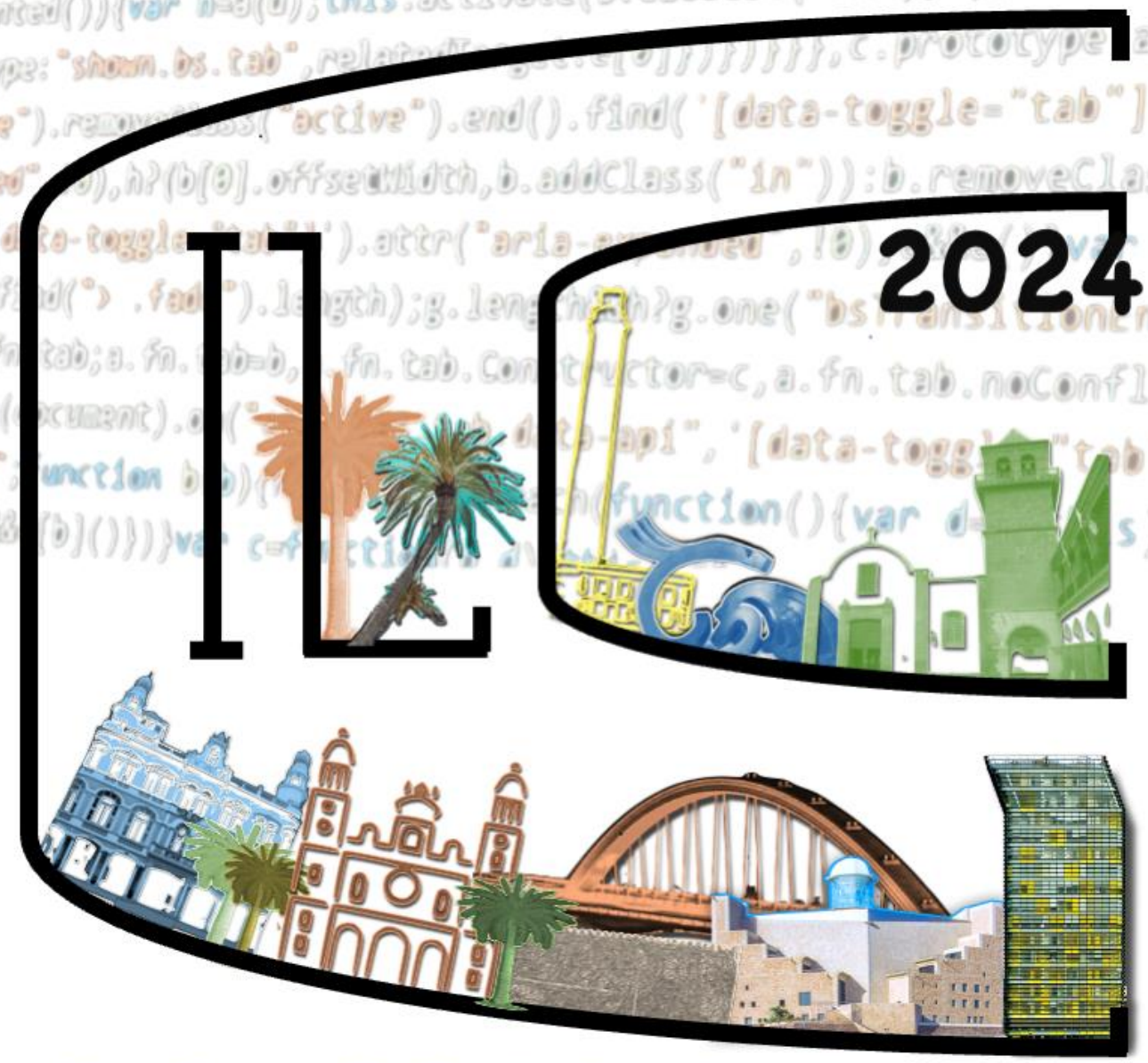


15th International Corpus Linguistics Conference CILC2024

XV Congreso Internacional de Lingüística de Corpus CILC2024

Corpus Linguistics, (digital) discourse, and AI. Unlocking new horizons

Lingüística de corpus, discurso (digital) e IA. Abriendo nuevos horizontes

Las Palmas de Gran Canaria, Spain, 22-24 May 2024

Las Palmas de Gran Canaria, España, 22-24 mayo, 2024

ULPGC · UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

aelinco
Asociación Española de Lingüística de Corpus

ULPGC
Universidad de Las Palmas de Gran Canaria
Departamento de Filología Moderna, Traducción e Interpretación

ULPGC
Departamento de Filología Hispánica, Clásica y de Estudios Árabes y Orientales

ULPGC
FTI
FACULTAD DE TRADUCCIÓN E INTERPRETACIÓN

Discourse, Communication & Society
DiCoS

Cabildo de Gran Canaria

Elaborado en Canarias

Tirma

Buen Lugar

PROYECTO VALLE DE LA LENGUA

La Rioja

UNIVERSIDAD DE LA RIOJA

JOHN BENJAMINS PUBLISHING COMPANY

BOOK OF ABSTRACTS
XV Congreso Internacional de Lingüística de Corpus
CILC2024

ÍNDICE

Lugar de celebración
Venue

The Conference will be held at Edificio de Humanidades, Campus del Obelisco, calle Pérez del Toro 1, near Plaza de la Constitución. The campus is located in the city centre of Las Palmas de Gran Canaria.

Edificio de Humanidades “Agustín Millares Carló"
c/ Pérez del Toro, 1, 35003 Las Palmas de Gran Canaria, Las Palmas

Junta directiva de AELINCO/
AELINCO executive board

Presidente

Miguel Fuster Márquez
Universitat de València

Vicepresidente

Jorge Leiva Rojo
Universidad de Málaga

Secretaria

Ángela Almela Sánchez-Lafuente
Universidad de Murcia

Tesorera

Natalia Judith Laso Martín
Universitat de Barcelona

Vocal

Francisco José Álvarez Gil
Universidad de Las Palmas de Gran Canaria

Paneles y directores
Panels and directors

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

Luís Miguel Puente Castelo
Universidade da Coruña

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

Giovanni Garofalo
Università degli Studi di Bergamo

Estudios gramaticales basados en corpus/ Corpus-based grammatical studies

Iván Tamaredo Meira
Universidad Complutense de Madrid

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

Javier Fernández Cruz
Universidad de Málaga

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

Marlén Izquierdo Fernández
Universidad del País Vasco

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

Zeltia Blanco Suárez
Universidade de Santiago de Compostela

Lingüística computacional basada en corpus/ Corpus-based computational linguistics

Chantal Pérez Hernández
Universidad de Málaga

Corpus, adquisición y enseñanza de lenguas/ Corpora, language acquisition and teaching

Carmen Maíz Arévalo
Universidad Complutense de Madrid

Fines específicos y lingüística de corpus/ Special uses of corpus linguistics

María José Marín Pérez
Universidad de Murcia

Agradecimientos
Acknowledgements

En nombre de AELINCO, el Comité Organizador desea agradecer el generoso apoyo de las siguientes instituciones y empresas.
On behalf of AELINCO, the Organizing Committee would like to thank the generous support of the institutions and companies listed below.

Plenarios
Plenary speakers

Corpus exploitation for Natural Language Processing: what matters more? Quality or quantity? Human-crafted rules or Artificial Intelligence?

Prof. Ruslan Mitkov, University of Lancaster

The keynote speech will seek to illuminate two important questions relevant to corpus linguistics and natural language processing. The first question is what matters more – the quantity or quality of corpora? To shed light on this perennial question, the results from two NLP studies that exploit corpora of different qualities will be reported. The second question to be answered is whether human-crafted rules based on corpus evidence can compete with artificial intelligence methods.

The first study investigates (and compares) the impact of the size and the quality of comparable corpora on the specific task of extracting translation equivalents of verb-noun collocations from such corpora. The results of a comprehensive evaluation of different configurations of English and Spanish corpora will be reported.

The second ongoing study exploits three Holocaust datasets of different sizes and qualities for the task of Named Entity Classification. The results will be available for (and reported for the first time in) Mitkov's keynote speech at CILC2024 in Las Palmas de Gran Canaria. In addition to seeking to shed light on whether the quantity or quality of the data is more important, this ongoing study will answer another fundamental question: Which methodology works best? Classical rule-based NER approaches, deep learning methods, or Language Models (LLMs)? Can old-fashioned, corpus-based rules compete with the latest LLMs?

This is talk is sponsored by:

PROYECTO_ VALLE DE LA LENGUA

La Rioja

UNIVERSIDAD DE LA RIOJA

ULPGC Universidad de Las Palmas de Gran Canaria
Departamento de Filología Moderna, Traducción e Interpretación

¿(Cómo) ha evolucionado el uso de corpus en la enseñanza de la traducción?

Prof. Patricia Rodríguez Inés, Universitat Autònoma de Barcelona

Desde aquellas publicaciones de finales de los años 90 en las que se narraban experiencias didácticas puntuales de uso de corpus para enseñar a traducir, hasta hoy, ha llovido mucho, pero, ¿ha cambiado en esencia la forma de crear, seleccionar, presentar y explotar estos recursos a la hora de formar traductores?

Del interés en la aplicación de corpus en esta área dieron fe la serie de conferencias CULT (Corpus Use and Learning to Translate), celebradas en 1997, 2000, 2004 y 2015. En aquellos primeros tiempos aparecieron publicaciones abundantes, tanto en traducción directa (Piotrowska 1997; Bowker 1998; Pearson 1999; Maia 2000; Rodríguez López 2002; Zanettin 2002; Kübler 2003, etc.), como en traducción inversa (Zanettin 1998, 2001, 2002; Gavioli y Zanettin 2000; Corpas Pastor 2001, 2002; Varantola 2003, etc.), así como trabajos que abordaban ambas direcciones, distintos niveles de competencia y ámbitos de especialización (Rodríguez-Inés 2008, 2009, 2010, 2013, 2014). En muchas de estas referencias, y aunque no se hiciera explícito, lo que se conseguía con los ejercicios descritos no era solo utilizar el corpus para solucionar problemas concretos de traducción, sino, aún más importante, aprender a traducir descubriendo, pensando, evaluando, redefiniendo, reflexionando.

"[…] however paradoxically, corpora can and should be employed to problematise rather than simplify the task of (future) translators. The greatest pedagogic value of the instrument lies, we suggest, in its thought-provoking, rather than question-answering, potential" (Bernardini, Stewart y Zanettin 2003).

Los avances tecnológicos, la disponibilidad de corpus y repositorios de textos digitales, las iniciativas para la construcción colaborativa de corpus, la integración de corpus en software de traducción, y la interactuación con otras disciplinas, entre otros factores, han propiciado el crecimiento de los estudios basados en corpus en las últimas décadas. Con el paso de los años la aplicación de corpus en la enseñanza de la traducción también se ha extendido y diversificado, si bien es cierto que no siempre se es consciente de que se está usando un "corpus", en el sentido laxo del término. Algunos ejemplos de esto son el enfoque de la web as corpus, las

memorias de traducción, los traductores automáticos, y ahora algunas aplicaciones de inteligencia artificial. No son corpus, stricto sensu, porque algunos no pretenden representar una lengua, un estadio o una variante de una lengua, pero como compilaciones de textos, tienen su utilidad, siempre y cuando se sea muy consciente de su composición.

Esta presentación mostrará usos de los corpus en traducción en la enseñanza de la traducción, aportando ejemplos de cómo éstos pueden ayudar no solo a aprender a traducir, sino a concienciar al alumnado sobre la importancia del factor humano y de su capacidad analítica e interpretativa, especialmente en un mundo cada vez más automatizado y tecnologizado.

This is talk is sponsored by:

ULPGC Universidad de Las Palmas de Gran Canaria — Departamento de Filología Moderna, Traducción e Interpretación

ULPGC FTI FACULTAD DE TRADUCCIÓN E INTERPRETACIÓN

ULPGC Departamento de Filología Hispánica, Clásica y de Estudios Árabes y Orientales

Interpersonality in Academic and digital genres,

Prof. María Luisa Carrió-Pastor, Universitat Politècnica de València

In this talk, I focus on the use of interpersonal strategies in academic and digital genres. The way writers communicate their findings in academic papers is crucial for convincing readers about the objectiveness and validity of research and it is also crucial in social media to attract attention and increase the number of followers. Thus, one of the challenges to be faced by human beings is to convince others of the importance of the validity of propositions. Specifically, interpersonal devices are used in different genres to attract the attention of readers or listeners and some categorizations have been proposed for their analysis, such as stance, evaluation, appraisal, metadiscourse, or voice. The use of corpus analysis to identify different patterns has been used by many researchers, who proposed categories and discussed their characteristics. Here I integrate the frameworks of metadiscourse, stance and appraisal theory to propose a notion broad enough to encompass diverse linguistic phenomena that can be considered interpersonal devices. Some examples from academic discourse and social media discourse are also shown to illustrate the validity of the proposal.

