosody, Dehé & Stathi (2016) have found that increasing grammaticalization is associated with decreasing prosodic prominence. On the other hand, reduction can also result from articulatory factors (speaking rate, phonological context), social context or item frequency. Therefore, an open question is how these factors interact in the usage of SKT along the grammaticalization cline. Do prosodic changes mark the earlier stages of the cline (as suggested by Dehé & Stahi, 2016: 939), and does phonetic reduction only occur at later stages? Are specific reduced variants (such as kinda) more strongly tied to a specific function than prosodic patterns (as would follow if forms are mentally stored but prosodic patterns are not)? A corpus-based quantitative analysis of 1,243 SKT tokens from the Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000–2005) and the Buckeye Corpus (Pitt et al., 2007) examines the interplay of function, phonetic form, prosody, duration, and contextual factors. We apply a structural equation model to capture the interrelations between variables. The results reveal a mixed picture: kind of exhibits a pattern consistent with grammaticalization-driven reduction but is also influenced by articulatory factors, challenging the entrenchment of kinda as a fully distinct variant. Despite showing some similar trends, realizations of sort of display greater variability. Our findings suggest that these items represent the grammaticalization of a constructional pattern led by kind of and echoed by less frequent forms like sort of, which trails behind.

References

Ajmer, K. (1984). Sort of and kind of in English conversation. *Studia Linguistica*, *38*, 118–128.

Boye, K., & Harder, P. (2012). A usage-based theory of grammatical status and grammaticalization. *Language*, *88*(1), 1–44.

Brems, L., & Davidse, K. (2010). The grammaticalisation of nominal type noun constructions with kind/sort of: Chronology and paths of change. English Studies, *91*(2), 180–202.

Dehé, N., & Stathi, K. (2016). Grammaticalization and prosody: The case of English sort/kind/type of constructions. *Language*, *92*(4), 911–946.

Denison, D. (2011, September 15–16). The construction of SKT. Paper presented at the 2nd Vigo-Newcastle-Santiago-Leuven International Workshop on the Structure of the Noun Phrase in English (NP2). Retrieved from https://www.escholar.manchester.ac.uk/uk-ac-man-scw:172513

Du Bois, J. W., Engelbertson, R., Chafe, W. L., Meyer, C., Thompson, S. A., & Martey, N. (2000–2005). *Santa Barbara Corpus of Spoken American English*, *Parts 1-4*. Philadelphia, PA. Retrieved December 1, 2022, from https://www.linguistics.ucsb.edu/research/sbcorpus.html

Pitt, M. A., Dilley, L. C., Johnson, K., Kiesling, S., Raymond, W. D., Hume, E., & Fosler Lussier, E. (2007). *Buckeye Corpus of Conversational Speech* (2nd release). Columbus, OH: Department of Psychology, Ohio State University (Distributor). Retrieved from https://www.buckeyecorpus.osu.edu

Reichelt, S. (2021). Recent developments of the pragmatic markers kind of and sort of in spoken British English. *English Language and Linguistics*, *25*(3), 563–580.

Hate speech analysis in Trump's 2024 presidential campaign

Belén Zapata Barrero, Antonio Joaquín Segura García

Centro Universitario de la Defensa, San Javier

Driven by the speed with which information is shared through social media networks, hate speech has become a relevant field of study when it comes to analysing the transmission of social meaning in interpersonal communication (Eckert, 2008, 2012; Silverstein, 2003; Subramanian, Sathiskumar, Deepalakshmi, Cho & Manikandan, 2023). Although said networks allow for the sharing of interesting information, they also foster the spread of harmful language that negatively affects certain communities. In this respect, the detection of harmful language becomes a challenge, since the boundaries between freedom of expression and regulation have become increasingly blurred (Jonker & Gomstyn, 2024).

The present study aims to (i) contribute to the understanding of the dynamics of hate speech, abusive language and profanity in order to ensure effective and responsible communication in political contexts and (ii) subsequently propose a methodology to detect and quantify harmful language to enhance the analysis of political discourses.

We focus on the speech production of Donald Trump across his race for the 2024 U.S. presidential elections, for which a corpus of his speeches delivered in different geographical locations of the U.S. was compiled. Mass media sources were employed as instruments for the obtention of the informant's speeches. They were processed and later quantitatively analysed implementing Large Language Models (LLMs), which were used for content safety classification. Further extra-linguistic factors were also taken into account, such as relevant events taken place in the context of the speeches analysed.

Results suggest that (i) LLMs can be effectively used for the detection and quantification of harmful language in the compiled corpus and that (ii) extra-linguistic factors can condition the level of harmful language used in a speech.

We conclude that LLMs are a valuable tool when it comes to political discourse analysis and, specifically, detecting potentially harmful content.

References

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453–476.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, *41*, 87–100.

Jonker, A., & Gomstyn, A. (December 24, 2024). Purifying AI: HAP filtering against harmful content. *IBM*. https://www.ibm.com/think/insights/hap-filtering

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language & communication*, *23*, 193–229.

Subramanian, M., Sathiskumar, V. E., Deepalakshmi, G., Cho, J., & Manikandan, G. (2023). A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, *80*, 110–121.

Panel 7

Corpus-based computational linguistics
Lingüística computacional basada en corpus

Relevancia de las particularidades técnicas en la interpretación de los resultados de las aplicaciones de búsqueda en corpus

Mario Barcala

NLPgo Technologies, S.L.

Las aplicaciones de consulta en corpus son utilizadas de manera habitual para realizar investigaciones lingüísticas (McEnery, 2012) (Rojo, 2021). Sin embargo, los/as investigadores/as que las usan no siempre son conscientes de que las particularidades técnicas de cada aplicación pueden afectar a los resultados obtenidos. Es decir, bajo la apariencia de interfaces similares, las aplicaciones de consulta en corpus pueden presentar diferencias significativas en su funcionamiento que, a la postre, afectan a la interpretación de los datos (Hiltunen, 2024) (Sass, 2022) (Rojo, 2021).

Este trabajo tiene como objetivo exponer algunas de las particularidades más relevantes que se pueden encontrar en las aplicaciones de búsqueda en corpus y que pueden influir en la interpretación de los resultados de las búsquedas. Exponemos brevemente, a continuación, algunas de estas problemáticas, que serán ilustradas con ejemplos concretos obtenidos de los corpus ESLORA (Vázquez et al., 2024), CORPES XXI (Real Academia Española, 2024), CORGA (Centro Ramón Piñeiro para a Investigación en Humanidades, 2024) y CODOLGA (Centro Ramón Piñeiro para a Investigación en Humanidades, 2023):

1. Palabras buscables: No todas las palabras incluidas en los textos de un corpus son accesibles para las consultas. Por ejemplo, es común que los fragmentos de texto en lenguas diferentes a la lengua principal del corpus se excluyan de las búsquedas, por lo que se hace necesario evaluar los resultados obtenidos antes de afirmar que una palabra no está presente en un corpus.

2. Anonimización en corpus orales: En los corpus de transcripciones orales, a menudo se incluyen palabras inventadas durante el proceso de anonimización de los participantes. Aunque estas palabras no forman parte del léxico natural, deben ser tratadas adecuadamente para evitar interferir en los resultados de búsqueda de los/as investigadores/as.

3. Adaptaciones en ediciones críticas: En los corpus basados en ediciones críticas, es habitual que la aplicación de consultas realice ciertas adaptaciones para mejorar la accesibilidad. Por ejemplo, suele ser esperable que si buscamos una palabra, obtengamos los resultados de esa palabra (independientemente de si incluye alguna marca de intervención editorial en su interior). No obstante, esto plantea desafíos a la hora de mantener la fidelidad a las marcas originales que aparecen en los documentos, lo que puede llevar a inconsistencias si no se tratan adecuadamente.

4. Interrupciones en la posición de las palabras: Cuando se buscan secuencias de palabras consecutivas, es crucial que la aplicación de consultas gestione correctamente su posición. Por ejemplo, ¿los resultados de una búsqueda de dos palabras consecutivas deberían incluir los casos en los que hay algún signo de puntuación entre ellas?

En conclusión, mediante la exposición de diferentes ejemplos concretos, en este trabajo demostramos que tener conocimiento de algunas particularidades técnicas de las

aplicaciones de consulta en corpus es fundamental para poder interpretar correctamente los resultados obtenidos de las mismas.

Referencias

Centro Ramón Piñeiro para a Investigación en Humanidades: *Corpus Documentale Latinum Gallaeciae* (CODOLGA), versión 20 (2023), <http://corpus.cirp.es/codolga>.

Centro Ramón Piñeiro para a Investigación en Humanidades: *Corpus de Referencia do Galego Actual* (CORGA) [4.1], <http://corpus.cirp.gal/corga/> [Actualizado el 11/04/2024]

Hiltunen, T. (2024). Early newspapers as data for corpus linguistics: Issues in using the British Library Newspapers database as a corpus, *Studies in Corpus Linguistics* 118, 68-88. https://doi.org/10.1075/scl.118.05hil

McEnery, T. & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice.

Real Academia Española: *Banco de datos* (CORPES XXI). (Versión 1.1, octubre de 2024). Corpus del Español del Siglo XXI (CORPES). <https://www.rae.es/corpes/>

Rojo, G. (2021). Introducción a la lingüística de corpus en español.

Sass, B. (2022). Principles of corpus querying: A discussion note, *Acta Linguistica Academica*, volume 69: issue 4, 599-614. https://doi.org/10.1556/2062.2022.00581

Vázquez, V. et al (2024). ESLORA: *Corpus para el estudio del español oral* <http://eslora.usc.es>, versión 2.3 de octubre de 2024.

Propuesta computacional para el análisis integral de material sociolingüístico: El caso de Preseea-Valencia

Adrián Cabedo Nebot

Universitat de València

Este trabajo se enmarca en el desarrollo de herramientas computacionales avanzadas para el análisis integral de material sociolingüístico sonoro. Se presenta una propuesta que implementa un flujo de trabajo completamente automatizado para el procesamiento de datos de habla, desde la transcripción hasta el análisis estadístico y la evaluación mediante inteligencia artificial. Este proceso combina diversos métodos de transformación lingüística y técnicas computacionales en un entorno integrado.

El proceso comienza con la transcripción automática de las grabaciones de audio, utilizando Whisper, un modelo de reconocimiento automático del habla (ASR) de alta precisión (Radford et al., 2022). Posteriormente, se realiza un alineamiento forzado entre el audio y la transcripción, permitiendo una sincronización palabra por palabra y sonido por sonido. Para esta tarea, se emplean técnicas de alineamiento avanzado basadas en modelos acústicos previamente validados (Yuan & Liberman, 2008). Este alineamiento se complementa con un etiquetado morfológico automático, que proporciona anotaciones detalladas sobre las características gramaticales de las palabras presentes en el corpus.

A continuación, se extraen parámetros acústicos y prosódicos a través de Praat, tales como el tono fundamental (F0), la intensidad y la duración, además de los patrones fonológicos del habla (Boersma & Weenink, 2022). Estos datos se combinan con la información morfológica y las características del alineamiento, creando una base de datos estructurada que permite análisis cruzados de múltiples dimensiones del habla. La base de datos utilizada representa una muestra limitada del corpus PRESEEA-Valencia, conformada por 18 entrevistas, 36 participantes (distinguiendo entre entrevistadores y entrevistados), con una duración total de 14 horas y un registro de 124848 palabras.

La herramienta incluye una funcionalidad de análisis automatizado mediante inteligencia artificial. Este componente aplica modelos de aprendizaje automático para identificar patrones en los datos y evaluar la relación entre las características acústicas, prosódicas y morfológicas en función de variables sociolingüísticas específicas, como el nivel de instrucción o el contexto comunicativo. Los análisis y transformaciones se desarrollaron en un entorno flexible basado en software de código abierto, como R, que permite integrar múltiples métodos en una única solución (R Core Team, 2023).

La propuesta busca no solo agilizar el procesamiento de grandes volúmenes de datos, sino también garantizar una mayor precisión y reproducibilidad en los estudios sociolingüísticos. Su implementación, basada en software de código abierto, ofrece una solución accesible y adaptable para investigadores interesados en explorar fenómenos complejos del habla a partir de corpus orales.

Referencias

Boersma, P., & Weenink, D. (2022). Praat: Doing phonetics by computer (versión 6.2). http://www.praat.org/

Radford, A., Kim, J. W., Hallacy, C., et al. (2022). Whisper: OpenAI's Automatic Speech Recognition System. OpenAI.

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation. https://www.R-project.org/

Yuan, J., & Liberman, M. (2008). Speaker diarization using forced alignment with acoustic models. Interspeech 2008.

Analyzing rhetorical transitions in evaluative texts: A computational approach to discourse structure and strategy

Javier Fernández Cruz, Carla Fernández Melendres

Universidad de Málaga

In the digital age, opinion texts wield considerable influence, particularly online, where they shape public sentiment and decision-making processes. Despite their importance, the complexity of discourse in opinion texts presents challenges for sentiment analysis (SA), a key tool in natural language processing. This study investigates the rhetorical organization and polarity lexicon of 150+ opinion texts, using an annotation framework that includes polarity, functional discourse units (FDUs) (Egbert et al., 2021), entities, and opinion holders. This study focuses on a specific feature of the annotated corpus: transitions in Functional Discourse Unit (FDU) structure. FDUs are autonomous, thematically unified, and fulfill distinct communicative functions. Using a combination of computational techniques (Tang, 2024), we employed computational techniques to calculate transition frequencies and probabilities between different FDU types across evaluative texts. This methodology facilitated the construction of transition matrices, which were later computed through probabilistic techniques to reveal discourse dynamics and identify rhetorical strategies in the editorials.