This is talk is sponsored by:

ULPGC Universidad de Las Palmas de Gran Canaria — Departamento de Filología Moderna, Traducción e Interpretación

aelinco Asociación Española de Lingüística de Corpus

Programa
Programme

	MIÉRCOLES 22 DE MAYO DE 2024 WEDNESDAY 22 MAY 2024

9:00-9:30	**RECOGIDA DE CREDENCIALES - *REGISTRATION*** ENTRADA EDIFICIO PRINCIPAL - *ENTRANCE MAIN BUILDING*
9:30-10:00	**ACTO DE APERTURA - *OPENING SESSION*** SALÓN DE ACTOS DEL EDIFICIO DE PROFESORADO - *AUDITORIUM OF THE EDIFICIO DE PROFESORADO*

10:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 2
Diseño, elaboración y tipología de corpus - *Corpus design, compilation and types* Panel moderado por/ Panel chaired by: **Mercedes Cabrera**		
10:00-10:30	Placing the Coruña Corpus in the world: The case of CEGeT Isabel Moskowich & Begoña Crespo, Universidade A Coruña	
10:30-11:00	Diseño y elaboración de un corpus textual divulgativo de reproducción asistida para evaluar el nivel de percepción de las pacientes Ana Reyes Herrero	
10:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 3
Discurso, análisis literario y corpus - Discourse, literary analysis and corpora Panel moderado por/ Panel chaired by: **Giovanni Garofalo**		
10:00-10:30	Deborah, Linguist *yet* Professor Michael. How British corpora reflect gender-relation through forms of address Michael Pace-Sigge	
10:30-11:00	El discurso biosanitario en torno a la salud de la mujer: un análisis asistido por corpus de los recursos metadiscursivos y evaluativos Giovanni Garofalo & Luisa Chierichetti	
10:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 4
Estudios gramaticales basados en corpus - *Corpus-based grammatical studies* Panel moderado por/ Panel chaired by: **Iván Tamaredo Meira**		
10:00-10:30	Detección de errores frecuentes en la escritura académica de estudiantes chilenos: tendencias para una plataforma inteligente Anita Ferreira Cabrera	
10:30-11:00	El acusativo preposicional en catalán al final de la Edad Moderna (1833-1903). Un estudio de corpus Josep E. Ribera i Condomina	
10:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 13
Lexicología y lexicografía basadas en corpus - Corpus-based lexicology and lexicography Panel moderado por/ Panel chaired by: **Javier Fernández Cruz**		
10:00-10:30	Análisis léxico de la influencia del francés en textos de gastronomía en lengua inglesa del siglo XIX escritos por mujeres Gabriel Díez Abadie & Rocío Gragera Retuerto	
10:30-11:00	Estrategia combinada para la definición del vocabulario arquitectónico contemporáneo francés: un enfoque basado en corpus y tecnología IA Zaida Bartolomé-Díaz	
10:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: *15*
Lingüística computacional basada en corpus - Corpus-based computational linguistics		

	Panel moderado por/ Panel chaired by: **María Chantal Pérez Hernández**	
10:00-10:30	Linguistic remix – Mapping the intertextual relationship of poetic texts with an n-gram approach Emese K. Molnár & Andrea Dömötör	
10:30-11:00	Old English text generation. A viable strategy of data augmentation? Ana Elvira Ojanguren López	
10:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: *17*
Corpus, adquisición y enseñanza de lenguas - *Corpora, language acquisition and teaching*		
	Panel moderado por/ Panel chaired by: **Carmen Maíz Arévalo**	
10:00-10:30	Examining the potential of AI in the annotation of corpus examples for language learning Iztok Kosem, Tanara Zingano Kuhn, Špela Arhar Holdt, Kristina Koppel, Carole Tiberius, Rina Zviel Girshin	
10:30-11:00	Spanish EFL learners' use of contrastive paratactic expressions across three CEFR levels and gender influence Carmen Maíz-Arévalo	
10:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: *18*
Usos y aplicaciones específicas de la lingüística de corpus - *Special Purposes and Corpus Linguistics*		
	Panel moderado por/ Panel chaired by: **María José Marín Pérez**	
10:00-10:30	El potencial de los corpus en el aula de chino: estado actual y uso complementario con ChatGPT Lingzhi Nie	
10:30-11:00	A corpus-based sociolinguistic study of distribution characteristics and influencing factors of *must* and *have to* in London teen language Tianai Zhang	
10:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: *5*
Variación y cambio lingüístico basados en corpus - Linguistic variation and change through corpora		
	Panel moderado por/ Panel chaired by: **Elena Quintana Toledo**	
10:00-10:30	What's done can't be undone': Verbal Contractions in Modern English Marta Pacheco-Franco & Javier Calle-Martín	
10:30-11:00	Change in motion: On the rise and development of mirative *wind up* Mario Serrano Losada	
10:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: *24*
Corpus, estudios contrastivos y traducción - Corpora, contrastive studies and translation		
	Panel moderado por/ Panel chaired by: **Marlén Izquierdo Fernández**	
10:00-10:30	Análisis contrastivo del uso de los verbos modales deónticos entre dos libros de recetas escritos por mujeres en el siglo XIX: "A Lady" (1818) y Beeton (1875) Isabel Soto Déniz	
10:30-11:00	Inocuidad alimentaria y traducción: evidencias terminológicas a partir de un corpus paralelo inglés-español Jorge Leiva Rojo	

11:00-11:30	**REFRIGERIO - *COFFEE BREAK*** PATIO EDIFICIO DEPARTAMENTAL - *DEPARTMENTAL BUILDING COURTYARD*

11:30-13:00	**SESIÓN 2 - *SESSION 2*** MAIN BUILDING	AULA - *ROOM*: 2
Diseño, elaboración y tipología de corpus - *Corpus design, compilation and types*		

	Panel moderado por/ Panel chaired by: **Luis Puente-Castelo**	
11:30-12:00	Designing a representative corpus of Maltese Joseph Buttigieg	
12:00-12:30	Hacia la construcción de un corpus de informes médicos en español: superando barreras lingüísticas en la salud de la mujer Ovidia Martínez Sánchez	
12:30-13:00	Written Learner Corpus of Brazilian Portuguese Shintaro Torigoe	
11:30-13:00	**SESIÓN 2 - *SESSION 2*** MAIN BUILDING	AULA - *ROOM*: *3*
Discurso, análisis literario y corpus - Discourse, literary analysis and corpora Panel moderado por/ Panel chaired by: **Giovanni Garofalo**		
11:30-12:00	Corpus linguistics applied to the language of law: A corpus-driven analysis of French parliamentary debates on immigration. Nadia Makouar	
12:00-12:30	Digital discourse on TripAdvisor: a genre analysis of negative hotel reviews written in Italian, French and Spanish Irene Cenni	
12:30-13:00	Navigating Gendered Terrain: An Analysis of Politeness and Impoliteness Strategies in Female Instructional Texts Walter Yared Armas Cáceres	
11:30-13:00	**SESIÓN 2 - *SESSION 2*** MAIN BUILDING	AULA - *ROOM*: 4
Corpus, estudios contrastivos y traducción - Corpora, contrastive studies and translation Panel moderado por/ Panel chaired by: **Marlén Izquierdo Fernández**		
11:30-12:00	The translation of English derivational adjective compounds in Romance languages: a corpus-based case study Raluca Nita & Ramón Martí Solano	
12:00-12:30	The translation of “yet” as an adverb and conjunction: English language and translation teaching and learning through an English-Spanish parallel corpus Sidoní López Pérez	
12:30-13:00	Ensayos clínicos y traducción automática neuronal: clasificación de errores según MQM-DQF Alicia Picazo Izquierdo & Adelina Gómez González-Jover	
11:30-13:00	**SESIÓN 2 - *SESSION 2*** MAIN BUILDING	AULA - *ROOM*: 13
Variación y cambio lingüístico basados en corpus - Linguistic variation and change through corpora Panel moderado por/ Panel chaired by: **Elena Quintana Toledo**		
11:30-12:00	A corpus-based research on the SQUARE lexical set in the dialect of nineteenth-century Lancashire Nadia Hamade Almeida	
12:00-12:30	A Multidimensional Analysis of British Legal Genres: Statute Law vs. Case Law Daniel Granados Meroño	
12:30-13:00	A relevance-theoretic account of the use of emoji in Twitter for Research Dissemination Purposes Silvia Murillo Ornat	
Usos y aplicaciones específicas de la lingüística de corpus - *Special Purposes and Corpus Linguistics* Panel moderado por/ Panel chaired by: **María José Marín Pérez**		
11:30-12:30	**SESIÓN 2 - *SESSION 2*** MAIN BUILDING	AULA - *ROOM*: *5*
11:30-12:00	Mapping pre-digital tourism communication of the town of Ferrara and its province between the 1960s and the 2000s: a quantitative and qualitative analysis Eleonora Federici	
12:00-	The impact of MT as a writing tool on EFL Academic Writing: a qualitative linguistic	

12:30	analysis Natalia Judith Laso Martín & Elisabet Comelles Pujadas
13:00- 13:30 **PÓSTER**	Corpus-based language tools: The case of English for Specific Purposes in the wine and olive oil tasting domains Lucía Sanz-Valdivieso

16:15	**SALIDA AL RECTORADO CAMINANDO /** **DEPARTURE TO THE RECTORATE BY WALKING** LUGAR DE REUNIÓN: ENTRADA DEL EDIFICIO PRINCIPAL MEETING POINT: ENTRANCE, MAIN BUILDING

17:30- 18:30	**SESIÓN PLENARIA - PLENARY SESSION** EDIFICIO DEL RECTORADO DE LA UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA PLANTA BAJA/ RECTORATE MAIN BUILDING **¿(Cómo) ha evolucionado el uso de corpus en la enseñanza de la traducción?** Prof. Patricia Rodríguez Inés Universitat Autònoma de Barcelona
18:30	**RECEPCIÓN DE BIENVENIDA - WELLCOME RECEPTION** PATIO DEL RECTORADO / RECTORATE COURTYARD

	JUEVES 23 DE MAYO DE 2024 THURSDAY 23 MAY 2024

09:30-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 2
Diseño, elaboración y tipología de corpus - *Corpus design, compilation and types* Panel moderado por/ Panel chaired by: **Luis Puente-Castelo**		
09:30-10:00	La anotación morfosintáctica de los corpus desde la perspectiva del usuario Eva María Domínguez Noya & María Paula Santalla del Río	
10:00-10:30	Creación del Corpus de errores en lengua catalana del nivel C1: El CELC Francesca Romero Forteza	
10:30-11:00	CoLaGe (*Corpus for the Study of Language and Gender in Spanish*): un corpus bidialectal del español oral Andrea Carcelén Guerrero, Gloria Uclés Ramada, Pekka Posio & Sven Kachel	
09:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 3
Discurso, análisis literario y corpus - Discourse, literary analysis and corpora Panel moderado por/ Panel chaired by: **Giovanni Garofalo**		
09:00-09:30	An exploration into modality and intonation in peer interaction Mercedes Cabrera-Abreu & Eva Estebas-Vilaplana	
09:30-10:00	Presence and absence of *laughter* and *gestures*. examples from the BNC-spoken 2014 and dickens' novels Michael T.L. Pace-Sigge	
10:00-10:30	Análisis de la metáfora del personaje de Xue Baochai en *Sueño en el pabellón rojo* a través de corpus Lili Wang & Yanli Zhang	
10:30-11:00	Deconstructing news narratives of child sexual abuse: Unveiling the underlying boundaries Nuria Lorenzo-Dus & Sergio Maruenda Bataller	
09:30-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 4
Corpus, estudios contrastivos y traducción - Corpora, contrastive studies and translation Panel moderado por/ Panel chaired by: **Marlén Izquierdo Fernández**		
10:30-11:00	Repeat or diversify? A multi-factorial study of English-to-Polish translation of reporting verbs in literary novels Łukasz Grabowski	
09:30-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 13
Variación y cambio lingüístico basados en corpus - Linguistic variation and change through corpora Panel moderado por/ Panel chaired by: **Elena Quintana Toledo**		
09:30-10:00	Prescriptivism and pronominal case variation from 1710 to 1920 Miriam Criado-Peña	
10:00-10:30	Morphosyntactic and pragmatic variation in *if/si*-constructions: A corpus-based analysis of English and Spanish newspaper discourse Cristina Lastres-López	
10:30-11:00	On generalized *-s* in nineteenth-century representations of south-western dialects Javier Ruano	
10:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 14
Lingüística computacional basada en corpus - Corpus-based computational linguistics Panel moderado por/ Panel chaired by: **María Chantal Pérez Hernández**		
10:00-10:30	Climate change impact on global health Stefania M. Maci	

10:30-11:00	Evaluating Large Language Models (LLMs) in Annotating Specialized Opinion Texts: A GPT-Fueled Analysis of Functional Discourse Units (FDUs) Javier Fernández-Cruz, Carla Fernández-Melendres & Irina Muñoz Toala	
09:30-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 15
Corpus, adquisición y enseñanza de lenguas - *Corpora, language acquisition and teaching* Panel moderado por/ Panel chaired by: **Carmen Maíz Arévalo**		
09:30-10:00	Hedging in academic writing. A corpus-based analysis of Clinical and Experimental Medicine MA and PhD theses in the *MoreThesis Corpus* Marina Bondi & Silvia Cavalieri	
10:00-10:30	Vocabulary additions in teacher talk: What kinds of words do language instructors add to the textbook? Nausica Marcos Miguel, Silvia Aguinaga Echeverría & Oihane Muxika Loitzate	
10:30-11:00	Automatic Readability Analysis of Small-range Corpus Based on Native Chinese Textbooks for Junior High School Lizhen Hao & Yiyi Zhao	
09:30-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 18
Usos y aplicaciones específicas de la lingüística de corpus - *Special Purposes and Corpus Linguistics* Panel moderado por/ Panel chaired by: **María José Marín Pérez**		
09:00-09:30	Using corpora to inform sociological studies of translation Paolo Canavese	
09:30-10:00	How close are we to guilt-free flying? sustainability in airlines' disclosures. Is the provision of data enough to guarantee transparency? Franca Poppi & Judith Turnbull	
10:00-10:30	Estrategias de atenuación en terapias de pareja Josefa Contreras-Fernández	
10:30-11:00	English travels by train: a corpus linguistic analysis of an Italian on-board magazine Valentina Di Francesco	