Rhetorical transitions are key to understanding the coherence and structure of evaluative texts (McKeown, 1985). Computational approaches, often grounded in Rhetorical Structure Theory (RST), have advanced discourse analysis by leveraging models such as the CODRA Framework (Joty et al., 2015), HILDA Parser (Hernault et al., 2010), and top-down neural architectures (Zhang et al., 2020). These methods parse discourse structures, identify relationships between units, and support the generation of discourse

strategies tailored to communicative goals. Challenges remain, including capturing inter-sentence relations and refining parsing strategies (dependency vs. constituency). Future directions focus on dynamic algorithms and improved models for natural language understanding. This research enhances applications in discourse analysis and NLP.

The analysis of transition probabilities between Functional Discourse Units (FDUs) in opinion texts reveals consistent structural patterns, with certain transitions occurring more frequently. These patterns reflect a strategic discourse organization where context is sustained, ideas are elaborated with additional information, and the flow progresses from presenting facts to evaluations and future projections. This highlights the logical and persuasive nature of these texts, essential for their influence on public opinion.

By visualizing and analyzing these transition probabilities, we provide a novel method for examining the discourse structure of annotated text. This analysis not only contributes to the understanding of rhetorical strategies in journalistic texts and tourism reviews but also bridges computational analysis with traditional discourse studies, offering new insights into the analysis of textual structure.

References

Egbert, J., Wizner, S., Keller, D., Biber, D., McEnery, T., & Baker, P. (2021). Identifying and describing functional discourse units in the BNC Spoken 2014. *Text & Talk*, 41(5–6), 715–737. https://doi.org/10.1515/text-2020-0053

Hernault, H., Prendinger, H., duVerle, D., & Ishizuka, M. (2010). HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue Discourse*, 1, 1-33. https://doi.org/10.5087/dad.2010.003.

Joty, S., Carenini, G., & Ng, R. (2015). CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, 41, 385-435. https://doi.org/10.1162/COLI_a_00226.

McKeown, K. (1985). Discourse Strategies for Generating Natural-Language Text. *Artif. Intell.*, 27, 1-41. https://doi.org/10.1016/0004-3702(85)90082-7.

Tang, J. (2024). Understanding English Rhetorical Strategies Based on Neurosemantic Analysis. *Applied Mathematics and Nonlinear Sciences*. https://doi.org/10.2478/amns-2024-2819.

Zhang, L., Xing, Y., Kong, F., Li, P., & Zhou, G. (2020). A Top-down Neural Architecture towards Text-level Parsing of Discourse Rhetorical Structure. , 6386-6395. https://doi.org/10.18653/v1/2020.acl-main.569.

Analysis of lexical diversity levels in human-originated texts and texts originated by ChatGPT in English and Spanish

Francisco Javier Ferre-Fuertes, Ángela Almela Sánchez

Universidad de Murcia

Over the last decade, artificial intelligence has undergone a significant upgrade, Large Language Models being the key point of it (Hayawi, Shahriar and Samuel, 2024). As a result, generative artificial intelligence appeared, becoming ChatGPT, from OpenAI, one of the most outstanding (Goar, Singh, and Singh, 2023; Hayawi *et al.*, 2024).

The tenets behind ChatGPT advancement rely on the use of learning algorithms to learn from what it is trained on (Goar *et al.*, 2023). To generate similar responses to those from humans, ChatGPT shows a wide range of syntactic structures and a large inventory of vocabulary. The present study focusses on lexical diversity to measure how skilful

ChatGPT is in comparison to humans, in order to point out, if possible, not only a weakness of the machine, but a useful pattern which could lead to the identification of artificially generated texts.

This project considers previous research in the field, that of Reviriego, Conde, Merino-Gómez, Martínez and Hernández (2023) being one of the most relevant. They analysed lexical diversity parameters between human-originated texts and ChatGPT-generated texts using the Root Type/Token Ratio method. Other researchers suggested other possibilities, such as the MTLD or the HD-D methods (Torruella and Capsada, 2013). Following Reviriego *et al*. (2023), this study uses the STTR method and the Mass index method to analyse lexical diversity using as sample texts four writing registers —economy, legal, medicine and general— in English and Spanish.

In order to obtain the samples, ChatGPT was asked to write some texts using those four writing registers, whereas human sample texts were obtained from newspapers in the case of Spanish samples, and from the ProQuest database in the case of English samples. All texts were gathered in single files depending on their source, their register and their language. In order to obtain the number of types and tokens, every single file was processed through *WordSmith* 7.0 (Scott, 2008). All the sample texts were analysed and cleaned before going through *WordSmith* to avoid misleading results. *WordSmith* showed diverse data, among which the most relevant was the number of types, the number of tokens and the STTR results for the different samples. Given the complexity of the Mass index method, the results from it were obtained from *Microsoft Excel* using the corresponding formula and the numbers of types and tokens previously mentioned.

The results showed the differences between human-generated texts and texts obtained from ChatGPT, the latest being the ones with the lowest lexical diversity. At the same time, it was also possible to appreciate differences at lexical diversity levels when looking at English sample texts and Spanish samples. The consistency of the results for both measurements points out that both methods are valid, although the STTR seems to be more precise. The knowledge obtained from this project might be applicable for fake authorship identification of texts created by any generative artificial intelligence.

References

Goar, V., Yadav, N. S., & Yadav, P. S. (2023). Conversational AI for natural language processing: A review of ChatGPT. *International Journal on Recent and Innovating Trends in Computing and Communication*, 11, 109-117. https://doi.org/10.17762/ijritcc.v11i3s.6161

Hayawi, K., Shahriar, S., & Mathew, S. S. (2024). The imitation game: Detecting human and AI-generated texts in the era of ChatGPT and BARD. *Journal of Information Science*. https://doi.org/10.1177/01655515241227531

Reviriego, P., Conde, J., Merino-Gómez, E., Martínez, G., & Hernández, J. A. (2023). Playing with words: Comparing the vocabulary and lexical richness of ChatGPT and humans. https://doi.org/10.48550/arXiv.2308.07462

Scott, M. (2008). WordSmith Tools Version 7. Lexical Analysis Software.

Torruella, J., & Capsada, R. (2013). Lexical statistics and typological structures: a measure of lexical richness. *Procedia-Social and Behavioural Sciences*, 95, 447-454. https://doi.org/10.1016/j.sbspro.2013.10.668

Mapping multiple meanings: Polysemy analysis with BERT's contextualized word embeddings

Reyes Gago Sosa

Universidad de Cádiz

The Bidirectional Encoder Representations from Transformers (BERT) is a generative artificial intelligence model designed to understand and generate human language by learning patterns in large volumes of text that was introduced by Devlin et al. (2029). BERT's architecture is based on a neural network that captures contextual meaning from both directions in a sentence, allowing it to produce Contextualized Word Embeddings (CWE). These are numerical vector representations of words that encode semantic meaning based on context. Unlike traditional word embeddings, which assign the same vector to a word regardless of its context, CWE creates distinct vectors for each instance of a word based on its specific use.

In this investigation, we employ BERT's CWE to examine its capacity to represent English polysemous words within the vector space. For this purpose, we utilize the SemCor 3.0 corpus (Miller et al., 1993), a semantically annotated English corpus containing 352 texts from the Brown corpus. Each text is labeled at the word level with WordNet 1.6 senses and later mapped to WordNet 3.0 senses.

The current research seeks to address the following research questions: How effectively can BERT disambiguate polysemous words? Does BERT have the ability to differentiate between distinct senses of the same term? The primary objective of this experimentation is to capture and visualize meaning differentiation through embeddings. Additionally, we aim to visualize the representations of each appearance of the selected words in the corpus to identify patterns in the distribution of their semantic vectors.

The methodology followed in this study is outlined as follows: The SemCor 3.0 corpus was pre-processed and BERT's tokenizer applied. Tokens were assigned contextual embeddings from BERT's final layer, as recommended by Yenicelik et al. (2020). Ten polysemous English words (nouns and verbs), e.g., *party*, *bank*, *run*, etc., were selected based on their frequency and WordNet sense tags.

Afterward, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the embeddings, enabling visualization and facilitating the use of the k-Nearest Neighbors (k-NN) algorithm, which classifies words based on their proximity in the vector space using Euclidean distance. Cross-validation was used to evaluate the model's generalization ability, while accuracy and F-score metrics were employed to assess its performance during the classification task.

Provisional results suggest that BERT is more effective at distinguishing homonymy than polysemy, with homonymy being defined as the convergence of unrelated meanings onto the same phonological form (Klein & Murphy, 2001). When handling polysemy, BERT shows improved performance in differentiating senses that have undergone significant semantic extension.

In conclusion, BERT demonstrates a certain degree of word sense knowledge, as evidenced by its moderate success in k-nearest neighbor (KNN) classification tasks. However, the limitations of WordNet sense classification are evident, particularly in its tendency to label very specific distinctions between senses. This raises questions about whether BERT introduces a more flexible, context-dependent notion of meaning compared to the rigid, linguistically defined categories of WordNet.

References

Akbik, A., Blythe, D. A. J., & Vollgraf, R. (2018). In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1638–1649). Association for Computational Linguistics. https://aclanthology.org/C18-1139

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, MN: Association for Computational Linguistics.

Harris, Z. (1954). Distributional structure. *Word*, 10(23), pp. 146–162.

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, 45(2), pp. 259–282. https://doi.org/10.1006/jmla.2000.2753

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., & Li, L. (2020). On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9119–9130). Online: Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop* at ICLR, 2013.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6 (1), pp. 1–28.

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro*, New Jersey, March 21-24, 1993.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). New Orleans, LA: Association for Computational Linguistics.

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *ArXiv*, abs/1909.10430.

Yenicelik, D., Schmidt, F., & Kilcher, Y. (2020). How does BERT capture semantics? A closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 156–162). Online: Association for Computational Linguistics.

LIWC-22 for emotion recognition: A corpus-based study

Paula Jiménez Sancho, Ángela Almela

Universidad de Murcia

A growing body of research indicates that patterns in linguistic choices can disclose underlying cognitive and emotional processes, as well as psychological dynamics and personality traits. This is clearly beneficial for applied linguists and researchers in psychology, as it allows gaining valuable insight into the human mind, but in order to conduct good research on these issues, we need to use ground truth data that relates directly to the question. One of the most widely used tools in the field is LIWC (Boyd et al.,

2022); this tool uses large dictionaries with words that are manually classified into different categories.

As this word-by-word automatic categorization has proved problematic in some recent studies (e.g., Almela et al., 2024), the aim of this study is to examine the performance of the last version of LIWC (LIWC-22) so as to identify its limitations and provide possible improvements. To proceed with the study, the free access corpus "Sentiment Polarity Dataset Version 2.0" created by Bo Pang and Lillian Lee was used. This corpus can be accessed through the site for movie-review data of the University of Cornell. The dataset includes two thousand movie reviews that were previously processed and labelled as negative or positive through an explicit star and numerical rating method, which provides a less subjective classification. This previous classification makes it possible to compare and verify the performance of the tool under examination for text analysis in terms of emotion detection. The focus of the study is on the "positive emotion" and "negative emotion" categories of the model, together with the categories that are included within them. The analysis obtained with LIWC is firstly compared in a more general scope to the classification done in the "Sentiment Polarity Dataset Version 2.0" corpus. Subsequently, it is thoroughly examined to provide a human-based analysis of the emotion conveyed in the text. In case of a wrong classification of the emotion by the model, the reasons behind the mistake are examined, in order to conclude where the problem is. The two main problems found in the preliminary analysis are the lack of a context-based analysis and the wrong classification of several emotions. The former results in emotion detection errors, especially in cases of word ambiguity, mixed emotions or irony. Moreover, the way LIWC classifies its categories also implies a problem in emotion identification, due to possible overgeneralization. These limitations are used to propose different possible improvements that could be implemented in future versions of LIWC, especially regarding the classification method and the decontextualized analysis conducted.

References

Almela, Á., Cantos-Gómez, P., Granados-Meroño, D., & Alcaraz-Mármol, G. (2024). LACELL at EmoSPeech-IberLEF2024: Combining Linguistic Features and Contextual Sentence Embeddings for Detecting Emotions from Audio Transcriptions. Proceedings of IberLEF 2024. Aachen: CEUR. https://ceur-ws.org/Vol-3756/EmoSPeech2024_paper4.pdf

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. University of Texas at Austin.

Gong, Y., Shin, K., & Poellabauer, C. (2018). *Improving LIWC using soft word matching*. https://doi.org/10.1145/3233547.3233632.

Kahn,J., Tobin, J., Massey, A., Anderson,J. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology*, 120(2), 263-286.

McDonnell, M., Owen, J. E., & Bantum, E. O. (2020). Identification of Emotional Expression With Cancer Survivors: Validation of Linguistic Inquiry and Word Count. *JMIR Formative Research*, 4(10). https://doi.org/10.2196/18246.

Yakut, I., Pan, S. (2022). Incorporating LIWC in Neural Networks to Improve Human Trait and Behavior Analysis in Low Resource Scenarios. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 4532–4539.

Detección automática de metáforas en francés: el caso del discurso político de Emmanuel Macron

Paula Pruaño Fuentes, Reyes Gago Sosa

Universidad de Cádiz

Según Lakoff y Johnson (2020), contamos con un sistema conceptual de naturaleza metafórica. Estos conceptos estructuran todo lo que pensamos, hacemos y percibimos. ¿Podría un programa detectar metáforas con la misma facilidad que nosotros? ¿Y si lo entrenamos?

La detección efectiva y automática de metáforas conceptuales supone un gran desafío en tareas de procesamiento del lenguaje natural debido a su subjetividad y abstracción. Se ha logrado un avance significativo en los últimos años en el abordaje de esta tarea gracias al desarrollo computacional, particularmente en el aprendizaje profundo. Sin embargo, los modelos de tareas de clasificación más recientes han sido generalmente entrenados con datos en inglés.

Nuestra hipótesis de partida es que un modelo multilingüe como XLM-RoBERTa ajustado para la detección de metáforas (Wachowiak et al., 2022) puede mostrar un desarrollo considerable en francés aunque no haya sido la lengua de entrenamiento. Además, este desempeño podría mejorar si el modelo se reentrena previamente con datos específicos en español como los del corpus CoMeta (Sánchez-Bayona y Agerri, 2022), debido a la similaridad lingüística entre ambas lenguas.

Para la elaboración de nuestro corpus nos hemos basado en los enunciados metafóricos pronunciados por Emmanuel Macron. Los discursos seleccionados pertenecen al comienzo del segundo quinquenio. Para ello, recurrimos a la página oficial de la Presidencia Francesa: *Élysée*. La anotación de nuestro corpus ha seguido las directrices de MIPVU (Steen et al., 2010) utilizadas en el corpus VUAM, que utiliza un etiquetado binario a nivel de token (B-METAPHOR/O).

Encontramos relevante nuestra investigación por dos motivos: la originalidad, ya que no existe ningún dataset en francés de acceso libre; y la agilización, puesto que, los modelos de IA pueden facilitar el trabajo manual.

En cuanto a la metodología experimental empleada para el ajuste fino del modelo, en primer lugar, cargamos el conjunto de datos CoMeta (Sánchez-Bayona y Agerri, 2022) desde la librería Datasets de Hugging Face (Wolf et al., 2019) y lo preprocesamos utilizando el tokenizador correspondiente al modelo *lwachowiak/Metaphor-Detection-XLMR* (Wachowiak et al., 2022).

Después, configuramos el entrenamiento del modelo con parámetros específicos de épocas, tasa de aprendizaje y tamaño de lote. Para la evaluación cuantitativa, utilizamos métricas estándar como recall o f-score. A continuación, evaluamos el modelo ajustado en el conjunto de datos de prueba de CoMeta, así como en nuestro corpus en francés. Esto nos permite como paso final llevar a cabo un análisis cualitativo del rendimiento del modelo en ambos conjuntos de datos.

Gracias a la metodología anterior, pretendemos obtener resultados similares en ambas lenguas, que ronden el 0.60 en las métricas estándar en el reconocimiento de la etiqueta B-METAPHOR. Somos conscientes que las diferencias de temáticas de los dataset y la claridad de los ejemplos puede influir significativamente en los resultados.

Aunque anticipamos un resultado aceptable, no creemos que el modelo pueda lograr reconocer la metáfora de manera excelente debido a la complejidad de la tarea y a su vinculación con el lenguaje figurado, por lo que seguirá siendo necesaria una revisión humana posterior.

Referencias

Lakoff, G., & Johnson, M. (2020). Metáforas de la vida cotidiana. Cátedra.

Macron, E. (2022, 13 de julio). *Discours aux armées à l'Hôtel de Brienne*. 13 de julio de 2022. Élysée. Recuperado el 16 de diciembre de 2024 de https://www.elysee.fr/emmanuel-macron/2022/07/13/discours-aux-armees-a-lhotel-de-brienne-1.

Macron, E. (2022, 15 de septiembre). *Discours du Président Emmanuel Macron aux Préfets*. Élysée. Recuperado el 16 de diciembre de 2024 de https://www.elysee.fr/emmanuel-macron/2022/09/15/discours-du-president-emmanuel-macron-aux-prefets.

Macron, E. (2022, 20 de septiembre). *Discours du Président de la République devant l'Assemblée générale de l'Organisation des Nations unies*. 20 de septiembre de 2022. Élysée. Recuperado el 16 de diciembre de 2024 de https://www.elysee.fr/emmanuel-macron/2022/09/20/discours-du-president-de-la-republique-devant-lassemblee-generale-de-lorganisation-des-nations-unies.

Macron, E. (2022, 21 de septiembre). *Discours du Président de la République à la septème conférence de reconstitution des ressources du Fonds mondial de lutte contre le sida, la tuberculose et le paludisme*. 21 de septiembre de 2022. Élysée. Recuperado el 16 de diciembre de 2024 de https://www.elysee.fr/emmanuel-macron/2022/09/21/discours-du-president-de-la-republique-a-la-septieme-conference-de-reconstitution-des-ressources-du-fonds-mondial-de-lutte-contre-le-sida-la-tuberculose-et-le-paludisme.

Macron, E. (2022, 30 de junio). *Discours du Président Emmanuel Macron à la Conférence des Nations unies sur les océans de Lisbonne*. 30 de junio de 2022. Élysée. Recuperado el 16 de diciembre de 2024 de https://www.elysee.fr/emmanuel-macron/2022/06/30/discours-du-president-emmanuel-macron-a-la-conference-des-nations-unies-sur-les-oceans-de-lisbonne.

Macron, E. (2022, 30 de noviembre). *Discours du Président de la République en l'honneur de la communauté française résidente aux États-Unis*. 30 de noviembre de 2022. Élysée. Recuperado el 16 de diciembre de 2024 de https://www.elysee.fr/emmanuel-macron/2022/11/30/discours-du-president-de-la-republique-en-lhonneur-de-la-communaute-francaise-residente-aux-etats-unis.

Macron, E. (2022, 30 de septiembre). Inauguration du Foirail à Pau par le Président Emmanuel Macron. 30 de septiembre de 2022. Élysée. Recuperado el 16 de diciembre de 2024 de https://www.elysee.fr/emmanuel-macron/2022/09/30/inauguration-du-foirail-a-pau-par-le-president-emmanuel-macron.

Macron, E. (2022, 9 de noviembre). *À Toulon, le Président présente la Revue nationale stratégique*. 9 de noviembre de 2022. Élysée. Recuperado el 16 de diciembre de 2024 de https://www.elysee.fr/emmanuel-macron/2022/11/09/a-toulon-le-president-de-la-republique-presente-la-revue-nationale-strategique.

Sanchez-Bayona, E., & Agerri, R. (2022). Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, 228–240. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Steen, G., Dorst, L., Herrmann, J., Kaal, A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*.

Wachowiak, L., Gromann, D., & Xu, C. (2022). Drum up support: Systematic analysis of image-schematic conceptual metaphors. *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, 44–53.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

On the polarisation after the DANA in Valencia: A corpus of news about Carlos Mazón and Teresa Ribera

Guillem Soler, Paolo Rosso

Universidad Politécnica de València

This study aims to analyse the media polarization in Spain following the devastating floods that took place in Valencia on October 29, 2024, which caused 224 deaths, extensive damage, and significant public concern. The research focuses on two key political figures, Carlos Mazón, leader of the Valencian Popular Party (PP) and President of the Generalitat Valenciana, and Teresa Ribera, Third Vice President of Spain and Minister for the Ecological Transition (PSOE). Both were central to the disaster response and quickly became subjects of intense media scrutiny. The study examines how newspapers with different political orientations framed these politicians and the events, exploring whether media bias influenced public perceptions during a crisis (López-Rico, González-Esteban & Hernández-Martínez, 2020).

The analysis is based on a corpus of news articles from two Spanish newspapers of opposite political leanings: *El País* and *El Confidencial*, collected over a 15-day period following the DANA event (October 29–November 13, 2024) (Razgovorov & Tomás Díaz, 2019). The dataset includes 20 articles about Teresa Ribera from *El País*, 23 from *El Confidencial*, 76 articles about Carlos Mazón from *El Confidencial*, and 81 articles about Mazón from *El País*. The main research questions focus on (i) whether newspapers with different political leanings adopt distinct journalistic approaches to the same events and politicians, and (ii) how they use emotional and tonal elements such as polarity of sentiment (positive, neutral, negative), clickbait techniques, and informational tones.

Articles were automatically extracted and pre-processed to standardise their format and prepare them for computational analysis. Sentiment and emotion analyses were conducted using pysentimiento's Artificial Intelligence models (Pérez et al., 2023) to measure the emotional and sentiment content of the texts. To evaluate tonal differences, a classification pipeline using Meta's Llama Large Language Model (Touvron et al., 2023) was applied to distinguish between literary, clickbait, and informational tones (Crespo-Martínez et al., 2024). Finally, discriminant analysis was employed to identify significant differences in how the two newspapers framed the same politicians and events.

Preliminary results show differences in the coverage of the two newspapers. *El Confidencial* highlights positive aspects of Carlos Mazón's leadership with a supportive tone, while *El País* adopts a critical stance, often using eye-catching headlines. In the case of Teresa Ribera, *El Confidencial* portrays her in a direct but somewhat negative way, with a serious tone, while *El País* mixes positive comments with occasional criticism, offering a more balanced view. These preliminary results illustrate how political trends influence information, adapting narratives to the audience's expectations (Castromil & Chavero, 2012).

These results are provisional, as the study represents the first phase of a broader project conducted within the framework of the FairTransNLP research initiative, funded by MCIN (AEI and ERDF/EU) under Grant PID2021-124361OB-C31. Future work will expand the corpus to include articles from additional Spanish newspapers, such as *elDiario.es* and *El Mundo* (Martínez & Serrano-Contreras, 2023). This expanded corpus will

offer a more comprehensive understanding of how media polarization operates, contributing to the study of variation in linguistic and narrative strategies within Spanish media.

References

Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alonso Alemany, L., & Martínez, M. V. (2023). *pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks* (Versión 1). Research Square. https://doi.org/10.21203/rs.3.rs-3570648/v1

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: *Open and efficient foundation language models*. arXiv. https://doi.org/10.48550/arXiv.2302.13971

Crespo-Martínez, I., Melero-López, I., Mora-Rodríguez, A., & Rojo-Martínez, J.-M. (2024). Política, uso de medios y polarización afectiva en España. *Revista Mediterránea De Comunicación*, 15(2), e26681. https://doi.org/10.14198/MEDCOM.26681

Razgovorov, P., & Tomás Díaz, D. (2019). Creación de un corpus de noticias de gran tamaño en español para el análisis diacrónico y diatópico del uso del lenguaje. *Procesamiento del Lenguaje Natural*, 62, 29–36. https://doi.org/10.26342/2019-62-3

López-Rico, C. M., González-Esteban, J. L., & Hernández-Martínez, A. (2020). Polarización y confianza en los medios españoles durante el Covid-19. Identificación de perfiles de audiencia. *Revista Española De Comunicación En Salud*, 77-89. https://doi.org/10.20318/recs.2020.5439

Castromil, A. R., & Chavero, P. (2012). Polarización política y negativismo mediático: Similitudes y diferencias en la prensa de derecha y la de izquierda en las elecciones autonómicas y municipales de 2011. *REDMARKA. Revista Digital de Marketing Aplicado*, 1(8), 55-81. https://doi.org/10.17979/redma.2012.01.08.4734

Martínez, R. L., & Serrano-Contreras, I. J. (2023). Ten years of immigration: Ten years of polarisation? A media analysis of the Andalusian case (2012–2022) through natural language processing. In M. Musiał-Karg & Ó. G. Luengo (Eds.), *Digital communication and populism in times of Covid-19* . Studies in Digital Politics and Governance. Springer. https://doi.org/10.1007/978-3-031-33716-1_4

"If this be Irish gratitude, I could wish myself a Frenchman": Exploring (im)politeness in intimate discourse in the context of historical letter writing through an analysis of reproaches in CORIECOR

David Sotoca Fernández

Universidad de Extremadura

This presentation performs an initial approach to historical (im)politeness strategies in the context of letter writing using corpus linguistics tools. It resorts to a sub-corpus of CORIECOR (Amador-Moreno 2022) which contains letters written by Irish immigrants that relocated to the US and their intimates (Clancy 2016) and contained a total of 595 letters. This research focuses on reproaches (Tulimirovic 2023; Albelda Marco 2023) and it applies Archer's model for (im)politeness studies (2017) to all instances of this speech act extracted from the data. This results in two different sub-corpora, containing what is here defined as "Face Enhancing Reproaches" (FER) and "Face Aggravating Reproaches" (FAR) respectively. FER contained a total of 84 reproaches that amounted to 5002 words.

Similarly, FAR comprised 74 reroaches, adding up to 4892 words. These two sub-corpora are contrasted through the SketchEngine software with a focus on FER. Keywords and concordance lists were used systematically to shed light on the most frequent words and expressions used to encode reproaches with an ambiguous face-enhancing intent. A keyword list was used first to retrieve a list of salient lemmas by assigning FER as the focus corpus and FAR as a reference corpus, with a focus on "rare" adjusted to a value of 0.00001, a minimum frequency of a value of 10, and no limit for maximum frequency. Also, both "keyword settings" and "N-Grams settings" were adjusted to "lemma" in order for the software to retrieve not only specific words but also all their possible variations. This way, Sketch Engine retrieved a list of lemmas with a frequency of at least 10 instances in FER that were predominant because of their absence in FAR. The most predominant lemmas in FER (still, before, what, your, account, home, uneasy, dear, like and wish) are analyzed and discussed in relation to their function for (im)politeness purposes. This research sheds light on the face-enhancing value of these linguistic items to encode this speech act within the specific context in question.

References

Albelda Marco, M. (2023). Rhetorical Questions as Reproaching Devices. *Journal of Language Aggression and Conflict* 11(2): 176–199

Amador-Moreno, C. P. (2022). Contact, Variation and Change: Mapping the History of Irish English through CORIECOR. *Nexus 22(2)*: 49–56.

Archer, D. (2017). (Im)politeness in Legal Settings. In J. Culpeper, M. Haugh and D. Z. Kádár (eds). *The Palgrave Handbook of Linguistic (Im)politeness* (pp. 713–737). London: Palgrave MacMillan.

Clancy, B. (2016). *Investigating Intimate Discourse*: *Exploring the spoken interaction of families, couples and friends*. London: Routledge.