11:00-11:30	**REFRIGERIO - *COFFEE BREAK*** PATIO EDIFICIO DEPARTAMENTAL - *DEPARTMENTAL BUILDING COURTYARD*

11:30-11:50	**PRESENTACIÓN DE SESIÓN PLENARIA - PLENARY SESSION INTRODUCTION** SALÓN DE ACTOS EDIFICIO PRINCIPAL DE HUMANIDADES Prof. Francisco Javier Martín Arista, Universidad de La Rioja Plan de Transformación de la Universidad de La Rioja
11:50-12:50	**SESIÓN PLENARIA - PLENARY SESSION** SALÓN DE ACTOS EDIFICIO PRINCIPAL DE HUMANIDADES **Corpus exploitation for Natural Language Processing: what matters more? Quality or quantity? Human-crafted rules or Artificial Intelligence?** Prof. Ruslan Mitkov University of Lancaster

12:55-13:55	**SESIÓN 2 - *SESSION 2*** MAIN BUILDING	AULA - *ROOM*: *2*
Diseño, elaboración y tipología de corpus - *Corpus design, compilation and types* Panel moderado por/ Panel chaired by: **Luis Puente-Castelo**		
12:55-13:25	Experimenting a constructivist approach to annotation Edmond Cane	

13:25-13:55 **PÓSTER**	Repositorios y cuestiones extralingüísticas de los materiales cronísticos áureos para la creación de un corpus M. Teresa Cáceres Lorenzo, Anabel Mederos Cedrés & Yaiza Santana Alvarado Inteligencia artificial para textos del siglo XVI: resultados de una investigación en la acción José Ignacio Salas-Cáceres & Yaiza Santana Alvarado Análisis visual de las investigaciones sobre el *Corpus paralelo de Sueño en el pabellón rojo* en la plataforma CNKI Lili Wang, María Teresa Cáceres-Lorenzo & Yanli Zhang	
12:55-13:55	**SESIÓN 2 - *SESSION 2*** MAIN BUILDING	AULA - *ROOM*: 3
Discurso, análisis literario y corpus - Discourse, literary analysis and corpora Panel moderado por/ Panel chaired by: **Giovanni Garofalo**		
12:55-13:25	A corpus-assisted analysis of intersectional representations of parenthood in migration-themed picture books Izaskun Elorza & María Birlea	
13:25-13:55	A Corpus-Linguistic Approach to Gendered Negations in Science Fiction: Cross-Linguistic Patterns in English and Chinese Yaqin Wang & Xinying Chen	
12:55-14:55	**SESIÓN 2 - *SESSION 2*** MAIN BUILDING	AULA - *ROOM*: 4
Usos y aplicaciones específicas de la lingüística de corpus - *Special Purposes and Corpus Linguistics* Panel moderado por/ Panel chaired by: **María José Marín Pérez**		
12:55-13:25	The relation between metaphor and evaluation revisited Radoslava Trnavac & Kattie Patterson	
13:25-13:55	Análisis de las dificultades de uso de los verbos con preposición del alemán en el contexto de aprendizaje hispanohablante mediante el corpus de aprendices MERLIN Alexander Gahr	

16:00-17:30	**SESIÓN 3 - *SESSION 3*** MAIN BUILDING	AULA - *ROOM*: B02
Discurso, análisis literario y corpus - Discourse, literary analysis and corpora Panel moderado por/ Panel chaired by: **Mercedes Cabrera**		
16:00-16:30	Análisis contrastivo del uso de partículas aproximadoras y sus valores discursivo-pragmáticos en el habla coloquial de Argentina y Chile Lissette Mondaca Becerra	
16:30-17:00	Citing others in the Coruña Corpus Margarita Mele Marrero	
17:00-17:30	Dependency Directions as a Stylometric Tool: Distinguishing Genres in Czech through Syntactic Analysis Miroslav Kubát & Xinying Chen	
16:00-17:30	**SESIÓN 3 - *SESSION 3*** MAIN BUILDING	AULA - *ROOM*: B03
Estudios gramaticales basados en corpus Panel moderado por/ Panel chaired by: Elena Quintana Toledo		
16:00-16:30	Allostructions along the Dynamic Model: Lexical specificity in the variable expression of pronominal subjects Ivan Tamaredo Meira	
16:30-17:00	Exploring Colloquialization in English as a Lingua Franca (ELF): Multivariate Analysis of Necessity Modals in Spoken and Written ELF Chunyuan Nie	
17:00-17:30	Words are syntactically distributed for efficient use: Evidence from syntactic neighborhood density Phillip G. Rogers	

16:00-17:30	**SESIÓN 3 - *SESSION 3*** MAIN BUILDING	AULA - *ROOM*: B04
Lexicología y lexicografía basadas en corpus - Corpus-based lexicology and lexicography Panel moderado por/ Panel chaired by: **Margarita Sánchez**		
16:00-16:30	Metáforas conceptuales y marcos semánticos: análisis de la percepción de los movimientos sociales franceses en el corpus de prensa español Estéfano Rodríguez-Peláez	
16:30-17:00	"You'll really appeal to your customers and *push out* your products." The phrasal verb use in European student business case competitions Siyang Zhou & Xinyue Zhang	
17:00-17:30	Aportaciones de los corpus a la elaboración de un vocabulario fundamental del español Isabel Sánchez López	

17:30-18:00	**REFRIGERIO - *COFFEE BREAK*** PATIO EDIFICIO DEPARTAMENTAL - *DEPARTMENTAL BUILDING COURTYARD*

18:00-19:30	**SESIÓN 4 - *SESSION 4*** MAIN BUILDING	AULA - *ROOM*: B04
Discurso, análisis literario y corpus - Discourse, literary analysis and corpora Panel moderado por/ Panel chaired by: **Margarita Sánchez**		
18:00-18:30	Mean Dependency Distance in Contemporary Czech Language - a Genre Analysis Xinying Chen & Miroslav Kubát	
18:30-19:00	La descripción del personaje en la prosa del esperpento de Valle-Inclán: aproximación desde la lingüística de corpus Andrés Ortega Garrido	
19:00-19:30	Analyzing the rhetoric of Díaz-Ayuso and Lauren Boebert's political tweets on X: A corpus-assisted analysis Mª Milagros del Saz Rubio	
18:00-19:30	**SESIÓN 4 - *SESSION 4*** MAIN BUILDING	AULA - *ROOM*: B03
Usos y aplicaciones específicas de la lingüística de corpus - *Special Purposes and Corpus Linguistics* Panel moderado por/ Panel chaired by: **Mercedes Cabrera**		
18:00-18:30	Linguistic structures and pragmatic mechanisms of persuasion in fake news Radoslava Trnavac, Nele Põldvere & Silje Susanne Alvestad	
18:30-19:00	El aspecto léxico de pacientes con Alzheimer: un análisis de corpus desde la Gramática del Papel y la Referencia Alejandro Suárez Rodríguez	
19:00-19:30	Análisis de los calcos en un corpus de textos escritos en francés por estudiantes españoles: implicaciones para la enseñanza de lenguas extranjeras en secundaria Cristina María Santana-Peñate	

19:30	**SALA DE GRADOS - EDIFICIO ANEXO** **ASAMBLEA GENERAL DE AELINCO** ***GENERAL ASSEMBLY OF AELINCO***

21:00	**CENA DE GALA - *GALA DINNER*** **HOTEL AC IBERIA LAS PALMAS** Avenida Alcalde José Ramírez Bethencourt, 8, 35003 Las Palmas de Gran Canaria, Las Palmas

	VIERNES 24 DE MAYO DE 2024 FRIDAY 24 MAY 2024

09:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 2
Discurso, análisis literario y corpus - Discourse, literary analysis and corpora Panel moderado por/ Panel chaired by: **Giovanni Garofalo**		
09:00-09:30	Navigating Work and Family: A Critical Discourse Analysis of the Representation of Women's Multiple Roles in Chinese Sina Weibo Reports Qianqian Wu	
09:30-10:00	From Ulyanov to Lenin: a corpus-based discourse analysis of Vladimir Lenin's works Mikhail Mikhailov	
10:00-10:30	Logical markers in Academic Writing Virginia Mattioli	
10:30-11:00	Gender, True Crime and Journalism: A Comparative Study of Media Portrayals in the Yorkshire Ripper Case (1975-1985) Elena Castellano-Ortolà	
10:00-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 3
Corpus, estudios contrastivos y traducción - Corpora, contrastive studies and translation Panel moderado por/ Panel chaired by: **Marlén Izquierdo Fernández**		
10:00-10:30	Nueva terminología en el campo de los videojuegos: un estudio de corpus Carmen Luján García	
10:30-11:00	Data-driven empirical translation studies: relating error annotations and metadata in a learner corpus Marlén Izquierdo	
09:30-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 4
Variación y cambio lingüístico basados en corpus - Linguistic variation and change through corpora Panel moderado por/ Panel chaired by: **Elena Quintana Toledo**		
10:00-10:30	El efecto de definitud en español: nuevos datos dialectales y viejos enfoques teóricos Jorge Agulló	
10:30-11:00	A Corpus-Based Study on the Diachronic Changes of Morphological Richness and Word Order Freedom in Romance Languages Siqi Liu & Jianwei Yan	
09:30-11:00	**SESIÓN 1 - *SESSION 1*** MAIN BUILDING	AULA - *ROOM*: 15
Usos y aplicaciones específicas de la lingüística de corpus - *Special Purposes and Corpus Linguistics* Panel moderado por/ Panel chaired by: **María José Marín Pérez**		
09:00-09:30	English travels by train: a corpus linguistic analysis of an Italian on-board magazine Valentina Di Francesco	
09:30-10:00	Análisis comparativo de corpus sobre el aspecto léxico de individuos sanos y pacientes con Alzheimer desde la Gramática del Papel y la Referencia Alejandro Suárez Rodríguez	
10:00-10:30	A corpus-assisted discourse study of museum exhibit labels Xiaoyu Xu & Cecilia Lazzeretti	
10:30-11:00	CSR and (un)transparent communication of *equality* vs. *equity*: A mini-diachronic corpus-based analysis Federico Zaupa	

11:00-11:30	**REFRIGERIO - *COFFEE BREAK*** PATIO EDIFICIO DEPARTAMENTAL - *DEPARTMENTAL BUILDING COURTYARD*

11:30-12:30	**SESIÓN PLENARIA - PLENARY SESSION** **Interpersonality in academic and digital genres** Prof. María Luisa Carrió-Pastor Universitat Politècnica València

12:30-13:30	**SESIÓN 2 - *SESSION 2*** MAIN BUILDING	AULA - *ROOM*: *24*
Discurso, análisis literario y corpus - Discourse, literary analysis and corpora Panel moderado por/ Panel chaired by: **Giovanni Garofalo**		
12:30-13:00	The construction of VAW in the US Press: A corpus-informed discursive news values analysis Miguel Fuster-Márquez	
12:30-13:00	Helping find the writer quoting. A solution to mark authorial presence in the vicinity of quotations Luis Puente-Castelo and Isabel Moskowich	

A corpus-assisted analysis of intersectional representations of parenthood in migration-themed picture books

Izaskun Elorza, University of Salamanca

Maria Bîrlea, University of Salamanca (Lindes Research Group)

Migration has received a lot of social attention lately and this has been mirrored in the publication of many picture books dealing with this topic (Hope, 2008). In these books, readers are very often presented with the narrations of children that tell their experience during the migration process from their own point of view. Although unfortunately migrant children often have to migrate unaccompanied by their parents, the complexities of their migration experiences tend not to be fully represented in picture books (e.g., Gu & Catalano, 2022). In migration-themed picture books, it is very often the case that parents accompany children and support them, comforting them, or explaining to them how their new life is going to be like, thus participating in the child's adaptation process. A relevant question that arises, when considering the need to incorporate an intersectional approach to migration narratives (Güell & Parella, 2021), is whether parents are represented in terms of gender stereotyping roles in picture books about migration. Previous work on parents' representations in picture books point to an uneven representation of them, with fathers being more 'invisible' (Anderson & Hamilton, 2005), as well as the representation of both mothers and fathers according to traditional stereotyped gender roles (Sunderland, 2011).

This paper addresses gender and migrant intersectional representations of parenthood in picture books. In order to explore such representations, we have carried out a critical corpus-assisted analysis (Baker & McGlashan, 2020; Partington, 2008) of a sample of multimodal migrants' narratives comprising 34 picture books. Parenthood has been scrutinised in this corpus by studying their visual and verbal presence/absence in the corpus, as well as the verbal interactions between parents and children. Our approach to the analysis relies on Halliday's (1978) and Kress and Van Leeuwen's (2006) combined approaches to multimodal texts.

Existing literature about parents' empathy in different contexts situate mothers as more empathetic than fathers (e.g., Garner & Parker, 2018 on parents reading picture books with their children, or Drotbohm, 2005 on migrant teenagers' feelings about their parents' empathy). In addition to the presence of different ways of referring to mothers and fathers in the corpus, as well as their relative frequencies, our study has explored concordance patterns. Our purpose was to look for patterns associated with the lemmas [mother] and [father], with special attention to the cases where they are interacting verbally with their children, either directly (direct speech) or indirectly (indirect speech).