Tulimirovic, B. (2023). Reproach as a core value: the analysis of the communicative potential of the routine formulae 'qué broma es esta' and 'de qué vas'. *Pragmalingüística* 31: 579–603.4

El uso de grafos de conocimiento a partir del corpus de *Don Quijote de la Mancha* para crear un sistema de QA basado en RAG

Yanco Amor Torterolo Orta, Sofía Roseti, Antonio Moreno Sandoval

Universidad Autónoma de Madrid y UNED / Universidad Autónoma de Madrid / Universidad Autónoma de Madrid

El propósito del presente trabajo académico es emplear un corpus compilado a partir del *magnus opus* de Miguel de Cervantes Saavedra, *Don Quijote de la Mancha*, para crear un *retriever* (recuperador de información) que sea capaz de recuperar el contexto adecuado que contenga la respuesta a la pregunta planteada. Este recuperador le proporcionará esta información a un LLM (*Large Language Model*) para generar la respuesta a dicha pregunta. A esto se lo conoce como RAG (Retrieval-Augmented Generation). Dado que el modelo emplea la información recuperada para responder a las preguntas planteadas, esta técnica trata de impedir que el modelo se invente las respuestas, lo que se denomina *alucinaciones*, y mejorar la explicabilidad del resultado.

En cuanto a metodología, el primer paso consiste en compilar el corpus. Se parte del libro en formato TXT. La información contenida en la obra se convierte a CSV o EXCEL de una manera semiautomatizada. Se registra información estructural, la cual se centra en los

párrafos. Cada párrafo contiene el número del párrafo, de capítulo y un identificador. Por otro lado, se registra información narrativa, no solo estructural, lo cual es un proceso más laborioso y lingüístico. Esta información incluye qué personajes intervienen, el sexo y clase social de estos, el turno de diálogo, etc. Tras ello, esta información se vuelca en una base de datos Neo4j. Se trata de una base de datos basada en grafos, que contiene grafos de conocimiento (*knowledge graphs*). Gira en torno a los nodos y relaciones, llamadas aristas, que existen entre ellos. Los nodos normalmente son entidades, como personajes o lugares, pero también pueden ser párrafos, oraciones, *chunks*, etc., lo cual permite estructurar la información y emplear representaciones vectoriales (*embeddings*) para las búsquedas por similitud semántica.

Una vez cubierta la parte del corpus y contando con la base de datos, los siguientes pasos son aplicar diversas técnicas de recuperación de información, como la vectorización, el uso de la información adicional de los nodos y las aristas (conocida como *metadatos*), búsquedas híbridas con keywords, etc. Se tratará de innovar y aportar al estado de la cuestión probando diversas técnicas actuales.

Cabe destacar que la compilación del corpus y su conversión al formato EXCEL ya se realizó en un trabajo previo, y la conversión a Neo4j ya se ha realizado para la presente comunicación. Los siguientes pasos descritos se elaborarán en lo sucesivo, pero partiendo de la experiencia de trabajos anteriores centrados en RAG. Como hipótesis, y con base en experiencias previas, se puede anticipar que al sistema le costará responder a preguntas cuya respuesta sea implícita porque no podrá recuperar correctamente el contexto. Las respuestas que necesiten múltiples elementos a lo largo de distintos contextos también tenderán a ser problemáticos. Estos son desafíos que se intentarán solventar a largo plazo. En definitiva, el resultado será un sistema que responda preguntas sobre *El Quijote*, y se prevé mejorar los resultados de trabajos anteriores.

Referencias

Cervantes Saavedra, M., Sevilla Arroyo, F., y Rey Hazas, A. (2001). *Don Quijote de la Mancha* (1ª ed.). Alianza Editorial.

Chaudhri, V., et al. (2022). Knowledge graphs: Introduction, history and, perspectives. *AI Magazine*, 43(1), 17–29. https://doi.org/10.1002/aaai.12033

Ermer, M. (2023). *Visualizing literary narratives with a graph-centered approach*. https://knowledge.e.southern.edu/crd/2023/humanities/3/

Hou, Y., Zhang, R. (2024). *Enhancing Dietary Supplement Question Answer via Retrieval-Augmented Generation (RAG) with LLM.* https://www.medrxiv.org/content/10.1101/2024.09.11.24313513v1

Poliakov, M. y Shvai, N. (2024). *Multi-Meta-RAG: Improving RAG for Multi-Hop Queries using Database Filtering with LLM-Extracted Metadata.* https://doi.org/10.48550/arXiv.2406.13213

Yang, L., et al. (2024). Give us the Facts: Enhancing Large Language Models with Knowledge Graphs for Fact-Aware Language Modeling. *IEEE Transactions on Knowledge and Data Engineering*, 36(7), 3091–3110. https://doi.org/10.1109/TKDE.2024.3360454

Panel 8

Corpora, language acquisition and teaching
Corpus, adquisición y enseñanza de lenguas

Tagging translanguaging in a CLIL-Bio learner corpus

Marlén Izquierdo, Yolanda Ruiz de Zarobe

Universidad del País Vasco, UPV/EHU

Translanguaging, which is understood as both multilingual behaviour and a multilingual pedagogical approach, has long been the object of study in CLIL research, this being mostly "descriptive and focused on processes" (Moore, 2023, p. 38). The aim of this study is, however, to approach translanguaging not as a process but rather as the outcome of foreign language teaching-learning. To this end, we aimed to compile a learner corpus with texts produced in a CLIL classroom within a multilingual setting (the Basque Country), where English is the medium of instruction and the learners are native speakers of Spanish and Basque. Our research questions are: 1) how does the L1-input impact on the learners' L3 output? Do they come up with the expected target language or do they rely on translanguaging to complete the task? And 2) if translanguaging occurs, at what level? and is such translanguaging content-specific or rather, representative of general English use?

Ninety secondary-education learners of biology and geology in L3 English were asked to write two different texts about two environmental issues, namely, air pollution and climate change. Benefitting from pedagogic translanguaging, we designed two tasks with diagrams and prompts deliberately planned not only in English, but also in Spanish and Basque, on the assumption that the learners' L1 would help to make L3 input comprehensible. In one of the tasks they had to explain the conceptual relations in a diagram and in the other they had to expose previously known scientific knowledge. Nothing was elicited in particular so that the learners could build their discourse as naturally as possible.

As a result, we compiled CLIL-Bio EL3 Corpus, a learner corpus of 180 texts, enriched with the learners' metadata and tagged for translanguaging. Tagging was done manually and encoded to enable searches in AntConc 4.3.1 (Anthony 2024). The tagging procedure reveals nearly 200 instances of translanguaging at three different levels, mainly lexical, but also syntactical and morphological. It was also found that the L1 most learners alternate English with was Spanish, which, according to the learners' metadata is the majority's L1, with Basque having an L2 role. The implications of the study are not only pedagogical but also methodological in that it might open a research avenue where CLIL and SLA finally meet Corpus Linguistics.

References

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc

Moore, P. (2023). Translanguaging in CLIL. In D. L. Banegas & S. Zappa-Hollman (Eds). *The Routledge Handbook of Content and Language Integrated Learning* (pp. 28-42). London: Routledge.

Proficiency level and the expression of exemplification in EFL opinion essays: A cross-sectional learner corpus study

Carmen Maíz-Arévalo, Belén Díez-Bedmar

Universidad Complutense de Madrid, Universidad de Jaén

The use of exemplification in academic writing is a key skill to "empower clarity and argumentation" (Triki, 2024: np). L2 writers, therefore, are expected to master exemplification, especially in specific genres such as the argumentative essay. Exemplification is, in fact, one of the descriptors in the Common European Framework of Reference (2001) and its Companion Volume (2020) as part of pragmatic competence (textual competence). Most research in the field, however, has focused on the performance of exemplification by expert writers (Hyland, 2007; Triki, 2024) or on the contrast between novel and expert writers, specifically comparing the written production of MA and PhD EFL students with that of renowned authors (Guziurová, 2022; Su et al. 2022; Su and Lu, 2022; Su and Ye, 2023; Vijayakumar, 2024, among others). Spanish EFL students have received less attention (cf. Mur-Dueñas, 2011; Paquot, 2008). This paper aims to redress this imbalance by conducting a learner corpus study with students' written production by L1 Spanish students at a proficient level (C1 level) as well as those at two intermediate levels (B1 & B2 levels). More specifically, we intend to answer the following research questions: (i) What variety can be observed in the use of exemplification linguistic patterns depending on the students' level (B1, B2, C1)? And (ii) How frequently do these students employ such patterns to express exemplification depending on their level? In line with previous research (Mur-Dueñas, 2011), it is hypothesized that students at higher levels will deploy a wider variety of linguistic patterns to express exemplification and use exemplification to a more frequent extent than lower-level ones. To this end, a total of 104,629 words consisting of 511 opinion essays written by L1 Spanish EFL learners belonging to three different levels (B1, B2 and C1) were analysed. The dataset (part of the FineDesc Learner Corpus) is a convenience sample consisting of 187 texts belonging to B1, 182 to B2 and 142 to C1. Following a mixed-methods approach, the expression of exemplification was coded and then statistically tested with SPSS. Among other things, results show that, as expected, students both at B1 and B2 level resort to a limited number of expressions (i.e., "for example", "for instance" and "such as") while higher level students display a wider range of strategies to introduce examples than the other two groups. Interestingly, however, they still fail to use common expressions like "namely" or "to give an example", which are rather frequent among native writers.

References

Guziurová, T. (2022). Glossing an argument: Reformulation and exemplification in L2 Master's theses. *Topics in Linguistics*, *23*(2), 18–35.

Hyland, K. (2007). Applying a Gloss: Exemplifying and reformulating in academic discourse. *Applied Linguistics*, *28*(2), 266–285.

Mur-Dueñas, P. (2011). An intercultural analysis of metadiscourse features in research articles written in English and in Spanish. *Journal of Pragmatics*, *43*(12), 3068-3079.

Paquot, M. (2008). Exemplification in learner writing. In *Phraseology in Foreign Language Learning and Teaching* (pp. 101–119). John Benjamins.

Su, H., & Lu, X. (2022). Assessing pragmatic performance in advanced L2 academic writing through the lens of local grammars: A case study of 'exemplification.' *Assessing Writing*, *54*, 100668.

Su, H., & Ye, J. (2024). Local grammar approach to investigating advanced Chinese EFL learners' development of communicative competence in academic writing: The case of 'exemplification.' *Corpus-based Studies Across Humanities*, *1*(1), 157–181.

Su, H., Zhang, Y., & Chau, M. H. (2022). Exemplification in Chinese English-major MA students' and expert writers' academic writing: A local grammar-based investigation. *Journal of English for Academic Purposes*, *58*, 101120.
Triki, N. (2024). Exemplification and reformulation in expert linguists' writings: Elaborative metadiscourse between disciplinarity and individuality. *Journal of English for Academic Purposes*, *71*, 101407.
Vijayakumar, C. (2024). Exemplification in student essay writing: A study of learner corpus of essay writing (LCEW). *International Journal of Applied Linguistics*, *34*(1), 1514–1532.

Multi-word verbs in the texts of ELT coursebooks: A contrastive analysis with large scale L1 English corpus studies

Elaine Millar

Universidad de Cantabria

Second language acquisition research has shown that learners benefit from frequent exposure to multi-word units, with each encounter in contextualised input helping them to accumulate and consolidate their knowledge (Siyanova-Chanturia. Conklin & van Heuven, 2011; Schmitt & Schmitt, 2020). While the EFL classroom is an important source for such input, there are doubts whether the language in these settings sufficiently resembles that of 'real world' discourse (Pérez-Paredes, Mark & O'Keeffe., 2020). This paper explores the extent to which this may be an issue concerning contextualised exposure to multi-word verbs, by examining content in intermediate to advanced level ELT coursebooks. Two research questions are asked in the study: 1) How often will learners encounter multi-word verbs in the reading passages and listening transcripts, and to what extent is this frequency similar to L1 English discourse? and 2) Which multi-word verb lemmas and senses will they encounter, and to what extent are these similar to L1 English discourse? A total of 12 CEFR B1-C1 titles from four international ELT publishers were selected for analysis. Using #LancsBox 6.0 corpus software, verb-particle and verb-preposition constructions were first identified in the texts via simple and complex part-of-speech tag searches. Then, they were contrastively analysed with data from largescale L1 English corpus studies (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liu, 2011). The analysis revealed that, overall, the multi-word verbs identified in the texts were comparable with L1 discourse in terms of their relative, lemma-token, and sense-based frequencies. The frequencies were found to increase steadily as the coursebooks' target proficiency levels rose, and a higher proportion of items were identified in the listening transcripts than in the reading passages. This suggests the elaboration of texts for the materials involved a degree of purposeful omission and/or addition of items for pedagogical purposes. However, the very nature of EFL coursebooks, which consist mainly of short texts, means that this input alone is unlikely to foster incidental multi-word verb acquisition. Nevertheless, these findings show that the reading passages and listening transcripts of these popular coursebooks can serve as a reliable starting point for focused multi-word verb instruction exploring language in context.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2004). *Longman Grammar of Spoken and Written English* (4th ed.). Pearson Education Limited.
Gardner, D., & Davies, M. (2007). Pointing out Frequent Phrasal Verbs: A Corpus-Based Analysis. *Source: TESOL Quarterly*, *41*(2), 339–359. https://doi.org/10.1002/j.1545-7249.2007.tb00062.x

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A Pedagogical List of Phrasal Verbs and their Most Frequent Meaning Senses. *Language Teaching Research*, *19*(6), 645–666. https://doi.org/10.1177/1362168814559798

Liu, D. (2011). The Most Frequently Used English Phrasal Verbs in American and British English: A Multicorpus Examination. *TESOL Quarterly*, *45*, 661–688. https://doi.org/10.2307/41307661

Pérez-Paredes, P., Mark, G., & O'Keeffe, A. (2020). *The impact of usage- based approaches on second language learning and teaching*. Cambridge Education Research Reports.

Schmitt, N. & Schmitt, D. (2020). *Vocabulary in Language Teaching* (2nd ed.). Cambridge University Press

Siyanova- Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). Seeing a phrase 'time and again' matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 776–784. https://doi.org/10.1037/a0022531

Developmental patterns in lexical richness of university students of English as a Foreign Language

Pauline Moore

Universidad Autónoma del Estado de México

An important area of language learner development is the enrichment and growth of vocabulary in the second language. There are several available measures to determine vocabulary size like the Vocabulary Knowledge Scale (Paribakht and Wesche, 1997), the Vocabulary Size Test (Nation & Beglar, 2007), the Productive Levels Test (Laufer & Nation, 1999), among others. However, these measures generally measure knowledge of a word as the ability to identify or produce synonyms or definitions of the item which evaluates vocabulary learning as an isolated and context-independent skill. However, this discrete knowledge of words does not always mean that a learner can access and use a word under the demands of interaction. Corpus studies provide us with access to samples of learner oral production that offer a broader picture of productive vocabulary knowledge and use under natural conditions of communication in a second language. Our research question was: How do lexical diversity and lexical sophistication vary across time in university students of English as a Foreign Language?

We will present data from MexLeC corpus, a longitudinal spoken learner corpus, calculating the lexical diversity and lexical sophistication of students' oral production across three years of language instruction and examining patterns of development in the acquisition of lexis. We used TAALED (Kyle, Crossley and Jarvis, 2021) to calculate MATTR and MTLD as measures of lexical diversity which are not affected by variation in text length while learner proficiency improves. For lexical sophistication, several measures are presented from an analysis using TAALES (Kyle & Crossley, 2015; Kyle, Crossley & Berger, 2018). These include linguistic considerations of word frequency both for individual words and for n-grams to show development in the use of larger lexical units; semantic aspects like polysemy and hypernymy; and psycholinguistic measures, like concreteness, familiarity and age of acquisition.

Our preliminary results suggest that lexical diversity improves at a faster rate than lexical sophistication and that initial gains in lexical diversity are more significant for function words than for content words. As well as revealing important patterns in learner language development these results can provide benchmark values for the field of automated language assessment. The results also offer important feedback for the teaching process and can be used to inform the design of teaching materials for English as a Foreign Language.

References

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), pp. 757-786. doi: 10.1002/tesq.194.

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030-1046. https://doi.org/10.3758/s13428-017-0924-4.

Kyle, K., Crossley, S. A., and Jarvis, S. (2021). Assessing the validity of lexical diversity using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322. doi:10.1093/applin/16.3.307

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *Language Teacher*, *31*(7), 9–13.

Paribakht, T. S. and Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady and T. Huckin (Eds.) *Second Language Vocabulary Acquisition* (pp. 174-200). Cambridge: Cambridge University Press.

L1 Influence on L2: Teaching impersonal and personal passives to address learner

Emma Nemishalyan

Universidad Francesa en Armenia

Research in second language acquisition (SLA) and learner corpus studies has increasingly employed corpus-based methods to examine grammatical patterns in second language (L2) learning. While extensive research has explored linguistic transfer in English acquisition among speakers of Chinese, Spanish, and German, little attention has been given to Armenian learners. This study addresses this gap by investigating the use of personal and impersonal passive constructions among Armenian learners of English, with a focus on the role of linguistic transfer and broader cognitive and pedagogical influences. Given that Armenian lacks direct equivalents for these passive structures, this research critically examines whether this linguistic difference leads to systematic underuse or avoidance of English passives and how cognitive constraints and instructional practices further shape these patterns.

Drawing on linguistic transfer theory and Contrastive Interlanguage Analysis (CIA), this study investigates passive construction usage among Armenian learners relative to native English speakers. A specialized learner corpus was compiled following the International Corpus of Learner English (ICLE) standards, consisting of argumentative essays written by upper-intermediate to advanced Armenian learners in untimed settings with access to reference tools. To ensure reliable comparisons, the Louvain Corpus of Native English Essays (LOCNESS) was used as a reference corpus. While this study focuses on argumentative writing, we acknowledge that register and task type may influence passive usage patterns, necessitating further research across different genres.

Through both quantitative and qualitative analysis, findings reveal a significant underuse of impersonal passives and an avoidance of personal passives. While the absence of direct syntactic equivalents in Armenian plays a role, this study further identifies cognitive processing constraints—such as complexity avoidance—and pedagogical factors, including insufficient exposure to passives in instructional materials, as contributing to these trends. These findings align with prior research on passive avoidance in other L1 backgrounds but suggest unique challenges for Armenian learners due to typological differences between Armenian and English.

The results carry important pedagogical implications for English language teaching, emphasizing the need for targeted instruction that explicitly integrates passive constructions into curricula for Armenian learners. By bridging theoretical insights on linguistic transfer with practical teaching approaches, this research offers a more nuanced understanding of passive voice acquisition in SLA and proposes strategies to facilitate the effective integration of complex grammatical structures in L2 learning.

An empirical approach to analysing language and content teachers' perceptions of student productions

Michael O'Donnell, Alex Hope

Universidad Autónoma de Madrid

Many high schools in Spain are working with a CLIL approach, integrating content and language teaching, e.g., teaching Biology in English as a means for the student to develop both language and content skills at the same time. In some contexts, writing produced by students in these integrated classrooms may be assessed separately by content and language teachers. It is not clear if these teachers restrict themselves to their specialisation, or whether content teachers take into account language issues, and language teachers take into account content issues. This study attempts to explore this issue, assessing to what degree a set of high school biology teachers draw upon content vs language criteria in their assessment, and similarly for the language teachers in these classrooms. As part of a larger project, responses to biology prompts were collects from students in 1st and 4th year from various high schools across Spain. 25 responses from each year were randomly selected, and then content and language teachers ranked the responses using No More Marking, a comparative judgement tool in which teachers chose the texts they prefer from a sequence of paired texts (Jones, 2016). Separate rankings were produced for content and language teachers. The software then converted these rankings to a scaled score from 0 to 100. Texts were also processed within UAM Corpustool (O'Donnell, 2008), in particular, to produce statistics for each response in relation to a number of variables:

- Fluency, simplistically measured in terms of number of words in the response.
- Accuracy, measured in terms of language errors per 1000 words. All texts were manually annotated for errors, in terms of errors in vocabulary, syntax, phrasing, and register.
- Complexity, measured in a range of different ways, including number of elements per noun phrase, number of elements per verb phrase, use of marked structures, etc.
- Content appropriateness, measured in degree of use of the key vocabulary of the subject, and separately, percent of key vocabulary that appear in the response.

We assumed that the more important the feature was for the teacher-type's assessment, the stronger the correlation would be between that feature and the ranking of responses. The paper presents these findings, showing that, as expected, content teachers pay more attention to content issues, and language teachers to language-related factors. However, the results also show that each teacher-type also seems to pay attention to the factors more relevant to the other teacher-type, although to different degrees.

References

Jones, N. (2016). 'No More Marking': An online tool for comparative judgement. *Cambridge English: Research Notes*, *63*, 12-15.

O'Donnell, M. 2008. The UAM CorpusTool: Software for corpus annotation and exploration". In C. M. Bretones Callejas et al. (Eds.), *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente* (pp. 1433-1447). Almería: Universidad de Almería.

A speech corpus as a research resource to diagnose L2 English pronunciation and set teaching priorities: Introducing USAL-EP&P

Jorge Andrés Prieto Prat

Universidad de Salamanca

This paper presents the *USAL-EP&P* spoken corpus, a project that could turn a compilation of 1st-year university students' SFS-WASP recorded files into a resource for qualitative and quantitative analysis of speech accuracy, fluency and other sound aspects of L2 English learning / acquisition.

Non-native English speech corpora are far less common than written or native spoken ones (Raab et al. 2008), and their processing requires painstaking tagging before they can be processed by ASR systems for different purposes, like EFL pronunciation assessment (e.g. Chen 2023). Fewer still are those transcribed phonetically. Once given shape as such, the *USAL-EP&P* corpus would include statistics and visual interfaces, potentially suitable for machine reading. However, narrow transcription of the files and machine browsability are not in the scope of this project. Besides, legal and technical aspects must be sorted out first.

There were practically no precedents of a spoken corpus-based part in the examination process of English Phonetics and Phonology in Spanish universities, at least until 2012, when recordings started in the degree of English Studies at the USAL. After a pilot stage, students began to record readings of short scripts that year as a half-optional, conditional activity to have their L2 English pronunciation diagnosed, not disregarding fluency and intonation, which was a first for many.

Quite unintentionally, audio files with their spectrograms and marked scripts stockpiled for over ten years, reaching about 1700 when the activity ceased in 2023. Should all that information be shelved away, or disposed of like written exams? Or could that raw material become a researchable spoken corpus instead? If so, it must be defined in its attributes, analyzed, organized, broken down statistically, and then put to work.

Over a 10-year, 8-course period, hundreds of future teachers received detailed feedback on their L2 English pronunciation, gaining awareness of their key role. What were their shortcomings and strengths, and how could they improve? Is negative transfer from L1 to blame for all L2 deficiencies? What errors must be addressed in transitional pronunciation? Where is the borderline between accent colouring and plain inaccuracy? Can ASR testing outperform human raters? (Wester et al. 2001).

This project will try to answer queries like those by setting its theoretical foundations, supported by authorities, both present (Collins & Mees 2013 for one) and traditional, to continue with a contrastive analysis of segmental conflicts in the shift from Spanish to English (Mott 2011, Gómez González & Sánchez Roura 2016, etc.), illustrating them with pronunciation challenges in the corpus scripts. Next, there will be a description of the procedure followed to record and assess the audio clips that will eventually make up the spoken corpus. Finally, there should be some examples of how to browse this corpus and analyze features of interest.

The conclusions aim at revising conventions and misconceptions affecting teaching priorities, spoken models for EFL/ESL/EIL (Jenkins 2000, Lindsey 2016), and phonemic transcription.

Once this unsought compilation reaches corpus status, it should remain as a teaching tool and an object of research at the University's discretion.

References

Chen, H. C. (2023). What a Spoken Learner Corpus Tells Us: Construction and Application of a Pronunciation Programme for English-Language Teachers. *The Southeast Asian Journal of English Language Studies*, *29*(2).

Collins, B. and Mees, I. (2013). *Practical Phonetics and Phonology*. Abingdon: Routledge.

Gómez González, M.A., & Sánchez Roura, T. (2016). *English Pronunciation for Speakers of Spanish. From Theory to Practice*. Boston / Berlin: De Gruyter Mouton.

Jenkins, J. (2000). *The Phonology of English as an International Language*. Oxford: Oxford University Press.

Lindsey, G. (2019). *English after RP. Standard British Pronunciation Today*. Cham, Switzerland: Palgrave Macmillan.

Monroy Casas, R. (2024). *British English Pronunciation for Spanish-Speaking University Students*. Almería: Editorial Círculo Rojo.

Mott, B. (2011). *English Phonetics and Phonology for Spanish Speakers*. Barcelona: Publicacions i Edicions Universitat de Barcelona.

Raab, M., Gruhn, R., & Noeth, E. (2008). Non-Native Speech Databases. Conference paper. DOI: 10.1109/ASRU.2007.4430148. Source: IEEE Xplore. *Automatic Speech Recognition & Understanding*, 2007. ASRU.

Wester, M., Kessens, J., Cucchiarini, C., & Strik, H. (2001). Obtaining Phonetic Transcriptions: A Comparison between Expert Listeners and a Continuous Speech Recognizer. *Language and Speech* 2001, *44*(3), 377-403.

Innovación en el proceso de enseñanza y aprendizaje del alemán como lengua extranjera: Corpus PaGeS como puente entre lenguas

Mar Soliño Pazó

Universidad de Salamanca

Corpus PaGeS es una herramienta esencial para analizar patrones lingüísticos complejos del alemán, como las estructuras subordinadas, partículas modales y verbos separables, que suelen ser difíciles para hispanohablantes debido a las diferencias tipológicas entre ambos idiomas. Además, facilita su comprensión mediante contextos reales, lo que enriquece el aprendizaje lingüístico y pragmático. Destaca por su utilidad en identificar patrones lingüísticos recurrentes que permiten analizar fenómenos complejos del alemán, como las estructuras subordinadas, las partículas modales y los verbos separables, aspectos que suelen ser desafiantes para hablantes de español debido a las diferencias tipológicas entre ambas lenguas (Klein, 2013; García Mayo & Fernández, 2020). Desde el ámbito didáctico, el corpus ofrece ejemplos auténticos que reflejan usos discursivos genuinos del alemán, permitiendo a los docentes abordar fenómenos como la modalidad epistémica, deóntica o dinámica, esenciales para desarrollar competencias comunicativas interculturales. Partículas como *doch*, *ja*, *wohl*, y *eben* permiten expresar matices específicos y la actitud del hablante, aspectos clave para desarrollar la competencia comunicativa intercultural. Estas partículas, sin equivalentes directos en español, presentan retos en los procesos de aprendizaje y traducción, donde pueden ser omitidas o reinterpretadas.

Las tres preguntas clave de esta investigación son:

1. ¿Cómo pueden los fenómenos lingüísticos contrastivos entre alemán y español, identificados en el corpus PaGeS, contribuir al diseño de materiales didácticos eficaces?
2. ¿Cómo influye la integración del corpus en el aula en el desarrollo de estas competencias en hablantes de español que aprenden alemán?
3. ¿Qué patrones de uso, frecuencia y contexto emergen en partículas modales, estructuras subordinadas y verbos separables del alemán, y cómo pueden ser tratados didácticamente?

La hipótesis plantea que los corpus paralelos facilitan el aprendizaje de estructuras complejas en alemán como lengua extranjera (ALE) y que el análisis de fenómenos pragmáticos, como la modalidad, impulsa el desarrollo de competencias comunicativas e interculturales.