Although there is an even distribution of the presence of [mother] and [father] in the corpus, with very few cases of books where parents are not present, results show a preference for mothers as caretakers. In line with previous research, results indicate that there is a greater tendency for mothers to be represented as supporters of their children, and children tend to go to them whenever they need answers and affection. Fathers' roles are more related to sanctioning children's behaviour. These results seem to point to a stereotyped representation of parents' gender roles in migration-themed picture books. Our analysis attempts to contribute to the study of literary representations of parenthood from a corpus linguistics perspective (e.g., Geybels, 2024). However, further studies with larger samples are needed in order to gain a better understanding of how intersectionality is realised in migration-themed picture books, and how it affects multimodal representations of parents.

References

Anderson, D. A., & Hamilton, M. (2005). Gender role stereotyping of parents in children's picture books: The invisible father. *Sex Roles, 52*(3-4), 145–151. DOI: https://DOI.org/10.1007/s11199-005-1290-8

Baker, P., & McGlashan, M. (2020). Critical discourse analysis. In S. Adolphs & D. Knight (Eds.), *The Routledge handbook of English language and the digital humanities* (pp. 220-241). Routledge.

Drotbohm, H. (2005). Small heroes. Rap music and selective belongings of young Haitian immigrants in Montreal. In J. Knorr (Ed.), *Childhood and migration: From experience to agency* (pp.155-174). Transcript Verlag.

Garner, P. V., & Parker, T. S. (2018). Young children's picture-books as a forum for the socialization of emotion. *Journal of Early Childhood Research, 16*(3), 291-304. DOI: https://DOI.org/10.1177/1476718X18775760

Geybels, L. (2024). "Weird, but lovely" A digital exploration of age in David Almond's oeuvre. In V. Joosen et al. (Eds.), *Age in David Almond's oeuvre: A multi-method approach to studying age and the life course in children's literature* (pp. 59-91). Routledge. DOI: https://DOI.org/10.4324/9781003369608-4

Gu, X. &, Catalano, T. (2022). Representing transition experiences: A multimodal critical discourse analysis of young immigrants in children's literature. *Language and Education*, *71*. DOI: https://DOI.org/10.1016/j.lingEd.2022.101083

Güell, B., & Parella, S. (2021). *Guidelines on how to include the gender perspective in the analysis of migration narratives.* Bridges: Assessing the Production and Impact of Migration Narratives and CIDOB: Barcelona Center for International Affairs. DOI: https://DOI.org/10.5281/zenodo.5040804

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning.* Edward Arnold.

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*(4), 295-304. DOI: https://doi.org//10.1007/s10583-008-9072-x

Kress, G., & Van Leeuwen, T. (2006). *Reading images: The grammar of visual design* (Second Edition). Routledge.

Partington, A. S. (2008). The armchair and the machine: Corpus-assisted discourse studies. Essay. In C. T. Torsello, K. Ackerley & E. Castello (Eds.), *Corpora for university language teachers* (pp. 189–213). Peter Lang.

Sunderland, J. (2011). *Language, gender and children's fiction.* Bloomsbury.

A corpus-assisted discourse study of museum exhibit labels

Xiaoyu Xu, The Education University of Hong Kong

Cecilia Lazzeretti, Free University of Bozen Bolzano

This study is rooted in the seismic shift of museums towards inclusivity and accessibility, whereby each exhibit label is no longer just a piece of text, but a means to create a richer, more engaging experience for everyone. Museums in Hong Kong have embraced this transformative ethos, but museum text creators are unsure how to craft English exhibit labels that resonate with diverse audiences. Lingering in the shadows are questions about the right linguistic tools to prepare a new generation of museum writers. Consequently, this research aims to empower Hong Kong's museums with essential linguistic tools and insights they need to create inclusive and accessible texts.

While previous studies on exhibit labels (e.g., Ravelli, 2006; Blunden, 2017) have primarily relied on qualitative discourse analysis approaches, this study employs a corpus-assisted discourse studies approach (CADS), integrating techniques and tools developed within corpus linguistics. This approach involves 'oscillating' (Mautner, 2007, p. 66) or 'shunting' (Partington & Marchi, 2015, p. 231) between the quantitative and qualitative components of the analysis, creating a 'useful synergy' between discourse studies and corpus linguistics (Subtirelu & Baker, 2018, p. 107). CADS enables analysts to examine a corpus as a whole rather than focusing solely on certain texts that may, by chance or on purpose, validate what one meant to convey all along. It thus refutes the 'cherry-picking' allegation often levelled against discourse studies. Furthermore, researchers appreciate the contribution that CADS makes towards 'unpacking what makes discourse tick' and 'the greater confidence it gives the analyst in interpreting the results' (Gillings et al., 2023, p. 1).

To pinpoint features engaging audiences, we collected 100 exhibit labels from one local art museum and conducted a corpus-assisted discourse study (CADS). Frequent keywords were generated using *WordSmith* 8.0, and their functions in context were analysed through close reading of concordances. We found that personified verbs (e.g., the hotel expresses…; the building helped…; its success allowed…) gave objects and abstract concepts a human or lifelike quality, making knowledge more personally relatable to the audiences. We also found that verbs depicting material processes (e.g., arrive, pass through, live) collocated with personal pronouns (e.g., we live) or human agents (e.g., guests pass through) helped to construct narratives that led the reader into an imagined world or scenario. Other interesting features included the use of evaluative language (e.g., this is not a record player) to anticipate and engage in audiences' potential reactions. These findings offer practical insights for enhancing the inclusive text-writing skills of museum professionals, informing training materials and serving as benchmarks for digital curatorial writing tools. Ultimately, the integrated inclusive strategies derived from this study aim to enhance the overall museum experience.

References

Blunden, J. (2017). The sweet spot? Writing for a reading age of 12. *The Museum Journal, 60*(3), 291-309.
Brown, K., & Mairesse, F. (2018), The definition of the museum through its social role. *Curator, 61,* 525-539.
Gillings, M., Mautner, G., & Baker, P. (2023). *Corpus-assisted discourse studies.* Cambridge University Press.
Mautner, G. (2007). Mining large corpora for social information: The case of elderly. *Language in Society, 36*(1), 51-72.
Partington, A. (2010). Modern diachronic corpus-assisted discourse studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora, 5*(2), 83-108.
Ravelli, L. (2006). Genre and the museum exhibition. *Linguistic and the Human Sciences, 2*(2), 299-317.
Subtirelu, N. C., & Baker, P. (2017). Corpus-based approaches. In J. Flowerdew & J. E. Richardson (Eds.), *The Routledge handbook of critical discourse studies* (pp. 106-119). Routledge.

A corpus-based research on the SQUARE lexical set in the dialect of nineteenth-century Lancashire

Nadia Hamade Almeida, University Camilo José Cela

The Lancashire dialect has been extensively represented in literature. *The Late Lancashire Witches* (1634), *That Lass O' Lowrie's* (1877) or *Hard Times* (1854) are instances of the Lancashire dialect representation. Regional literature is traditionally classified into dialect literature and literary dialect. The first type refers to those works that are completely written in a non-standard variety. As a result, dialect literature is mainly addressed to a limited readership: those readers who are capable of reading and understanding the vernacular variety represented. Alternatively, literary-dialect texts are largely composed in Standard English or the prestige variety except the characters' dialogues that are marked with a particular dialect. However, as literary-dialect writers were not linguists or dialect experts, they were not thoroughly rigorous and meticulous in dialect depiction as that was not their principal concern.

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 believed to be useful tools for linguists in dialect study (Ruano-García 2007, p. 111; Beal 2011, p. 204). This is because a meticulous analysis of the nonstandard spellings may help obtain phonological information of a regional variety at a

certain period. For this reason, this paper relies on these texts to examine the Lancashire vernacular variety.

As a complete insight into the Lancashire dialect would be beyond the scope of the present paper, this study focuses on sounds and spellings related to the SQUARE lexical set, according to the classification Wells (1982a, p. 155) provides for words related to RP [ɛə]. This paper attempts to set out and explain the distinct dialect pronunciations and the possible coexistence of sounds within the same group of words, considering historical and sociolinguistic reasons.

In this endeavor, a corpus compilation consisting of nineteen literary-dialect works written by five distinct authors was elaborated and manually approached. As the dialect is uniquely represented in the characters' speech, this study principally examines these dialogues. The different deviant spellings related to SQUARE were taken as primary sources and attributed to their possible sounds in the Lancashire dialect. In this regard, García-Bermejo Giner (1999, p. 252) considers that a comparison between the standard and the non-standard orthography is of great value when researchers attempt to undertake a phonological study via literary-dialect texts.

The results of this research reveal the existence of two distinct dialect sounds within the SQUARE lexical set. The findings also show that while one of the pronunciations could be a regressive form during the nineteenth century due to the scant data gathered, the second would be a stereotyped sound as it is usually attributed to the Lancashire vernacular.

References

Beal, J. (2011). *English in modern times*. Hodder Education

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.

A Corpus-based Sociolinguistic Study of Distribution Characteristics and Influencing Factors of Must and Have to in London Teen Language

Tianai Zhang, University of Oxford

'Must' and 'have to' are modal verbs that indicate obligation and necessity. Based on the Bergen Corpus of London Teenage Language (COLT), this study uses quantitative and qualitative methods to address the following three research questions.

1) To what extent do semantic features (i.e. epistemic modality and root modality) influence the distribution of ‘must’ and ‘have to’ in the London teen language?

2) To what extent do social factors (i.e. gender and locality) influence the choices of ‘must’ and ‘have to’ in the London teen language?

3) What are the distribution characteristics of ‘must’ and ‘have to’ used by London teenagers across different semantic features and social factors?

See below: Figure 1 The conceptual framework for exploring ‘must’ and ‘have to’ in the London teen language

The conceptual framework displayed in the figure above provides a lens through which variables should be explorEd. COLT contains examples of the natural spoken language of individuals aged between 1- 59 years old, but most speaker data in the corpus is from teenagers aged between 13-17 years old. The data was collected and processed through the retrieval functions of COLT, Excel, and SPSS. The main findings are summarised as follows. First, in the spoken language of London teenagers, semantic features significantly impacted the distribution of the two modal verbs. ‘Must’ was used more to express epistemic modality, with ‘have to’ instead used to express the root modality. Second, gender made no significant impact on the use of ‘must’ and ‘have to’ by London teenagers. Even considering data from the 21-59 age group in COLT, the effect of gender was negligible. Third, the socioeconomic status of the locality participants resided in had a significant influence on teenagers’ choices, which had not been observed in previous studies based on all age groups of London speakers. Upper-class London teenagers were more likely to use ‘must’, while middle-class teenagers were more likely to use ‘have to’. Teenagers of lower socioeconomic status showed similar patterns in the usage of ‘must’ and ‘have to’, but without obvious preference.

Gender and locality can shape the power structure in society, while language usage is closely related to power dynamics and cultural meanings. Furthermore, the results of this study support the argument that modal verbs in native English varieties are experiencing democratisation. The findings show that at least in the British capital, the modals ‘must’ and ‘have to’ do not reflect the existing power inequalities between genders. However, modal verbs used by people from different localities reinforce the social stratification based on socioeconomic hierarchies in London.

As an interdisciplinary study of teen language, this study benefits individuals who are interested in the intersection of linguistics and society, researchers in linguistics, and language teachers. Through looking at the results of this study, language enthusiasts can learn more about the relationship between certain patterns of language usage and demographic background. Additionally, the study forms a strong informative basis for further investigation studies on including social factors in theoretical or corpus linguistics. Finally, for teachers who teach English as a second or foreign language, understanding the use of

modal verbs by native speakers can be a valuable reference in the teaching of non-native speakers.

A Corpus-Based Study on the Diachronic Changes of Morphological Richness and Word Order Freedom in Romance Languages

Siqi Liu, Hangzhou Normal University

Jianwei Yan, Zhejiang University

The diachronic changes of morphosyntactic features from Latin to Romance is a topic of great interest in the field of linguistics (e.g., Schwegler, 1990; Ledgeway, 2012; Liu & Xu, 2012), since the relationship between Romance languages and their common ancestor, Latin, provides an excellent opportunity to study how language changes over time. However, while there have been some explorations of these typological changes (e.g., Gulordava & Merlo, 2015; Haspelmath & Michaelis, 2017), there is still a noted lack of quantitative investigation of the overall features of morphological richness and word order freedom. The current study adopts the large-scale corpora of Universal Dependencies (UD) with morphological and syntactic annotations to delve into how the morphology of Romance languages has changed compared to Latin and what the observed diachronic drifts are in word order freedom within this language family.

Specifically, Latin and 7 Romance languages (8 languages, 23 corpora in total) are extracted from UD as the main focus. These 7 Romance languages belong to subbranches of Ibero-Romance (Spanish, Portuguese, Galician, Catalan), Gallo-Romance (French), Italo-Romance (Italian), and Eastern Romance (Romanian). In addition, 4 corpora of Chinese (both Modern and Classical) are used as the benchmark of a typical analytic language. Moreover, the metrics of moving-average mean size of paradigm (MAMSP) based on the forms and lemmas of each corpus and cosine similarity of constituent orders (COSS) based on main clauses and subordinate clauses (Author, 2021) are utilized to quantify the linguistic features under discussion.

Firstly, the quantitative analysis reveals significant diachronic changes in morphology from Latin to Romance languages. Latin had a highly inflectional system with extensive case marking on nouns and verbs conjugated for person, number, tense, mood, etc., while Romance languages have reduced or lost some inflectional categories altogether. This reduction can be seen through decreased form counts for lemmas compared to Latin.

In addition, the quantitative methods provide evidence for the observation that syntax underwent significant changes during the evolution from Latin to Romance languages.

Whereas Latin had relatively free word order due to its rich case system allowing for flexible constituent placement within sentences; most Romance languages exhibit more rigid subject-verb-object (SVO) order in main clauses and even extending to subordinate clauses as well.