Este trabajo emplea una metodología mixta. El análisis cuantitativo utiliza herramientas digitales como AntConc y Sketch Engine para identificar patrones de frecuencia y distribución de fenómenos lingüísticos clave. Paralelamente, se realiza un análisis cualitativo para explorar ejemplos concretos de fenómenos pragmáticos como las partículas modales, y evaluar su transferencia al aula a través de actividades basadas en PaGeS.

Se anticipa que el uso de PaGeS en el aula de ALE mejora significativamente las competencias pragmáticas y estructurales de los aprendientes hispanohablantes, esperando que los estudiantes desarrollen una mayor habilidad para usar partículas modales y/o estructuras complejas, lo que impulsa una interacción más natural y efectiva en alemán. Igualmente, los materiales didácticos diseñados a partir del corpus fomentan un aprendizaje contextualizado y comunicativo, promoviendo tanto la competencia lingüística como la intercultural. El análisis contrastivo apoyado en este corpus permite diseñar materiales y actividades didácticas auténticas que destacan la expresión de la modalidad y otros aspectos pragmáticos esenciales. Este enfoque fomenta una enseñanza significativa del alemán, promoviendo un aprendizaje profundo y contextualizado.

Los resultados preliminares de las intervenciones didácticas reflejan un impacto positivo en el aprendizaje y ofrecen valiosas perspectivas para la investigación futura en lingüística contrastiva y enseñanza de lenguas extranjeras.

Referencias

Bybee, J., Perkins, R., Pagliuca, W. (1994). *The evolution of Grammar. Tense, Aspects, and Modality in the languages of the world*. The University of Chicago Press.

Doval Reixa, I. (2018). Das PaGeS-Korpus, ein Parallelkorpus der deutschen und spanischen Gegenwartssprache. *Revista de Filología Alemana*, *26*, 181-197. http://dx.doi.org/10.5209/RFAL.60148

Ellis, R. (2015). *Understanding Second Language Acquisition*. Oxford University Press.

García Mayo, M. P., & Fernández, J. (2020). *Studies in Bilingualism and Second Language Acquisition*. John Benjamins.

Klein, W. (2013). *Grammar in Dialogue: The German Modal Particles*. De Gruyter.

Martos Ramos, J.J. (2001), Una aproximación temporal a las partículas modales alemanas. *Philologia Hispalensis*, *15*(2), 169-178. https://doi.org/10.12795/PH.2001.v15.i02.11

Métrich, R. (1998). *Wie übersetzt man eigentlich Partikeln*. In W. Börner and K. Vogel (Eds.), *Kontrast und Äquivalenz* (pp. 194-207). Narr Verlag.

O`Sullivan, E. & Rössler, D. (1989). Wie kommen Abtönungspartikeln in deutsche Übersetzungen von Texten, deren Ausgangssprachen für diese keine direkten Äquivalente haben. En H. Weyt (Ed.), *Sprechen mit Partikeln* (pp. 204-216). De Gruyter.

Soliño Pazó, M.M. (2022). El manejo de un corpus paralelo en el aula de alemán como lengua extranjera. Descubrir PaGeS. En F. Robles & K. Siebold (Eds.), *El español y el alemán en contraste y sus implicaciones didácticas. Nuevas aportaciones desde la gramática, la traducción y la lingüística de corpus* (pp. 247-262). Narr Francke Attempto Verlag.

Towards a manageable corpus for an English-for-publication-purposes course

Julia Williams Camus

Universidad de Cantabria

The use of corpora in English for academic purposes (EAP) research is well-established and the technological advances seen in recent decades have made it possible to analyse large corpora to study the structure and lexico-grammatical patterns of research articles. While the growing availability of data enhances the reliability of research findings, there remains a challenge on how to transfer these results from EAP research to practice. In addition, for certain less researched disciplines information about writing conventions may not be available (Swales, 2019).

A common problem among practitioners is the lack of corpus accessibility for the development of course materials, which could be partially solved if EAP teachers resort to the creation of *ad-hoc* corpora for their courses. This would be particularly desirable in English-for-publication-purposes courses targeted at doctoral or researchers from the same discipline (Charles, 2018). However, given that corpus compilation and analysis can be time-consuming, the question remains as to what size is sufficient for the practitioner to gain insight into the specific disciplinary characteristics to develop suitable teaching materials.

This study presents an analysis of self-reference in two corpora of research articles from the field of agricultural sciences, Agri10 (n=10; 67,026 running words) and Agri30 (n=30; 202,193 running words), to determine whether robust findings can be obtained from the smaller sample. The analysis is based on Tang and John's (1999) self-reference taxonomy of six writer-identity roles arranged in increasing authorial power and risk to others in the scientific community, namely: *representative*, *guide*, *architect*, *recounter*, *opinion-holder* and *originator*. The difference in the overall presence of self-reference between the two corpora was 2.9%. However, the normalised figure for *we* was 10.9% higher in the larger corpus whereas *our* was 10.8% lower. The 21 instances of *us* in Agri30 compared to 5 in Agri10 represented a normalised 39.2% increase.

The results for the individual sections varied with differences of less than 10% for Results (7.9%), Discussion (3.1%) and Conclusion (5.6%), around 10% for Abstract (10.5%) and Introduction (10.8%), but as high as 18.7% for Methods. The figures were higher in Agri10 than in Agri30 for Abstract, Introduction, Results and Discussion, and lower for Methods and Conclusion. Similarly, the results for the different identity roles varied with differences below 10% for *opinion-holder* (9.9%) and *originator* (8.8%), over 10% for *representative* (13.8%) and *recounter* (14.4%), and of 39.2% for *guide*, which had the lowest number of tokens (n=11) of the roles. Again, the figures tended to be higher in the smaller Agri10 corpus, with only the figure for *recounter* being higher in Agri30.

These results suggest that ten articles could be sufficient to provide reasonably solid data on which to base the preparation of materials on a relatively frequent feature of academic discourse such as first-person reference. The number also is also small enough for the analysis not to take up an excessive amount of time.

References

Charles, M. (2018). Using do-it-yourself corpora in EAP: A tailor-made resource for teachers and students. *The Journal of Teaching English for Specific and Academic Purposes*, *6*(2), 217-224.

Swales, J. M. (2019). The futures of EAP genre studies: A personal viewpoint. *Journal of English for Academic Purposes*, *35*, 75-82.

Tang, R., & John, S. (1999). The 'I' in identity: Exploring writer identity in student academic writing through the first person pronoun. *English for Specific Purposes*, *18*, S23-S39.

Panel 9

Special uses of corpus linguistics
Usos y aplicaciones específicas de la lingüística de corpus

Modelling and annotating specialised attributes in humanitarian discourse: A pilot study

Santiago Chambó, Pilar León Araúz

Universidad de Granada

The Humanitarian Encyclopedia (HE; humanitarianencyclopedia.org) combines expert knowledge from entry authors with corpus-based systematic conceptual analyses provided by a team of linguists. Corpus-based conceptual analysis aims to 1) mitigate possible biases and content gaps in entries due to the diverse backgrounds of entry authors, and 2) to detect conceptual variation (Author, 2017; Hampton, 2020) among humanitarian organisations by leveraging corpus metadata.

At the HE, concepts are analysed using Frame-based Terminology (FBT; Faber, 2015, 2022), an approach inspired by Frame Semantics (Fillmore, 2008), which models meaning into semantic triples from knowledge rich contexts (KRCs; Condamines, 2022) extracted from domain-specific corpora. A KRC is a passage that contains textual evidence of conceptual description for at least one semantic relation. For example, the conceptual proposition SLEEPING MAT *type_of* NON-FOOD ITEM was abstracted from the following KRC: "Hundreds of thousands of displaced people are in evacuation centres or makeshift shelters, lacking essential non-food items (NFIs) such as sleeping mats and hygiene kits" (OCHA, 2024). FBT provides well-founded guidelines to describe entities like NON-FOOD ITEM and processes like DISPLACEMENT, but it remains vague about how to describe attributes systematically (Author, 2020). Examples of key humanitarian attributes include AID DEPENDENCE, IMPACT, NEUTRALITY and VULNERABILITY.

Attributes, also referred to as qualities or properties, constitute bundles of values ascribed to entities and processes, which can be represented with binary, discrete and continuous scalar structures (Nirenburg & Raskin, 2004; Oshima et al., 2019). COLOUR is a prototypical attribute (Hansen & Chemla, 2017) whose values can be represented lexically by adjectives (e.g., red) or numerically (e.g., 625-740 nm wavelength on the visible light spectrum). Attributes vary in complexity and level of abstraction. Complex attributes are composed of simpler attributes (e.g., MOMENTUM as the product of MASS and VELOCITY) (Gilreath, 1995, p. 45). Attributes like COLOUR, whose values can be perceived directly, can be described as concrete (Löhr, 2022, p. 559) However, humanitarian attributes like VULNERABILITY are more abstract and require complex composite indicators to be assessed (Sattar et al., 2020).

Studying conceptual variation affecting humanitarian attributes requires clarification on how to systematically represent their structure, values and ascription. Following Finlayson & Erjavec's (2017) recommendations, this contribution presents a preliminary semantic annotation scheme for attributes produced in a pilot study. Firstly, a phase zero annotation scheme was defined based on a tentative metamodel developed through a multidisciplinary literature review on attributes and their encoding in natural language. Secondly, a sample of KRCs on ACCOUNTABILITY, EFFECTIVENESS and VULNERABILITY was extracted from a corpus of English humanitarian documents obtained from ReliefWeb (Authors, 2024). Thirdly, the sample of KRCs was annotated with INCEpTION (Klie et al., 2018) by a team of annotators using the MAMA cycle (Pustejovsky

et al., 2017), an iterative process that refines the annotation scheme by adjusting the original metamodel based on interim findings. This is expected to increase the quality and accuracy of an annotation scheme before conducting a larger study. It also allows pre-existing theories on attributes to be corroborated empirically and potential unattested semantic relations to be identified. Expected findings will inform a subsequent study on a larger selection of humanitarian attributes.

References

Author, (2017)
Author, (2020)
Authors, (2024)
Condamines, A. (2022). How the Notion of "Knowledge Rich Context" Can Be Characterized Today. *Frontiers in Communication*, *7*. https://www.frontiersin.org/articles/10.3389/fcomm.2022.824711
Faber, P. (2015). Frames as a framework for terminology. In H. J. Kockaert, & F. Steurs (Eds.), *Handbook of Terminology: Volume 1* (pp. 14–33). John Benjamins Publishing Company. https://benjamins.com/catalog/hot.1.fra1
Faber, P. (2022). Frame-based Terminology. In P. Faber & M. C. L'Homme (Eds.), *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge* (Vol. 23, pp. 353–376). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.23.16fab
Fillmore, C. J. (2008). Frame semantics. In D. Geeraerts (Ed.), *Cognitive Linguistics: Basic Readings* (pp. 373–400). De Gruyter Mouton. https://doi.org/10.1515/9783110199901.373
Finlayson, M. A., & Erjavec, T. (2017). Overview of Annotation Creation: Processes and Tools. In N. Ide, & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 167–192). Springer Netherlands. https://doi.org/10.1007/978-94-024-0881-2
Gilreath, C. T. (1995). Merons, Taxons, and Qualities: A Taxonomy of Aspects. *Terminology*, *2*(1), 17–59. https://doi.org/10.1075/term.2.1.03gil
Hampton, J. A. (2020). Investigating differences in people's concept representations. In T. Marques, & A. Wikforss (Eds.), *Shifting Concepts: The Philosophy and Psychology of Conceptual Variability* (pp. 67–82). Oxford University Press. https://doi.org/10.1093/oso/9780198803331.003.0005
Hansen, N., & Chemla, E. (2017). Color adjectives, standards, and thresholds: An experimental investigation. *Linguistics and Philosophy*, *40*(3), 239–278. https://doi.org/10.1007/s10988-016-9202-7
Klie, J. C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In D. Zhao (Ed.), *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (pp. 5–9). Association for Computational Linguistics. https://aclanthology.org/C18-2002
Löhr, G. (2022). What Are Abstract Concepts? On Lexical Ambiguity and Concreteness Ratings. *Review of Philosophy and Psychology*, *13*(3), 549–566. https://doi.org/10.1007/s13164-021-00542-9
Nirenburg, S., & Raskin, V. (2004). *Ontological Semantics*. Mit Pr.
OCHA. (2024). Philippines—Global Humanitarian Overview 2025. *Humanitarian Action*. https://humanitarianaction.info/document/global-humanitarian-overview-2025/article/philippines
Oshima, D. Y., Akita, K., & Sano, S. (2019). Gradability, scale structure, and the division of labor between nouns and adjectives: The case of Japanese. *Glossa: A Journal of General Linguistics*, *4*(1), Article 1. https://doi.org/10.5334/gjgl.737

Pustejovsky, J., Bunt, H., & Zaenen, A. (2017). Designing Annotation Schemes: From Theory to Model. In N. Ide, & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 21–72). Springer Netherlands. https://doi.org/10.1007/978-94-024-0881-2

Sattar, M. A., Biswas, A. A. A., Islam, M. T., Hossain, M. A., Siddeqa, M., Rahim, M. A., Amin, M. N., Touhiduzzaman, M., Rahman, M. A., & Aktar, S. (2020). Disaster vulnerability and mitigation of humanitarian issues in coastal Bangladesh: Local evidence and knowledge gaps. *Progress in Disaster Science*, *8*, 100138. https://doi.org/10.1016/j.pdisas.2020.100138

A corpus-based study of legal language & migration: Exploring the interaction between keyword and multidimensional analysis

Daniel Granados Meroño

Universidad de Murcia

The application of Corpus Linguistics to discourse analysis can be traced back to the early 1990s, with Caldas-Coulthard's (1993) corpus-assisted gender study. Since then, the number and variety of such studies has increased throughout the decades (Anthony, 2018; Baker, 2019, 2023; Baker & McEnery, 2005; Baxter, 2003; Belica, 1996; Hunston, 2002; Johnson, S. et al., 2003; Partington et al., 2013; to name but a few).