Furthermore, a diachronic examination of the evolution of Romance languages from Latin reveals a trade-off correlation between morphological complexity and syntactic flexibility. The Spearman's rank correlation coefficient between MAMSP and COSS (for both main and subordinate clauses) is moderate, and statistically significant, and the Spearman's rank correlation coefficient between MAMSP and COSS (for main clauses only) is also moderate, and statistically significant. It means that, as Latin developed into Romance languages, its rich system of morphological markings was progressively simplified, while word order became increasingly rigid. In other words, the diachronic analysis provides evidence for a compensatory relationship between morphological and syntactic change in the transition from synthetic Latin to analytic Romance.

The findings highlight the importance of examining linguistic evolution from a diachronic perspective and how quantitative analysis can aid in uncovering patterns and trends over time. Further research could explore other aspects of language change within this language family or compare it with other language families to gain more insights into the broader dynamics of linguistic evolution.

References

Gulordava, K., & Merlo, P. (2015). Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and ancient Greek. In *Proceedings of the Third International Conference on Dependency Linguistics* (pp. 121-130). Depling.

Haspelmath, M., & Michaelis, S. M. (2017). Analytic and synthetic. In *Language Variation-European Perspectives VI: Selected Papers from the Eighth International Conference on Language Variation in Europe* (pp. 3-21).

Ledgeway, A. (2012). From Latin to Romance: Morphosyntactic typology and change, vol. 1. Oxford University Press.

Liu, H., & Xu, C. (2012). Quantitative typological analysis of Romance languages. *Poznań Studies in Contemporary Linguistics, 48*(4), 597-625.

Author. (2021). Morphology and word order in Slavic languages: Insights from annotated corpora. *Вопросы языкознания,* (4), 131-159.

A Corpus-Linguistic Approach to Gendered Negations in Science Fiction: Cross-Linguistic Patterns in English and Chinese

Yaqin Wang, Guangdong University of Foreign Studies/University of Vienna

Xinying Chen, University of Ostrava

Negation, as a linguistic feature, has been one of the subjects of interest in gendered language studies, traditionally indicating a higher prevalence in female discourse as one of the markers of indirectness (Mulac, 1986; Mulac et al., 2001). Contrarily, recent analyses, such as those by Hancock et al. (2015), challenged this view, suggesting that negations do not significantly strengthen the model of gender perception. Most of those earlier discussions, furthermore, often relied on experimental approaches that measure perceptions of femininity and masculinity (e.g., Hancock & Rubin, 2015), rather than empirical corpus-linguistic methods that incorporate linguistic features.

Addressing this problem, our study adopts a corpus-linguistic framework to explore gendered language patterns through a keyword analysis of dependency types, particularly focusing on negations. This syntactic approach based on dependency grammar (De Marneffe & Nivre, 2019) examines binary word relationships that reveal the underlying structure within sentences. Based on a corpus of science fiction texts from female and male authors, the study contrasted works in English and Chinese. Specifically, the corpus contains of 47 English and Chinese novels recognized by the Hugo and Xingyun science-fiction awards, offering a balanced representation of gendered writing in both languages.

The research questions are:

1) Does a corpus-based linguistic analysis reveal differences in the use of negations between female and male writers across languages?

2) If yes, what are the patterns and distributions of negations in gendered texts?

Employing keyword analysis via the software *LancsBox X* (Brezina & Platt, 2023) and examining specific types of negations, we delved into the nuanced use of negations by authors of different genders. Results indicate a more pronounced presence of the dependency relation ‘neg’, signifying negation expressions, in texts by female authors—a pattern consistent across both the English and Chinese corpora. Different from Hancock & Rubin (2015)’s study to some degree, this finding hints at a crosslinguistic parallel in the usage of negations of female writers. Further, we juxtaposed the specific distributions of negations in gendered texts and deliberated on potential underlying factors influencing these patterns. By integrating a corpus-linguistic methodology with a cross-linguistic perspective, our study endeavors to shed some light on the intricate dynamics of gendered language use and its manifestations in negation across different cultural and linguistic contexts.

References

Brezina, V., Platt, W. (2023). *#LancsBox X* 3.0.0 [software package], lancsbox.lancaster.ac.uk

De Marneffe, M. C., & Nivre, J. (2019). Dependency grammar. *Annual Review of Linguistics, 5,* 197-218.

Hancock, A. B., Stutts, H. W., & Bass, A. (2015). Perceptions of gender and femininity based on language: Implications for transgender communication therapy. *Language and Speech, 58*(3), 315-333.
Mulac, A., & Lundell, T. L. (1986). Linguistic contributors to the gender-linked language effect. *Journal of Language and Social Psychology, 5*(2), 81-101.
Mulac, A., Bradac, J. J., & Gibbons, P. (2001). Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences. *Human Communication Research, 27*(1), 121-152.

A Multidimensional Analysis of British Legal Genres: Statute Law vs. Case Law

Daniel Granados Meroño, University of Murcia

Many scholars use the methodology first developed by Douglas Biber, known as multidimensional (MD) analysis, to contrast English oral and written discourse (1988). This method uses factor analysis to reduce variability, clustering the variables provided in 'factors' and showing the correlation between the variables in each factor. This has been extensively used to explore the variation in different types of discourse, such as Twitter (Clarke, 2022), academic language (Alamri, 2023; Biber, 2006; Pérez-Guerra & Smirnova, 2023) or literature (Grieve, 2023). This method is useful when dealing with a phenomenon probably explained by a high number of variables, being very likely correlated between each other. This is precisely the case of language phenomena (Biber, 1988), and that is why, since the implementation of it by Biber, it has been replicated in so many scopes of study.

However, no extensive study on legal discourse has ever used MD analysis, which is the purpose of this study. By applying Biber's MD Analysis to legal English, this study is aimed at testing (1) whether the conclusions regarding legal English discourse exist in studies on legal English from translation and genre studies, as well as discourse analysis approaches (Alcaraz & Hughes, 2002; Álvarez Álvarez, 2008; Bhatia, 2014; Goźdź-Roszkowski, 2011; Macías Otón, 2013; Moneva, 2013; Trosborg, 1995) are confirmed and (2) whether there are differences between two of the most relevant legal genres, judgments and legislation (acts or statute law).

This study used the BLRC corpus on British judicial decisions (Marín & Rea Rizzo, 2012) and a corpus on British legislation (BSLC) compiled *ad hoc*. This legal corpus of almost 14 million words contains legislation from the four main legislative bodies of the UK (the House of Commons, the Scottish Parliament, the Northern Ireland Assembly, and the Senedd), which makes it a very valuable source of linguistic data for teachers, researchers or legal professionals interested in the structure, idioms, vocabulary, or recurrent topics found in British statute law.

The factor analysis performed in R clustered the 67 linguistic variables obtained from the corpora into six factors (to be later interpreted as textual dimensions). The variables with the highest weights in the factors obtained were the following: factor 1 showed positive weights on predicative adjectives, THAT verb complements and BE as a main verb, and showed negative weights on independent clause coordination and nominalisations; factor 2 showed positives weights on amplifiers, adverbs, demonstrative pronouns and first person pronouns, and negative weights on nouns; factor 3 had positive weights on conditional adverbial subordination and present tense, while the average noun length, phrasal coordination and nouns were negatively weighted; factor 4 had average word length, attributive adjectives and phrasal coordination positively weighted and demonstratives on the negative side; factor 5 showed positive weights on past tense and third person pronouns, and negative weights on present tense and nouns; finally, factor 6 only had positive weights, which were BE as a main verb, predicative adjective, and present tense.

A t-test showed significant differences between the factor scores loaded in the judicial decisions and legislation corpora.

References

Alamri, B. (2023). A multidimensional comparative analysis of MENA and international English research article abstracts in applied linguistics. *SAGE Open, 13*(1), 21582440221145669. DOI: https://DOI.org/10.1177/21582440221145669

Alcaraz, E., & Hughes, B. (2002). *Legal translation explained,* vol. 4. Routledge.

Álvarez Álvarez, S. (2008). Elementos cohesivos en el lenguaje jurídico: Análisis contrastivo de las sentencias judiciales en lengua inglesa y española. *La traducción del futuro: mediación lingüística y cultural en el siglo XXI,* vol. 1 (pp. 407-418). DOI: https://dialnet.unirioja.es/servlet/articulo?codigo=5660324

Bhatia, V. K. (2014). *Analysing genre: Language use in professional settings.* Taylor & Francis.

Biber, D. (1988). Variation across speech and writing. In *variation across speech and writing.* DOI: https://DOI.org/10.1017/cbo9780511621024

Biber, D. (2006). University language. En *Scl.23*. John Benjamins Publishing Company.

Clarke, I. (2022). A Multi-dimensional analysis of English tweets. *Language and Literature, 31*(2), 124-149. DOI: https://DOI.org/10.1177/09639470221090369

Goźdź-Roszkowski, S. (2011). *Patterns of linguistic variation in American legal English: A corpus-based study.* Peter Lang.

Grieve, J. (2023). Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory, 19*(1), 47-77. DOI: https://DOI.org/10.1515/cllt-2022-0040

Macías Otón, E. (2013). Las expresiones binomiales en el lenguaje jurídico y su traducción en el aula de terminología (español-inglés/francés). *Paremia, 22,* 209-225.

Marín, M. J., & Rea Rizzo, C. (2012). Structure and design of the British law report corpus (BLRC): A legal corpus of judicial decisions from the UK. *Journal of English Studies, 10,* 131-145. DOI: https://DOI.org/10.18172/jes.184

Moneva, M. A. R. (2013). Cognition and context of legal texts: Spanish and English judgments compared. *Revista de Lingüística y Lenguas Aplicadas, 8,* 76-92.

Pérez-Guerra, J., & Smirnova, E. A. (2023). How complex is professional academic writing? A corpus-based analysis of research articles in «hard» and «soft» disciplines. *Vigo International Journal of Applied Linguistics, 20.* DOI: https://DOI.org/10.35869/vial.v0i20.4357

Trosborg, A. (1995). Statutes and contracts: An analysis of legal speech acts in the English language of the law. *Journal of Pragmatics*. https://DOI.org/10.1016/0378-2166(94)00034-C

A relevance-theoretic account of the use of emoji in Twitter for Research Dissemination Purposes

Silvia Murillo Ornat, University of Zaragoza

As well as their websites, international projects use social media in their dissemination and communication activities, specifically Facebook and Twitter/X (Gertrudix et al. 2021). In the Twitter/X accounts associated to their websites, tweets are relayed with several purposes (for instance, live reporting of conferences, and giving visibility to their research meetings and to their academic publications), and emoji are often introduced in the messages, with different functions. These graphic elements can be attitudinal or speech act clues for the intended interpretation of the messages, or they can contribute to the actual content of the messages (Scott 2022, Yus 2022).

In this paper I will examine a subset of the EUROPROtweets database (Pascual et al., 2020) that comprises the tweets of 10 Twitter accounts associated to H2020 research websites, coded through NVivo. In a corpus such as this it is possible to find enough examples to display the versatility of the emoji in the construction of messages in research communication processes. The emoji used in the messages are in general associated to the keywords of the projects, and their activities, contributing to the dissemination of their results and to their promotion. Bearing in mind this contextual information, I intend to advance a corpus-driven comprehensive explanation of how the emoji are used in order to build the messages, from a relevance-theoretic perspective. Emoji can be used instead of words, or they can be used next to words in a process that resembles verbal reformulations (words can be translated into emoji which appear afterwards, and also the other way round, emoji can be introduced first and then expressed in words); in fact, it is possible to identify examples with emoji in typical reformulation processes such as explanation, identification, specification and even conclusion or consequence, cf. Author. In my approach, I will draw from the notion of perceptual resemblance as used by Sasamoto (2022) and take into account Scott's (2022) and Yus's (2022) explanations for some of the uses of the emoji, along with some insights from previous accounts of reformulation processes in verbal language (Blakemore, 1996; Author).

For some authors, these latter 'reformulation' uses of emoji are redundant (Panckhurst & Frontini, 2020), or colourful (Siever & Siever, 2020) additions, unnecessary to understand the content of the text. On the other hand, for Logi & Zappavigna (2021), these occurrences form part of 'ideational emoji', interacting with text in order to create meaning. However, I will argue that these instances are ideational and interpersonal at the same time. Essentially, these emoji patterns display a mixture of reformulation processes and modal reinforcement, also present in some verbal reformulations; the emoji are not used in order to clarify the

messages, but they reinforce them and contribute to their engagement (c.f. Völkel et al., 2019). Some examples from the corpus also point to the polyfunctionality of emoji, as they may include additional discourse roles, such as topicalizing and discourse organizing functions.

References

Blakemore, D. (1996). Are apposition markers discourse markers? *Journal of Linguistics, 32*(2), 325-347.

Gertrudix, M., Rajas M., Romero-Luis J., & Carbonell-Alcocer, A. (2021). Comunicación científica en el espacio digital. Acciones de difusión de proyectos de investigación del programa H2020. *Profesional de la información/ Information Professional, 30*(1). DOI: https://DOI.org/10.3145/epi.2021.ene.04

Logi, L., & Zappavigna, M. (2021). A social semiotic perspective on emoji: How emoji and language interact to make meaning in digital messages. *New Media & Society, 0*(0). DOI: https://DOI.org/10.1177/14614448211032965

Panckhurst, R., & Frontini, F. (2020). Beyond the binary: Emoji as a challenge to the image-word distinction. In C. Thurlow, C. Dürscheid and F. Diémoz (Eds.), *Visualizing digital discourse: interactional, institutional and ideological Perspectives* (pp. 81-103). De Gruyter Mouton.