Regarding the phenomenon of migration, it has been examined as seen in public discourse (mostly in the media) with the aid of corpus linguistics techniques (Baker, Gabrielatos, & McEnery, 2013; Blinder & Allen, 2016; Gabrielatos & Baker, 2008). However, only Pérez-Paredes et al. (2017) have studied legal texts (UK legislation and official information) to explore how migrants are depicted within this domain.

The present study departs from previous work (Marín-Pérez, 2019), where a corpus of British judicial decisions revolving around the topic of migration was scrutinised with the aim of finding evidence of the use of evaluative vocabulary, using Systemic Linguistics as reference (Eggins and Slade, 1997; Martin, 2003; Rothery & Stenglin, 2000; White, 1999). This research showed that 1.08% of the most frequent types extracted from a corpus of 600 legal texts could be deemed evaluative. Once such items were analysed in detail and classified according to the Appraisal System paradigm, it was found that almost half of the items (47.30%) belonged in the category appreciation, followed by the categories: judgment (28%), amplification (14%) and affect (10.70%).

Nevertheless, the degree of subjectivity involved in the implementation of a method of analysis like Systemic Linguistics was relatively high (despite it being corpus-assisted). This is why this study aims to provide a quantitative insight into legal language through the implementation of Multi-dimensional Analysis (Biber & Conrad, 2013). This method has been used for the study of different genres and specialised discourse, including legal English (Goźdź-Roszkowski, 2011; Granados-Meroño, 2023; Huang & Sang 2024; Matulewska, 2014).

MAT (Nini, 2019) tools were used in this study for the annotation of the corpus and the obtention of the variables frequency, while R library 'psych' (Revelle, 2017) was used for the conduction of FA and the visualisation of its results. Preliminary results show the presence of 5 factors clustering 67 linguistic features, although only Factors 2 and 3 were significantly different across the judgments related to migration and the rest of judgments. As regards Factor 2, we find first person pronouns, demonstratives and the verbs seem and appear, which are negatively correlated to adverbials, nominalisations and average word length (AWL). This Factor might be related to the Involved vs Evaluative Focus dimension (Alamri, 2023; Ehret & Taboada, 2021; Huang & Sang, 2024; Sun & Cheng, 2017). In turn, Factor 3 is compounded by nominalisations, conditional adverbial

subordination, possibility modals, infinitives, necessity modals, the It pronoun, and wh-clauses, which are negatively correlated to nouns. Factor 3 was connected with the Current Information vs Procedural Focus dimension (Alamri, 2023). Factorial loadings in the corpora show that Migration judgments have a higher load on the negative side of both dimensions, that is, procedural and evaluative focus.

References

Alamri, B. (2023). A Multidimensional Comparative Analysis of MENA and International English Research Article Abstracts in Applied Linguistics. *Sage Open*, *13*(1). https://doi.org/10.1177/21582440221145669

Anthony, L. (2018). Visualisation in corpus-based discourse studies. In C. Taylor, & A. Marchi (Eds.), *Corpus Approaches to Discourse: A critical review* (pp. 197-224). Routledge.

Baker, P. (2019). Analysing Representations of Obesity in the Daily Mail via Corpus and Down-Sampling Methods. In J. Egbert, & P. Baker (Eds.), *Using Corpus Methods to Triangulate Linguistic Analysis* (pp. 85-108). Routledge.

Baker, P. (2023). *Using Corpora in Discourse Analysis*. 2nd edition. Blooomsbury Discourse Series.

Baxter, (2003). *Positioning gender in discourse: A feminist methodology*. Palgrave MacMillan.

Belica, C. (1996). Analysis of temporal changes in corpora. *International Journal of Corpus Linguistics*, *1*(1), 61-73.

Biber, D., & Conrad, S. (2013). Introduction: Multi-dimensional analysis and the study of register variation. In S. Conrad, & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 3–12). Routledge.

Blinder S., & Allen W. L. (2016). Constructing immigrants: Portrayals of migrant groups in British national newspapers, 2010-2012. *International Migration Review*, *50*(1), 3-40.

Caldas-Coulthard, C.R. (1993). From discourse analysis to critical discourse analysis: the differential re-presentation of women and men speaking in written news. In G. Fox, M. Hoey, & J. M. Sinclair (Eds.), *Techniques of Description* (pp. 196-208). Routledge.

Eggins, S., & Slade, D. (1997). *Analysing casual conversation*. Casell.

Ehret, K., & Taboada, M. (2021). Characterising online news comments: A Multi-Dimensional cruise through online registers. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.643770

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. *Journal of English Linguistics*, *36*(1), 5-38.

Goźdź-Roszkowski, S. (2011). *Patterns of Linguistic Variation in American Legal English: A Corpus-based Study*. Peter Lang.

Granados-Meroño, D. (2023). Judgments of the English and Spanish Supreme Courts: A corpus-based approach to the legal English and Spanish discourse using multi-dimensional analysis. *Journal of Research Design and Statistics in Linguistics and Communication Science*, *7*(1), 21–68. https://doi.org/10.1558/jrds.22453

Huang, Y., & Sang, Z. (2024). Linguistic variation in supreme court oral arguments by legal professionals: A novel multi-dimensional analysis. *Discourse Studies*, 14614456231221075. https://doi.org/10.1177/14614456231221075

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Johnson, S., Culpeper, J., & Suhr, S. (2003). From "politically correct councillors" to "Blairite nonsense": Discourses of political correctness in three British newspapers. *Discourse and Society*, *14*(1), 28-47.

Marín-Pérez, M.J. (2019). Exploring the unexpected in legal discourse: a corpus-based contrastive analysis of Spanish and British judgments on immigration. *ESP Today*, *7*(2), 161-183.

Matulewska, A. (2014). A review of 'Patterns of linguistic variation in American Legal English. A corpus-based study' by Stanislaw Gostanisław Goźdź-Roszkowski. *Comparative Legilinguistics*, *19*, 135–138. https://doi.org/10.14746/cl.2014.19.07

Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In *Multi-Dimensional Analysis: Research Methods and Current Issues* (pp. 67–94). Bloomsbury Academic.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. John Benjamins.

Revelle, W. (2017) *psych: Procedures for Personality and Psychological Research*, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.7.8.

Rothery, J., Stenglin, M. (2000). Interpreting literature: The role of APPRAISAL. In L. Unsworth (Ed.), *Researching language in schools and communities: Functional linguistics perspectives* (pp. 231-264). Casell.

Sun, Y., & Cheng, L. (2017). Linguistic variation and legal representation in legislative discourse: A corpus-based multi-dimensional study. *International Journal of Legal Discourse*, *2*(2), 315-339. https://doi.org/10.1515/ijld-2017-0017

White, P. R. R. (July 20, 1999). The language of attitude, arguability and interpersonal positioning. *The Appraisal Website*. Retrieved from: http://www.grammatics.com/appraisal

Constructing the Narrative: 'Aboutness' in the opening and closing statements of State v. Chauvin [2021]

Natalie Jones, Alison May

University of Leeds

Research on 'aboutness indicators' (Kehoe & Gee, 2012) typically pinpoint content words as the most appropriate focus for analysis, while this paper argues for the importance of considering function words as keywords in small specialised corpora. While we expect 'the' to be the most frequent lexical item used in any corpus, because it is the most frequent word in English there is potential for 'the' to be overlooked as having little to contribute to lexical meaning; however, using a small, specialised corpus, such as the opening and closing arguments in the Chauvin trial, it is possible to explore 'the' and its use in context. Specific focus on the use of 'the', particularly through collocation analysis, reveals how barristers attempt to position the jury to view the police, the victim, the defendant, and expert opinion, in speech events that bookend the trial. This shows the value of considering function words in establishing the aboutness of a text, evident in the exploration of collocational patterns and semantic prosodies.

This paper focuses on the opening statements and closing arguments in the State of Minnesota v. Derek Michael Chauvin trial [2021], seeking to explore the simple question; what are these texts about? While the opening statement and closing argument are different sub-genres of the trial genre with differing goals, they are comparable for their monologic voice and direct address to the jury. With this in mind, a combined approach using critical discourse analysis (CDA) and corpus linguistics (CL) using AntConc 3.5.9 (Anthony, 2020) and Lexical Feature Marker 7.0 (Woolls, 2020), is used to identify patterns in the data, generate (key) wordlists, and undertake collocation analysis, while positioning theory (Davies and Harré, 1990), and the concept of 'aboutness' (Phillips,

1985; Scott and Tribble, 2006: 55; Scott, 2017) are employed to establish and explore the 'key point' (Scott and Tribble, 2006; Scott, 2017) of each opening statement and compare this with the closing arguments. Examining the lexical and grammatical patterns in the opening statements and closing arguments, seeks to establish each text's 'aboutness', revealing ideas that are 'key' to the prosecution and defence stories.

While previous studies tend to focus on the opening and closing separately (e.g. Felton-Rosulek, L. 2010; 2015 and Chaemsaithong, K. 2017; 2018), this paper looks at the texts comparatively, considering the transformation of the crime narrative throughout the trial as 'textual travel' (Heffer, et al. 2013). From the beginning to the end of the trial, the jury's perception of narrative events and social actors develops, as the texts are '[...] shaped, altered, and appropriated during their journeys' (Heffer, et al. 2013, p. 4) within the institutional context. Focusing on each text's 'aboutness', we establish how the crime narrative and social actors' identities are positioned and transformed through the close analysis of 'adversarial speech' (Mertz, 2007).

References

Anthony, L. 2020. *AntConc* (3.6.9). [Software]. [Accessed 28 September 2023].

Chaemsaithong, K. (2018). Use of voices in legal opening statements. *Social Semiotics*, *28*(1), 90–107.

Chaemsaithong, K. (2017). Evaluative stancetaking in courtroom opening statements. *Folia Linguistica*, *51*(1), 103–132.

Court TV. (2021). *Trial Archives*. [Online]. [Accessed 14 September 2023]. Available from: https://www.courttv.com/

Davies, B. and Harré, R. (1990). Positioning: The Discursive Production of Selves. *Journal for the Theory of Social Behaviour*, *20*(1), 43–63.

Felton-Rosulek, L. (2015). *Dueling discourses: the construction of reality in closing arguments*. Oxford: Oxford University Press.

Felton-Rosulek, L. (2010). Prosecution and defense closing speeches The creation of contrastive closing arguments. In M. Coulthard, & A. May (Eds.), *The Routledge Handbook of Forensic Linguistics* (pp. 218–230). London: Routledge.

Heffer, C., Rock, F., & Conley, J. (2013). *Legal-Lay Communication: Textual Travels in the Law*. Oxford: University Press

Kehoe, A., & Gee, M. (2012). *Reader comments as an aboutness indicator in online texts: Introducing the Birmingham Blog Corpus.* [Online] [Accessed 5 November 2024]. Available from: https://varieng.helsinki.fi/series/volumes/12/kehoe_gee/.

Mertz, E. (2007). *The Language of Law School: Learning to "Think Like a Lawyer"*. New York: Oxford University Press.

Phillips, M. (1985). *Aspects of Text Structure: An Investigation of the Lexical Organization of Text*. Amsterdam: North-Holland.

Scott, M. (2017). *News downloads and aboutness* [Online]. [Accessed 15 July 2023]. Available from: https://www.youtube.com/watch?v=3FVa0KwtvLc

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Philadelphia and Amsterdam: John Benjamins Publishing.

Woolls, D. (2020). *Lexical Feature Marker* (7.0). [Software]. [Accessed September 28 2023].

From doctor to algorithm: How AI is reshaping communication in medicine

Stefania Maci

Università degli Studi di Bergamo

The integration of Artificial Intelligence (AI) into healthcare is revolutionizing medical communication, not only altering the dynamics of doctor-patient interactions and the relevant clinical documentation, but also medical decision-making processes. AI-powered diagnostic tools and chatbots (such as Babylon Health, Ada Health) are reshaping patient communication, often prioritizing efficiency over the nuanced exchange essential in traditional medical interactions. However, the deployment of these technologies raises significant concerns, including biases embedded in training datasets and the depersonalization of care through algorithmic decision-making.

By drawing on Corpus Linguistics and (Critical) Discourse Analysis, this paper investigates the extent to which AI technologies are reshaping medical discourse. The study focuses on three main research questions:

(1) How does AI influence the linguistic structures and registers in medical communication?
(2) What ethical and practical challenges arise from AI-mediated communication, particularly regarding bias, accessibility, and multilingual contexts?
(3) How can corpus-based insights guide linguists in contributing to the design of AI systems that enhance, rather than hinder, patient-centred care?

The corpus-driven analysis, based on datasets of AI-mediated communication (47436 interactions; 5,942,068 words), including chatbot transcripts and medical documentation generated by AI tools, reveals significant shifts in linguistic practices within medical communication. AI systems are found to prioritize standardization and machine-readability, leading to a reduction in the lexical and syntactic variation typically seen in traditional patient-doctor interactions. While this enhances processing efficiency, it risks excluding culturally specific and idiomatic expressions essential for accurate diagnosis and meaningful rapport (Lapata & Barzilay, 2005). Preliminary findings from chatbot transcripts reveal recurring patterns of pragmatic failures, where pre-programmed responses fail to address complex or non-linear patient narratives, often resulting in patient dissatisfaction or miscommunication.

Additionally, automated transcription and summarization tools streamline clinical workflows but often oversimplify contextual details, as evidenced by corpus-based error analysis of transcription outputs (Bickmore & Schulman, 2021). These findings underscore the ethical risks of AI systems perpetuating bias and depersonalizing interactions, undermining the trust and empathy critical to effective clinical care (Mittelstadt et al., 2016).