Pascual, D., Mur-Dueñas, P., & Lorés, R. (2020). Looking into international research groups' digital discursive practices: Criteria and methodological steps in the compilation of the EUROPRO digital corpus. *Research in Corpus Linguistics, 8*(2), 87-102. DOI: https://DOI.org/10.32714/ricl.08.02.05

Sasamoto, R. (2022). Perceptual resemblance and the communication of emotion in digital contexts: A case of emoji and reaction GIFs. *Pragmatics, 33*(3), 393-417. DOI: https://DOI.org/10.1075/prag.21058.sas

Scott, K. (2022). *Pragmatics online*. Routledge.

Siever, C. M., & Siever, T. (2020). Emoji-text relations on Instagram: Empirical corpus studies on multimodal uses of the iconographetic mode. In H. Stoeckl, H. Caple & J. Pflaeging (Eds.), *Shifts towards image-centricity in contemporary multimodal practices* (pp. 177-203). Routledge.

Völkel, S. T., Buschek, D., Pranjic, J., & Hussmann, H. (2019). Understanding emoji interpretation through user personality and message context. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI '19, New York, NY, USA.* Association for Computing Machinery.

Yus, F. 2022. *Smartphone communication. Interactions in the app ecosystem*. Routledge.

Allostructions along the Dynamic Model: Lexical specificity in the variable expression of pronominal subjects

Iván Tamaredo Meira, Complutense University of Madrid

Recent research on World Englishes has uncovered differences among dialects at the interplay of words and constructions (Mukherjee & Gries, 2009; Röthlisberger et al., 2017). Hoffmann (2014; 2021) put forward the Dynamic Model Productivity (DMP) hypothesis, which posits that the level of lexical specificity in speakers' constructional representations

fluctuates across varieties, mirroring their position in Schneider's (2007) Dynamic Model: speakers of less developed varieties tend to rely more on constructions that are partially lexically filled rather than on wholly schematic ones (see also Brunner & Hoffmann, 2020). However, only constructions which are not in competition with other alternative expressions have been investigated so far, so it is still unclear how syntactic alternations fit into the DMP hypothesis. The goal of the present paper, therefore, is to investigate whether this hypothesis can be extended from constructions to allostructions, that is, "variant structural realizations of a construction that is left underspecified" (Capelle, 2006, p. 18). In other words, the goal is to ascertain whether speakers of less developed varieties in Schneider's Dynamic Model also rely more on specific lexical items when they have to choose between competing syntactic structures.

To this purpose, the present study focuses on one syntactic alternation, namely the variable expression of subject pronouns (e.g. Schröter, 2019). This alternation can be formally (and schematically) represented as [(PRO +) VERB]: PRO stands for the pronominal subject and VERB represents the main verb of the clause. Speakers, however, sometimes omit the pronominal subject (thus the parentheses), which results in clauses without an overt subject. These two competing syntactic structures are illustrated in (1) and (2):

(1) [PRO + VERB]: Apple keeps bringing out incremental updates and upgrades for all its devices. It holds various events almost every quarter […]. (GloWbE, IN B, techinfolite.com)

(2) [VERB]: I enjoyed reading this, very well written and done. Showed some points which need to be talked about […]. (GloWbE, AU B, footyalmanac.com.au)

Instances of the two allostructions were retrieved from the Corpus of Global Web-Based English (GloWbE; Davies, 2013), in particular from the Australian, Bangladeshi, Canadian, Indian, Jamaican, Nigerian, Pakistani, and Singaporean national components, which contain data from English varieties in different phases of Schneider's Dynamic Model. A random sample of 3rd person null and overt pronominal subjects in initial clause position was extracted from the corpus and annotated for the lemma of the main verb of the clause, together with the GloWbE component they were retrieved from. Then, the data was analyzed by means of distinctive collexeme analysis (DCA; Gries & Stefanowitsch, 2004), a method which measures the degree of association between words and two or more competing variants of a construction. In this case, DCA was used to measure the degree of association of specific verbs in the VERB slot with either the [PRO + VERB] or [VERB] allostructions.

Preliminary results indicate that, on the whole, speakers of less advanced varieties also tend to rely more on specific lexical items when they have to choose between two alternative syntactic structures. This provides further support for the DMP hypothesis and shows that its scope of application should be extended from (simple) constructions to allostructions.

References

Brunner, T., & Hoffmann, T. (2020). The way construction in world Englishes. *English World-Wide*, *41*(1), 1–32.
Cappelle, B. (2006). Particle placement and the case for 'allostructions.' *Constructions.*
Davies, M. (2013). *Corpus of Global Web-Based English.* https://www.englishcorpora.org/glowbe/
Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics, 9*(1), 97–129.
Hoffmann, T. (2014). The cognitive evolution of Englishes: The role of constructions in the Dynamic Model. In S. Buschfeld, T. Hoffmann, M. Huber, & A. Kautzsch (Eds.), *The evolution of Englishes: The dynamic model and beyond* (pp. 160–180). John Benjamins.
Hoffmann, T. (2021). *The cognitive foundation of post-Colonial Englishes: Construction Grammar as the cognitive theory for the dynamic model. Elements in world Englishes.* Cambridge University Press.
Mukherjee, J., & Gries, S. T. (2009). Collostructional nativization in new Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide, 30*(1), 27–51.
Röthlisberger, M., Grafmiller, J., & Szmrecsanyi, B. (2017). Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics, 24*(2), 413–440.
Schneider, E. (2007). *Postcolonial English: Varieties around the world.* Cambridge University Press.
Schröter, V. (2019). *Null subjects in Englishes: A comparison of British English and Asian Englishes.* De Gruyter Mouton.

An exploration into modality and intonation in peer interaction

Mercedes Cabrera-Abreu, University of Las Palmas de Gran Canaria

Eva Estebas-Vilaplana, National University of Disatnce Education

A fall-to-mid pitch intonation in Spanish has traditionally been associated with meanings like 'uncertainty', 'undefined idea' (Navarro Tomás, 1944; XX and Prieto, 2008; Prieto and Roseano, 2010); 'incompleteness', 'continuation', 'politeness', 'command' (XX and Prieto, 2008; Prieto and Roseano, 2010); 'incomplete', 'inconclusive' (Labastía, 2018). As for their presence in syntactic constructions, these include: a subject followed by a non-restrictive relative clause, enumerations, some coordinated phrases, and before an appositional phrase (Navarro Tomás, 1944). These descriptions point towards a high frequency use of this fall-to-mid tone in tightly controlled speech, readings and context-offered elicitations. Yet, further research is still necessary to understand better the reasons behind its presence, both in terms of its syntactic distribution and its associated meanings. For the present study, we selected a sub-corpus of dialogues from the corpus *Glissando* (Garrido *et al.*, 2011), and identified acoustically all the instances of the fall-to-mid tone. The dialogues were produced by female and male participants, all of whom were speakers of European Standard Peninsular Spanish. Each conversation was maintained by speakers who shared a certain degree of familiarity so as to obtain a high level of naturalness. A total of 47 minutes of speech were analysed. Following the methodology typically adopted in studies of corpus linguistics, we tagged instances of modal meanings manifested in the use of the above specified tone to later compute the frequency and distribution of word categories bearing the tone as well as retrieve concordances. This allowed us to identify and qualify the modal meanings expressed by the presence of the fall-to-mid tone. At this stage of the inquiry, the

results seem to suggest that the fall-to-mid tone can be mapped into a wide range of word categories and syntactic constructions to express deontic and epistemic modality. Through this form of intonation, speakers presumably convey a series of interpersonal meanings that highlight their intentions, especially towards mitigation and politeness.

References

Estebas-Vilaplana, E. & Prieto, P. (2008). La notación prosódica del español: una revisión del Sp_ToBI. *Estudios de Fonética Experimental, 17*, 265-283.

Garrido, J. M., Aguilar, L., & Escudero, D. (2011). GLISSANDO, un corpus de habla anotado para estudios prosódicos en catalán y en español. In A. Hidalgo, Y. Congosto & M. Quilis (Eds.) *El estudio de la prosodia en España en el siglo XXI. Perspectivas y ámbitos. Quaderns de Filologia 75* (pp. 321-332). Universitat de València.

Labastía, L. (2018). *Entonación y estructura informativa en el español rioplatense* [Doctoral thesis: Universidad Nacional de Educación a Distancia].

Navarro Tomás, T. (1974). *Manual de entonación española.* Guadarrama (First edition 1944). Spanish Institute in the United States.

Prieto, P., & Roseano, P. (Eds.). (2010). *Transcription of intonation of the Spanish language.* Lincom Europa.

Análisis contrastivo del uso de los verbos modales deónticos entre dos libros de recetas escritos por mujeres en el siglo XIX: “A Lady” (1818) y Beeton (1875)

Isabel Soto Déniz, University of Las Palmas de Gran Canaria

Las recetas han sido portadoras de conocimiento tanto en el ámbito de la cocina como de la medicina desde hace siglos, además de desempeñar un papel crucial en la transmisión de cultura y dinámicas sociales a lo largo del tiempo. Este trabajo se centra en el análisis de textos producidos en la Inglaterra del siglo XIX, específicamente en el análisis de verbos modales deónticos presentes en dos libros de recetas escritos por mujeres de la época: "A Lady" (1818) y Beeton (1875). Se lleva a cabo un análisis contrastivo de cómo estas autoras emplean estos verbos modales, frecuentes en textos instructivos como las recetas debido a su naturaleza prescriptiva. Por otro lado, los verbos modales actúan como mecanismos interpersonales que añaden perspectiva a la información proporcionada. Los objetivos de esta presentación incluyen la identificación de los verbos modales deónticos en ambos textos, la detección de posibles similitudes o diferencias en su uso, y la observación de variaciones a lo largo de los 57 años entre la publicación del primero y el segundo texto. A pesar de los cambios significativos en el siglo XIX en áreas como la industrialización, tecnología y transporte, el papel tradicional de las mujeres en la sociedad como cuidadoras del hogar y curanderas persiste. Este papel resalta la importancia del género de recetas, considerando la cocina como una tarea esencial del hogar. La presentación comienza con una breve contextualización del siglo XIX en Inglaterra, el papel de la mujer y la gastronomía de la época, así como la definición del género de las recetas en palabras de Taavitsainen (2001). Se introducen conceptos fundamentales como género, tipo de texto y

registro, de la mano de Trosborg (1997), Alonso-Almeida (2008a, 2008b, 2013) o Alonso-Almeida y Álvarez-Gil (2020). Posteriormente, se explora la naturaleza de los verbos modales, abordando la definición de modalidad y sus distintas clasificaciones, como la modalidad deóntica, epistémica y dinámica, de acuerdo con autores como Palmer (2001) o Van der Auwera y Plungian (1998). El enfoque se desplaza hacia los verbos modales deónticos ofreciendo definiciones de académicos como Hoye (1997), así como proporcionando definiciones académicas y relacionándolos con el concepto de perspectiva o *stance,* ampliamente estudiado por Álvarez-Gil (2021). Para llevar a cabo este estudio, se emplea una metodología de lingüística de corpus utilizando el *Corpus of Women's Instructive Texts in English* (COWITE). Se combina el análisis cuantitativo y la inspección visual mediante herramientas de software para examinar las ocurrencias y realizar un análisis exhaustivo. Los posibles resultados incluirán tanto similitudes como diferencias en el uso de los verbos modales deónticos, arrojando luz sobre las dinámicas sociales de las mujeres del siglo XIX y demostrando si existe o no variabilidad en su uso a lo largo de las décadas.

References

Alonso-Almeida, F. (2008a). The pragmatics of and-conjunctives in Middle English medical recipes. A relevance theory description. *Journal of Historical Pragmatics*, 9(2), 171-199. https://doi.10.1075/jhp.9.1.02alo

Alonso-Almeida, F. (2008b). The middle English medical charm: Register, genre and text type variables. *Neuphilologische Mitteilungen*, 109(1), 9-38.

Alonso-Almeida, F. (2013). Genre conventions in English recipes, 1600-1800. En M. Di Meo & S. Pennell (Eds.), *Reading and Writing Recipe Books 1550-1800*, pp. 64-94. Manchester University Press.

Alonso-Almeida, F., & Álvarez-Gil, F. J. (2020). 'so that it may reach to the jugular'. Modal verbs in early modern English recipes. *Studia Neofilologiczne*, XVI, 61–88. https://doi.org/10.16926/sn.2020.16.04

Álvarez-Gil, F. J. (2021). Authority and deontic modals in Late Modern English. En I. Moskowich, I. Lareo & G. Rioboo-Camiña (Eds.), *"All families and genera". Exploring the Corpus of English Life Sciences Texts*, pp. 249–264. John Benjamins Publishing Company.

Taavitsainen, I. (2001). Middle English recipes. Genre characteristics, text type features and underlying traditions of writing. *Journal of Historical Pragmatics*, 2(1), 85–113.

Van der Auwera, J., & Plungian, V. A. (1998). Modality's semantic map. *Linguistic Typology* 2(1), 79-124.

Análisis comparativo de corpus sobre el aspecto léxico de individuos sanos y pacientes con Alzheimer desde la Gramática del Papel y la Referencia

Alejandro Suárez Rodríguez, University of Las Palmas de Gran Canaria

Esta comunicación propone la comparación de individuos sanos y de pacientes con Alzheimer en la etapa temprana provenientes del corpus de Peraita y Grasso (2010) y desde el enfoque funcionalista de la Gramática del Papel y la Referencia (GPR; Van Valin y

LaPolla, 1997; Van Valin, 2005). Para ello, nos proponemos analizar la frecuencia y distribución del aspecto léxico o *Aktionsart* tanto de los individuos sanos como de los pacientes diagnosticados con la enfermedad de Alzheimer y luego comparar los resultados.

El corpus seleccionado fue compilado por Peraita y Grasso (2010) en su investigación neuropsicológica de los pacientes con Alzheimer. En ella, analizan el deterior léxico-semántico de pacientes españoles y argentinos a partir de seis categorías semánticas: perro, pino y manzana (seres vivos) y coche, pantalón y silla (seres no vivos; Peraita y Grasso, 2010, p. 204). El corpus muestra las transcripciones de 211 españoles y argentinos, si bien nos hemos concentrado en aquellos individuos sanos y pacientes de la primera etapa procedentes de España. Al tomar una muestra representativa de los sujetos (para un intervalo de confianza del 95 %; López-Roldán y Fachelli, 2015), encontramos que emitieron conjuntamente 5277 predicados verbales, de los que analizaremos una muestra de, al menos, 352 verbos de individuos sanos y 285 verbos de la etapa temprana.

Nuestro marco teórico para determinar el tipo de *Aktionsart* es la Gramática del Papel y la Referencia (GPR; Van Valin y LaPolla, 1997; Van Valin, 2005), la cual se trata de una teoría funcionalista que intenta explicar la relación entre la sintaxis, la semántica y la pragmática. En concreto, la representación semántica de la GPR parte del predicado verbal y adapta la clasificación del aspecto léxico o *Aktionsart* de Vendler (1967). Para determinar el tipo de *Aktionsart* que debemos asignar a los verbos, la GPR propone ocho pruebas (Van Valin, 2005; González Vergara, 2006; Cortés Rodríguez, González Vergara y Jiménez Briones, 2012; Van Valin, 2018) que, aplicadas secuencialmente, nos permiten establecer el *Aktionsart* de cada predicado verbal. Por último, aplicamos estas pruebas a las muestras del corpus de Peraita y Grasso (2010) y calculamos la frecuencia y la distribución de los diferentes *Aktionsarten* mediante estadísticos básicos: media, mediana, moda, rango, varianza, desviación típica y coeficiente de variación.

Tras el análisis de las muestras por separado y en conjunto, observamos que los estados son el tipo de *Aktionsart* más usado por los sujetos en sendos subconjuntos, seguidos de las actividades y las realizaciones activas en proporciones variables. De hecho, hay mayor cantidad de estados, actividades y realizaciones causativas en la etapa temprana que en los individuos sanos, mientras que decrece el número de logros, realizaciones y realizaciones activas. Los resultados de estas muestras apuntan a una relación entre la enfermedad de Alzheimer y el tipo de verbo, si bien estos indicios deben ser corroborados en posteriores investigaciones.

Referencias

Cortés Rodríguez, F., González Vergara, C. y Jiménez Briones, R. (2012). Las clases léxicas. Revisión de la tipología de predicados verbales. En R. Mairal Usón, L. Guerrero y C. González Vergara (Coord.), *El funcionalismo en la teoría lingüística: La gramática del papel y la referencia* (pp. 59-84). Ediciones Akal.

González Vergara, C. (2006). La gramática del papel y la referencia: Una aproximación al modelo. *Onomázein*, *14*(2), 101-140.
López-Roldán, P. y Fachelli, S. (2015). *Metodología de la investigación social cuantitativa*. Universidad Autónoma de Barcelona.
Van Valin, R. D. y LaPolla, R. (1997). *Syntax: Structure, meaning and function*. Cambridge University Press.
Van Valin, R. D. (2005). *Exploring the syntax-semantics interface*. Cambridge University Press.
Van Valin, R. D. (2018). Some issues regarding (active) accomplishments. En R. Kailuweit, L. Künkel y E. Staudinger (Ed.), *Applying and expanding role and reference grammar* (pp. 71-94). Freiburg Institute for Advanced Studies, Albert-Ludwigs-Universität Freiburg.
Vendler, Z. (1967). *Linguistics in philosophy*. Cornell University Press.

Análisis de la metáfora del personaje de Xue Baochai en Sueño en el pabellón rojo a través de corpus

Lili Wang, University of Las Palmas de Gran Canaria

Yanli Zhang, Shanghai International Studies University

El Corpus paralelo de Sueño en el pabellón rojo (chino-inglés) fue creado por la Universidad Wenli de Shaoxing (China) en el año 2010. Fue el primer corpus en China que trató la novela Sueño en el pabellón rojo.

Sueño en el pabellón rojo, una de las cuatro grandes novelas clásicas chinas, escrita por Cao Xueqin en el siglo XVIII durante la dinastía Qing, es conocida como “el Quijote chino” (Guelbenzu, 2010). También es reconocida como la gran enciclopedia de la antigua sociedad china. La novela cuenta la tragedia del amor entre los jóvenes aristócratas feudales Jia Baoyu y sus primas Lin Daiyu y Xue Baochai. Los estudios sobre esta novela en China han sido variados y diversos desde su publicación en el año 1791. Muchos investigadores consideran que Sueño en el pabellón rojo es una novela autobiográfica del autor. Entre los muchos personajes y figuras retratados en Sueño en el pabellón rojo, Xue Baochai es sin duda la más difícil de caracterizar, pues la naturaleza polifacética de su personalidad ha provocado controversia (Liu, 2003; Lu, 2020).

Hemos elegido a Xue Baochai, la segunda heroína de la obra, como sujeto de nuestro trabajo. Hemos analizado este personaje y su metáfora a través del Corpus paralelo de Sueño en el pabellón rojo (chino-inglés).

Para llevar a cabo esta investigación hemos establecido los siguientes objetivos:

a. Presentar la novela clásica Sueño en el pabellón rojo

b. Analizar la metáfora del personaje de Xue Baochai a través del uso del corpus

c. Interpretar la metáfora detrás del personaje de Xue Baochai La metáfora conceptual fue propuesta por Lakoff y Johnson en 1980.

La metáfora es omnipresente en la vida cotidiana, no solo en el lenguaje sino también en el pensamiento y la acción. Los hablantes nativos de todos los idiomas utilizan una gran cantidad de metáforas cuando se comunican sobre el mundo (Lakoff y Johnson, 1980). Lo interesante de Sueño en el pabellón rojo es que el autor es muy bueno usando metáforas para expresar implícitamente sus ideas. Muchos investigadores se han propuesto analizar y estudiar los presagios y metáforas escritas por el autor en su obra desde varios ángulos.

La metodología que hemos usado para este trabajo es la siguiente:

a. Análisis cuantitativo y cualitativo

b. Método comparativo

A través de nuestro análisis de búsqueda del corpus encontramos que, en primer lugar, la época de la que trata esta novela es la dinastía Ming en lugar de la dinastía Qing, por lo que podemos decir que no es una novela autobiográfica del autor; en segundo lugar, los tres protagonistas tienen sus propias metáforas detrás de ellos. Baoyu representa el sello de jade, que es un símbolo del poder imperial de China. Daiyu representa al último emperador de la dinastía Ming, Chongzhen, también el último emperador de la etnia Han en China. Xue Baochai, nuestro sujeto de estudio, representa a Huang Taiji, el primer emperador de la dinastía Qing. En resumen, a través de nuestro estudio, hemos encontrado que Sueño en el pabellón rojo no es solo una historia sobre la caída de la dinastía Ming, sino que también es la tragedia de la etnia Han.

Referencias

Guelebenzu, J. (2010, 16 de noviembre). El 'Quijote' chino. *El País.* https://elpais.com/diario/2010/11/06/babelia/1289005937_850215.html

Lakoff, G., & Johnson, M. (1980). Conceptual metaphor in everyday language. *The Journal of Philosophy, 77*(8), 453-486. https://www.jstor.org/stable/2025464

Ren, L. (2010). *Corpus paralelo de Sueño en el pabellón rojo (chino-inglés).* http://corpus.usx.edu.cn/hongloumeng/index.asp

Análisis de las dificultades de uso de los verbos con preposición del alemán en el contexto de aprendizaje hispanohablante mediante el corpus de aprendices MERLIN

Alexander Gahr, University of Las Palmas de Gran Canaria(docente)/ University of Santiago de Compostela(doctorando)

Los verbos con preposición, también denominados verbos con complemento de régimen (preposicional), constituyen una dificultad para los aprendices hispanohablantes que estudian alemán como lengua extranjera. Dicha dificultad se debe a razones intralinguales,

referidas a las características inherentes a los verbos con preposición del alemán, e interlinguales, es decir, las divergencias que existen entre las lenguas alemana y española. También influyen factores externos, pero el presente estudio se centra en los factores lingüísticos que serán analizados mediante el corpus de aprendices MERLIN.

El primer objetivo de esta ponencia consiste en mostrar cómo se puede utilizar un corpus como MERLIN para analizar las dificultades de uso de los verbos con preposición por parte de estudiantes hispanohablantes. Tras la búsqueda de todas las frases con verbos con preposición en el corpus, se averiguará el porcentaje de las frases correctas y las incorrectas para obtener una visión general del uso de estos verbos.

El segundo objetivo de este estudio constituye el análisis detallado de los errores relacionados con los verbos con preposición hallados en el corpus con el fin de clasificarlos y tratar de explicarlos mediante los factores intra e interlinguales mencionados anteriormente.

El presente trabajo se fundamenta principalmente en los siguientes dos pilares: las gramáticas de las lenguas alemana y española, y la lingüística de corpus. En las gramáticas se definen los verbos con preposición y los criterios para diferenciar los sintagmas preposicionales regidos por los verbos de los complementos circunstanciales. Esto es relevante a la hora de identificar correctamente los verbos con preposición en el corpus de aprendices.

En el corpus existen diferentes tipos de errores (ortográficos, léxicos y gramaticales), pero este trabajo se centra en los relacionados directamente con las características específicas de los verbos con régimen preposicional. Con el fin de averiguar posibles transferencias del español al alemán, se contrastan gramáticas y diccionarios de ambas lenguas.

Entre las gramáticas del alemán destacan las de Engel (1996), de Zifonun *et a*l. (1997), de Eisenberg (2013) y de Duden (2022), así como el *Diccionario electrónico de las valencias del alemán* (E-VALBU) del Instituto de la Lengua Alemana. Respecto al español, para contrastar ambas lenguas, se usará principalmente el *Manual de gramática* (2010) y el *Diccionario panhispánico de dudas* (2005), ambos de la Real Academia Española, y la *Gramática contrastiva* de Cartagena/Gauger (1989). La investigación de corpus se basa en la tesis doctoral de Weber (2020) y las publicaciones de Köhler (2005), Granger 2012, Granger *et al*. (2015), Lemnitzer/Zinsmeister (2015), entre otros.

En el corpus de textos redactados en alemán por estudiantes hispanohablantes se han encontrado casi cien frases que contienen verbos con preposiciones regidas. Casi la mitad de estas frases presenta al menos un error relacionado directamente con el verbo (sin incluir su ortografía, conjugación o posición), el pronombre reflexivo en el caso de los verbos reflexivos, la preposición o el caso regido por la preposición (acusativo o dativo).

Con respecto a la clasificación de los errores hallados, destacan las omisiones de la preposición o del pronombre reflexivo, seguido de la confusión de la preposición. Estos errores dejan entrever transferencias negativas del español al alemán como una de las principales causas para las dificultades de uso de estos verbos.

Referencias

Cartagena, N. y Gauger, H. (1989). *Vergleichende Grammatik Spanisch-Deutsch*, 2 vols. Duden.

Eisenberg, P. (2013). *Grundriss der deutschen Grammatik: Der Satz* (4., aktualisierte und überarbeitete Auflage). Metzler.

Engel, U. (1996). *Deutsche Grammatik* (3., korrigierte Auflage). Groos.

Granger, S. (2012). How to use foreign and second language learner corpora. En S. Gass, y A. Mackey (Eds.), *Research methods in second language acquisition. A practical guide* (pp. 7-29). Wiley-Blackwell.

Granger, S., Gilquin, G. y Meunier, F. (2015). Introduction: learner corpus research: past, present and future. En S. Granger, G. Gilquin y F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 1-6). Cambridge University Press.

Köhler, R. (2005). Korpuslinguistik – zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven. *Journal for Language Technology and Computational Linguistics, 20*(2), 1-16.

Lemnitzer, L. y Zinsmeister, H. (2015). *Korpuslinguistik. Eine Einführung* (3. Auflage). Narr.

Real Academia Española (2010). *Nueva gramática de la lengua española. Manual*. Espasa.

Real Academia Española, y Asociación de Academias de la Lengua Española (2005). *Diccionario panhispánico de dudas*. Santillana Ediciones Generales.

Weber, T. (2020). *Präpositionen und Deutsch als Fremdsprache: Quantitative Fallstudien im Lernerkorpus MERLIN.* [Tesis doctoral: Universität Mannheim].

Wöllstein, A. y Dudenredaktion (Eds.). (2022). *Duden. Die Grammatik* (10., völlig neu verfasste Auflage). Band 4. Dudenverlag.

Zifonun, G., Hoffmann, L. y Strecker, B. (1997). *Grammatik der deutschen Sprache*. Band 1–3. De Gruyter.

Recursos en línea

Leibniz-Institut für Deutsche Sprache: Wörterbuch zur Verbvalenz. *Grammatisches Informationssystem grammis.* Permalink: https://grammis.ids-mannheim.de/verbvalenz.

MERLIN Corpus | *Resources for research and practice related to foreign language learning.* Dresden University of Technology, University Tübingen, European Academy Bozen. www.merlin-platform.eu.