By analysing the linguistic features of AI-mediated communication on a large scale and with a corpus-driven approach, gaps in AI training data can be detected and filled to better offer culturally inclusive, patient-centred designs. Furthermore, corpus-based insights can refine AI outputs for clarity, empathy, and contextual accuracy, ensuring that technological advancements align with the principles of equitable and effective medical communication. Ultimately, this paper underscores the importance of leveraging corpus-based methodologies to balance AI-driven innovation with the preservation of human connection in medicine

References

Bickmore, T. W., & Schulman, D. (2021). Automated communication for healthcare. *Annual Review of Biomedical Engineering*, *23*, 29-55.

https://doi.org/10.1016/j.artmed.2020.101822
Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1-21. https://doi.org/10.1177/2053951716679679
Lapata, M., & Barzilay, R. (2005). Automatic summarization of medical documents. *ACL Anthology*. Available at: https://aclanthology.org/2019.jeptalnrecital-recital.3.pdf

Food risk communication: A corpus-driven study

Camino Rea Rizzo, Antonio Fornet Vivancos

Universidad Politécnica de Cartagena, Universidad Rey Juan Carlos

The European Food Safety Authority (EFSA) is an agency of the European Union that provides scientific advice to risk managers and communicates on existing or emerging risks associated with the food chain. The agency establishes the scientific basis for laws and regulations aimed at protecting European consumers from food-related risks. The EFSA also assists in the field of risk communication to the different stakeholders involved, from the scientific community when presenting results of risk assessment, to risk managers disseminating information about a risk or even food crisis.

In fact, the EFSA published the comprehensive and detailed report "Technical Assistance in the Field of Risk Communication" (2021) as requested by the European Commission, where risk communication is defined as "the interactive exchange of information and opinions throughout the risk analysis process as regards hazards and risks, risk-related factors and risk perceptions, among risk assessors, risk managers, consumers, feed and food businesses, the academic community and other interested parties, including the explanation of risk assessment findings and the basis of risk management decisions" (2021, p. 14).

In this study, we will focus on the how the scientific community reports food-borne risk assessment, that is, the scientifically based process by which hazard is identified and characterized, the exposure to such hazard is assessed, and finally risk is characterized. As specified by the EFSA, risk assessment should conclude on "the likelihood that an agent will cause harm taking into account the nature of the hazard and the extent to which people, animals, plants and/or the environment are exposed to it" (Ibid), and account for the scientific uncertainties on the assessment conclusions. To this end, a corpus of abstracts from scientific articles published in high impact journals related to food safety and food microbiology has been compiled, reaching 345,546 tokens.

It was hypothesised that those abstracts would show a concern with uncertainty, as can be reasonably expected from all strands of risk communication (varied as the latter actually is). Indeed, risk itself has been defined not just as "the possibility of an adverse outcome" but also as "uncertainty over the occurrence, timing or magnitude of that adverse outcome" (Covello et al, 1993, p. 2). Uncertainty, in its turn, has been characterised by EFSA as "referring to all types of limitations in available knowledge that affect the range and probability of possible answers to [a risk] assessment question" (EFSA Scientific Committee, 2018, p. 11). Since limitations in knowledge are normally expressed in language by means of "probability phrases" and "verbal expressions" (Willems et al, 2020:2-3) such as chance, improbable or unlikely (among other verbs, nouns, adjectives and adverbs "covering linguistic meanings that indicate degree of epistemic support for a proposition"; Boye, 2016, p. 117), it can be likewise assumed that these phrases and expressions will be a regular feature of risk communication. Preliminary results from our study, however, suggest that those expressions may not be as frequent as expected in scientific food-borne risk characterisation.

References

Boye, K. (2016). The expression of epistemic modality. In J. Nuyts, & J. Van Der Auwera (Eds.). *The Oxford book of modality and mood* (pp. 117-140). Oxford University Press.

Covello, V. T., Merkhofer, M. W., Covello, V. T., & Merkhofer, M. W. (1993). Risk estimation. Risk Assessment Methods: *Approaches for Assessing Health and Environmental Risks*, 203-237.

EFSA, Hart, A., Maxim, L., Siegrist, M., Von Goetz, N., da Cruz C., Merten, C., Mosbach-Schulz, O., Lahaniatis, M., Smith, A., & Hardy, A. (2019). Guidance on Communication of Uncertainty in Scientific Assessments. *EFSA Journal*, *17*(1), 5520. https://doi.org/10.2903/j.efsa.2019.5520

Maxim L., Mazzocchi, M., Van den Broucke, S., Zollo, F., Robinson, T., Rogers, C., Vrbos, D., Zamariola, G., & Smith, A. (2021). Technical Assistance in the Field of Risk Communication. *EFSA Journal*. https://doi.org/10.2903/j.efsa.2021.6574

Willems, S., Albers, C., & Smeets, I. (2020). Variability in the interpretation of probability phrases used in Dutch news articles — a risk for miscommunication. *JCOM*, *19*(02), A03. https://doi.org/10.22323/2.19020203

Framing lithium to attract stakeholders: A corpus ecostylistic discourse analysis of lithium mining in Cáceres

Ana María Terrazas-Calero, Carolina Amador-Moreno

University of Bergen, Universidad de Extremadura

The emerging field of Ecolinguistics, which explores the role of language in the interactions between humans, other species, and the environment (Fill & Penz 2017) while also looking to demonstrate how linguistics can address key ecological issues (e.g. environmental justice, climate change, biodiversity loss, etc.), has benefitted from the application of corpus-analytical techniques (Poole 2022). This paper proposes a case study focusing on lithium mining which analyzes the discourse and frames used in corporate discourse with regard to changes in resource availability and energy solutions as an alternative to current, environmentally-damaging resources.

In 2020, the European Commission updated their Critical Raw Materials list to include lithium as one of the key materials the EU needs to obtain to bring security and sustainability to Europe, encouraging member states to legislate in favor of mining and treating lithium within the EU. In the context of Spain, Australian mining company, Infinity Lithium, identified the *Valdeflores-San José* mine as an unexploited lithium mine. Originally designed as an open-pit mine located close to the UNESCO World Heritage city of Cáceres (Spain), the development has encountered strong opposition from activists and locals who reject the project which they see as a threat to their socio-ecological environment through demonstrations, "NO A LA MINA" signs, and even an online petition, inter alia.

This case study, therefore, will explore the linguistic strategies and discursive frames *Infinity Lithium* uses in a corpus of Australian newspaper and magazine articles to address the protesters' concerns and garner their support. For this paper we have selected a subcorpus of articles which favor the company and its project (i.e. *ProINF* subcorpus) and which comprises 56,331 words. The goal of the paper is to investigate and identify 1) what are the frames *Infinity Lithium* utilizes to construct their corporate image regarding their *Valdeflores-San José* project, and 2) how do they linguistically frame their discourse to attract public and private 'stakeholders'. To do so, we will focus on their use of the key words *European* and *Extremadura*, which we will analyze using Stibbe's (2015) frame and metaphor identification method with the application of both quantitative and qualitative

Corpus Ecostylistics techniques. Preliminary findings suggest that the manner in which the company addresses both public and private 'stakeholders' is by using a variety of frames, which align with guidelines on good corporate governance and shareholder attraction, and which linguistically project the *San José* project, and their corporate image by extension, as a pioneering, energy-transition-facilitating, employment-producing, European endeavor.

References

Fill, A. F., & Penz, H. (eds.). (2017). *The Routledge Handbook of Ecolinguistics*. Abingdon: Routledge.

Poole, . (2022). *Corpus-Assisted Ecolinguistics*. New York: Bloomsbury.

Stibbe, A. (2015/2020). *Ecolinguistics Language, Ecology and the Stories We Live By*. Oxford: Routledge.

POSTERS

Riqueza cromática en textos jurídicos y narrativos

Yaiza Santana-Alvarado, Anabel Mederos-Cedrés, Mª Teresa Cáceres-Lorenzo

Universidad de Las Palmas de Gran Canaria

En la identificación del léxico americano se han encontrado resultados empíricos del uso de términos cromáticos en documentos preferentemente en prosa. Esta investigación se realiza como parte del proyecto AMERLEX de I+D+i PID2022-138801NB-I00 (https://amerlex.iatext.ulpgc.es). Para la elaboración de este proyecto se han realizado numerosas búsquedas para poder seleccionar e identificar el corpus de lemas americanizados que se utilizan en territorio americano. El objetivo de esta investigación es crear un lexicón de los términos cromáticos en el periodo seleccionado (1500-1740) que sirva de guía para posteriores búsquedas en AMERLEX.

El corpus elegido está constituido por la prosa de naturaleza legal, además de la que sirve para el acto comunicativo de contar, referir lo sucedido, un hecho o una historia, preferentemente, ficticios. Los documentos de la prosa jurídica son un medio de interpretar el gran conjunto de normas surgidas de las costumbres sociales y de la moral en territorio americano. En la narrativa encontramos textos ficcionales, como relatos breves, y en textos como noticias y crónicas. Para conseguir el objetivo se ha utilizado una metodología basada en la búsqueda del ejemplo y su subsiguiente análisis cuantitativo, con el fin de poder establecer una tendencia. En esta ocasión se ha realizado una indagación en el corpus del actual *Diccionario histórico de la lengua española* (CDH) hasta obtener como resultado la identificación de 81 términos cromáticos en la prosa jurídica y 157 términos en la narrativa, lo que nos permite distinguir entre voces panhispánicas, y otras que se especializan en América para nuevas realidades. Junto a esto encontramos el uso habitual de voces patrimoniales (*amarillo*, *azul*, *blanco*, *carmesí*, *colorado*, *dorado*, etc.)

Los resultados muestran que la creación de nuevas unidades léxicas para describir el color es un proceso que a menudo se inicia en la Edad Media (*albino*, *aloque*, *bermejura*, *celestre*, *negral*, *sabino*, etc.). A pesar de su origen medieval, las designaciones cromáticas se utilizan con mayor frecuencia en las producciones escritas sobre América que revitalizan estas palabras, ya que existe una necesidad comunicativa de distinguir las realidades a las que se refieren los textos. Son los casos de *pardo* que se emplea para designar una persona de tez oscura y rasgos aindiados, o *prieto* para una persona negra. Los dos términos son necesarios tanto en textos jurídicos como narrativos. Esta riqueza cromática de significados se utiliza a través de un proceso de adaptación semántica y formal que está presente en distintas tipologías textuales.

Léxico sobre el aspecto físico y moral en textos americanos (AMERLEX, 1500-1740)

Anabel Mederos-Cedrés, Yaiza Santana-Alvarado, Mª Teresa Cáceres-Lorenzo

Universidad de Las Palmas de Gran Canaria

AMERLEX es un proyecto de Humanidades Digitales (I+D+i PID2022-138801NB-I00) alojado en <https://amerlex.iatext.ulpgc.es> que tiene como finalidad la construcción de una base de datos, con su respectivo buscador, que ofrezca ejemplos empíricos textuales

del uso de los americanismos, al mismo tiempo que se reconocen las ideas que influyen en el vocabulario. En esta comunicación se presentan los resultados de investigación obtenidos tras la identificación y análisis de ejemplos de términos que se utilizan para describir el aspecto físico y moral en los documentos seleccionados en AMERLEX.

Se trata de documentos en los que se describe la realidad americana, por lo que muchos de ellos se identifican con lo que conocemos bajo el marbete de crónicas de Indias. Los textos escritos para describir la realidad americana representan la redacción de un número considerable de escritos de distintas tipologías que respondían al propósito del rey de España de estar continuamente informado por la necesidad de conocer "la entera noticia". De esta manera cada autor escribe sobre la realidad americana como testigo experto, en los que el léxico seleccionado coincide con lo que se considera el incipiente español americano. En este contexto, el vocabulario americano que se recopila en AMERLEX es predecesor del actual, que mantiene vitalidad cultural porque se registra en los repertorios lexicográficos dialectales publicados desde las distintas academias de la lengua.

Esto nos ha hecho indagar cerca de 100 textos producidos entre 1500-1740, periodo de auge económico de muchas de las urbes de América que lideraban una extensa red comercial y cumplían la función de organizadoras de una sociedad criolla enriquecida en la que conviven distintos grupos humanos (indígenas, mulatos, negros, peninsulares, otros europeos, etc.).

El estado de la cuestión sobre la historia léxico-semántica de los americanismos ha aportado datos parciales sobre el vocabulario que aparece en los distintos documentos. La elaboración de una base de datos extensa como AMERLEX hace posible una investigación panorámica. Las principales conclusiones afirman que el vocabulario se americaniza por necesidades sociales y se percibe la tendencia de la adaptación de la lengua colonizadora a la realidad americana y las preferencias de los hablantes por el indoamericanismo o el hispanismo americanizado.

El lexicón recopilado reúne 50 términos que son analizados a través de una metodología cuantitativa y cualitativa a través de su datación en distintas centurias. Con estos datos empíricos se obtienen resultados diacrónicos que se consideran la primera fase de una investigación que debe completarse con otras muestras de lenguas textuales.

CILC2025 gratefully acknowledges the financial support of the following funding bodies and sponsors

CILC2025 agradece los apoyos económicos de las siguientes instituciones y patrocinadores

DEPARTAMENTO DE FILOLOGÍA INGLESA

Filología USAL

VNiVERSiDAD Đ SALAMANCA

FACVLTAD Đ FiLOLOGíA

Fundación General

aelinco
Asociación Española de Lingüística de Corpus

GOBIERNO DE ESPAÑA

MINISTERIO DE CIENCIA E INNOVACIÓN

JOHN BENJAMINS PUBLISHING COMPANY

OXFORD UNIVERSITY PRESS

uncom