Análisis de los calcos en un corpus de textos escritos en francés por estudiantes españoles: implicaciones para la enseñanza de lenguas extranjeras en secundaria

Cristina María Santana-Peñate, University of Las Palmas de Gran Canaria

El presente estudio tiene como objetivo el análisis de los errores más comunes en la expresión escrita de estudiantes de educación secundaria obligatoria en Canarias en la asignatura de francés como lengua extranjera, a través del uso de la lingüística de corpus y la herramienta *AntConc*. Para realizar este estudio, se ha compilado un corpus de textos escritos por estudiantes de francés cómo lengua extranjera en diferentes cursos de educación secundaria de diversos centros educativos de Gran Canaria. Posteriormente se ha realizado un análisis detallado de los textos con la finalidad de identificar los errores más frecuentes en la expresión escrita y, en el caso de este capítulo, nos centraremos en unos de los errores más frecuentes que se han detectado y que están relacionados específicamente con el uso de calcos de la lengua materna (L1) en la expresión escrita en francés.

A partir de los análisis realizados, se han detectado diferentes tipos de calcos de la L1, como los calcos léxicos, sintácticos y semánticos, los cuales han sido identificados y clasificados de acuerdo con su frecuencia y gravedad. Entre los errores más frecuentes, se encuentran los calcos léxicos, que se refieren al uso inapropiado de palabras en francés que son similares a las de la L1, pero que no tienen el mismo significado. También se han encontrado errores sintácticos, relacionados con la estructura de las oraciones en francés y su equivalente en la L1. Asimismo, se han identificado errores semánticos, que se refieren al uso incorrecto de palabras que tienen un significado diferente en francés y en la L1. Estos errores son especialmente importantes, ya que pueden afectar significativamente la comprensión del mensaje en la escritura. En lo que se refiere a la metodología, tras la compilación del corpus y las búsquedas mediante el uso de la herramienta *AntConc*, se ha llevado a cabo un escrutinio visual de los casos para poder identificarlos y clasificarlos y en la sección de resultados proporcionaremos una selección de ejemplos para ilustrar los diferentes tipos de calcos que se han mencionado anteriormente: léxicos, sintácticos y semánticos.

Este estudio es relevante para la enseñanza del francés como lengua extranjera y sugiere la necesidad de diseñar estrategias pedagógicas específicas para abordar estos errores y mejorar la calidad de la expresión escrita en francés de los estudiantes. La toma de consciencia de los errores más comunes cometidos por los estudiantes permitirá abordarlos de manera efectiva. Por ello, se hará una propuesta metodológica para trabajar de manera efectiva los errores detectados que ayudarán a los estudiantes a evitar la transferencia de estructuras gramaticales y vocabulario inapropiados.

Análisis léxico de la influencia del francés en textos de gastronomía en lengua inglesa del siglo XIX escritos por mujeres

Gabriel Díez Abadie, University of Las Palmas de Gran Canaria

Rocío Gragera Retuerto, University of Las Palmas de Gran Canaria

Este estudio examina la influencia de la lengua francesa en la terminología culinaria en lengua inglesa del siglo XIX en textos manuscritos e impresos. Para esto, se llevará a cabo un análisis léxico del subcorpus del siglo mencionado incluido en el *Corpus of Women's Instructive Writing* (1550-1899) mediante el uso de metodología basada en la lingüística de corpus. Así, se busca identificar y analizar la presencia de términos y expresiones francesas en los textos que componen el corpus con el fin de comprender su impacto en la terminología culinaria de la época y su contribución a la fusión cultural en la cocina escrita por mujeres. La elección de esta compilación se debe a la novedad en el uso de textos exclusivamente por mujeres y que este corpus se ha etiquetado para el reconocimiento de las categorías gramaticales, incluyendo el etiquetado de palabras extranjeras. Un parte de este trabajo se nutre de la cuidada contextualización de los textos gastronómicos decimonónicos escritos en lengua inglesa y en lengua frances. De ahí, que usemos estudios sobre la estrecha relación cultural y gastronómica entre Francia e Inglaterra durante el siglo XIX (Flandrin y Montanari, 1999; Lehmann, 1983; Saillard, 2014). Se consideran aspectos clave como la transferencia de conocimientos culinarios entre Francia y el mundo angloparlante, la importancia de la terminología francesa en la definición de conceptos y técnicas culinarias avanzadas, y la influencia de estos términos en la percepción social de la cocina y la gastronomía.

El argumento principal se centra en cómo la aparición de términos franceses en el libro no solo refleja una influencia lingüística, sino también cultural, sobre la cocina anglosajona de este período. Los objetivos principales de este estudio son cuantificar la presencia de terminología en francés, analizar su contexto de uso y evaluar su impacto en las prácticas culinarias del siglo XIX. Este estudio se prevé como una etapa inicial en tanto que los resultados que se obtengan se compararán con el uso de vocablos franceses en una compilación de recetas escritas por hombres.

Como se expresó anteriormente, el análisis contextual es necesario para entender la función de este léxico en los textos para determinar cómo contribuye a la narrativa culinaria y a la transmisión de conocimientos gastronómicos, pero también para comprobar la aportación de estos usos al modo interpersonal (Halliday y Matthiessen, 2014). Esto significa que las escritoras pueden usar estos elementos léxicos como indicadores de posición para reivindicar un espacio epistémico en un mundo tradicionalmente liderado por los hombres. Este enfoque interdisciplinario sirve para tener una comprensión más profunda del uso de cambios de código, en este caso en lengua francesa, en estos textos de cocina escritos en lengua inglesa por mujeres, que podrían ser más frecuentes en el caso de las mujeres por el motivo mencionado, por lo que, en un próximo trabajo, se hace necesario

realizar este estudio en un corpus de estos textos escritos por hombre. Esto no menoscaba considerar la notable dinámica de intercambio cultural y lingüístico entre estas lenguas en un período clave de la historia de la gastronomía (Mennell, 1996; David, 1999), pero una frecuencia mayor según el género de los escritores demostraría una posible tendencia de uso cuya función descansaría en la manifestación de significado interpersonal.

Análisis visual de las investigaciones sobre el Corpus paralelo de Sueño en el pabellón rojo en la plataforma CNKI

Lili Wang, University of Las Palmas de Gran Canaria

María Teresa Cáceres-Lorenzo, University of Las Palmas de Gran Canaria

Yanli Zhang, Shanghai International Studies University

La Universidad Wenli de Shaoxing (China) creó el *Corpus paralelo de Sueño en el pabellón rojo (chino-inglés)* en el año 2010 (Ren, 2010). Fue el primer corpus en China que trató la novela clásica china *Sueño en el pabellón rojo*, conocida como "el Quijote chino" (Guelbenzu, 2010). Este corpus no solo nos proporciona el texto chino de esta novela, sino que también ofrece dos versiones de la traducción al inglés: la versión de Hawkes y Minford y la versión de Yang. Las publicaciones sobre este corpus en China se pueden encontrar en la biblioteca digital más grande de China: la CNKI. Por ello, los objetivos establecidos para este trabajo son los siguientes:

a. Presentar el *Corpus paralelo de Sueño en el pabellón rojo* y la biblioteca china CNKI.

b. Analizar los estudios sobre el *Corpus paralelo de Sueño en el pabellón rojo* en CNKI a través del análisis visual.

Para ello, necesitamos responder las siguientes dos preguntas:

1. ¿Qué son el *Corpus paralelo de Sueño en el pabellón rojo* y la platafora CNKI?

2. ¿Cuáles son las características de las investigaciones sobre el *Corpus paralelo de Sueño en el pabellón rojo* en la plataforma CNKI?

Para lograr los objetivos de la investigación, los métodos que utilizamos son principalmente la búsqueda de plataforma y el análisis cuantitativo y cualitativo y el análisis visual. El análisis visual es un medio para explorar y comprender datos, es la presentación visual de los datos abstractos. Se trata de un enfoque novedoso que se basa en la integración de procesos automatizados y de las habilidades únicas de los seres humanos, en un esfuerzo común para profundizar en problemas complejos que tienen que tratar con grandes cantidades de datos (Sánchez *et al*, 2011).

Gracias a nuestra investigación, encontramos que se han publicado más de 200 artículos científicos sobre el *Corpus paralelo de Sueño en el pabellón rojo* en la plataforma CNKI. Los temas más estudiados son la investigación sobre *Sueño en el pabellón rojo*, la construcción de un corpus y su investigación aplicada relacionada. También encontramos que 2012 fue el año con más trabajos publicados; que hay más estudios sobre la traducción de Hawkes y Minford que de la traducción Yang; que la distribución temática se centra en investigaciones en inglés; que el autor con más publicaciones es Liu Zequan y su equipo de la Universidad de Yanshan; y que, además, la revista más publicada es la revista *Sueño en el pabellón rojo.*

En resumen, aunque no hay una gran cantidad de artículos sobre el *Corpus paralelo de Sueño en el pabellón rojo* en CNKI, estos nos brindan un muy buen recurso para investigar dicha novela desde diferentes aspectos.

Referencias

CNKI. *Biblioteca digital.* https://www.cnki.net/index/

Ren, L. (2010). *Corpus paralelo de Sueño en el pabellón rojo (chino-inglés).* http://corpus.usx.edu.cn/hongloumeng/index.asp

Sánchez, R. T., Fontanillo, L. F., Marcos, A. E. y Herrero, C. S. (2011). Visual analytics: A novel approach in corpus linguistics and the Nuevo Diccionario Histórico del Español. En *Las tecnologías de la información y las comunicaciones: presente y futuro en el análisis de corpus: Actas del III Congreso Internacional de Lingüistica de Corpus* (pp. 335-342). Editorial Universitat Politècnica de València.

Analyzing the rhetoric of Díaz-Ayuso and Lauren Boebert's political tweets on X: A corpus-assisted analysis

Mª Milagros del Saz Rubio, Polytechnic University of València

This article focuses on the strategic tactic of 'going negative' (Druckman et al., 2020; Nai & Sciarinim, 2018) in order to elucidate whether its use is present in the social media discourse of two well-known female politicians on the micro-blogging site X. More precisely, a selection of 1,000 tweets from the account of the President of the Community of Madrid —Isabel Díaz Ayuso (@IdiazAyuso), and 1,000 from the account of Congresswoman Lauren Boebert (@laurenboebert) were analyzed in relation to their use of *negativity* as a rhetorical political strategy regarding various sensitive topics of high relevance such as immigration, the Israeli-Palestinen conflict, together with LGBTQ+ issues. Although the strategy of 'going negative' and the related tactic of character assassination are not new in the political scenario (see Ross & Caldwell, 2020; Samoilenko et al., 2020), no previous analysis tackles the deployment of these tactics contrastively on female politicians, at least to my knowledge.

Employing a corpus-based approach of comparative keyword analysis and drawing on insights from the Critical Discourse Analysis (Baker et al., 2008) and multimodality, this study employs the analytical framework of Appraisal, as introduced by Martin and White (2005), to investigate the semantic resources most commonly employed within Díaz Ayuso and Boebert's tweets to delegitimate other political actors (. Specifically, the focus is on the manner in which these resources are utilized to craft a discourse with a negative connotation toward other political actors by focusing on predication and delegitimation strategies (Van Leeuwen & Wodak, 1999). Results indicate that both Díaz-Ayuso and Boebert's tweets systematically draw on resources from the Appraisal framework with the rhetorical effect of attacking and de-legitimizing either their political opponents, i.e., Prime Minister Pedro Sánchez, in the case of the Spanish politician, and Joe Biden, in the case of Boebert alongside other less tangible entities, such as migrants or other minorities. Likewise, both politicians vilify their opponents while successfully judging their morality or ethics, capacity as leaders, associations with other actors, and accomplishments.

References

Nai A., & Sciarini P. (2018). Why "Going Negative?" Strategic and situational determinants of personal attacks in Swiss direct democratic votes. *Journal of Political Marketing, 17*(4), 382-417, DOI: https://doi.org//10.1080/15377857.2015.1058310

Ross A. S., & Caldwell D. (2020). "Going negative": An APPRAISAL analysis of the rhetoric of Donald Trump on Twitter. *Language & Communication*, *70*, 3-27. DOI: https://doi.org//10.1016/j.langcom.2019.09.003

Baker, P., Costa G., Majid K., M. McEnery A., & Wodak R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society*, *19*(3), 273–306. DOI: https://doi-org//10.1177/0957926508088962

Druckman, J. N., Kifer M. J., & Parkin M. (2010) Timeless strategy meets new medium: Going negative on congressional campaign web sites, 2002- 2006. *Political Communication*, *27*(1), 88-103.

Martin, J. & White, R. P. (2005) *The language of evaluation: Appraisal in English.* Palgrave Macmillan.

In S. Samoilenko, M. Icks, J. Keohane, & E. Shiraev (Eds.). (2020). *Routledge handbook of character assassination and reputation management.* Routledge.

Van Leeuwen, T., & Wodak, R. (1999). Legitimizing immigration control: A discourse-historical analysis. *Discourse Studies, 10*(1), 83–118.

Aplicaciones de los corpus paralelos a la investigación traductológica, a la enseñanza de la traducción y a la actividad traductora: estudio práctico basado en el corpus de acceso libre inglés > < español PaEnS

Inmaculada Serón Ordóñez, Pablo de Olavide University

Los corpus paralelos son una potente herramienta de investigación traductológica. Permiten obtener datos empíricos de naturaleza muy diversa: desde patrones de correspondencia (o no correspondencia) entre lenguas, con información cuantitativa (Bernardini, 2004: 28; Johansson, 2007), hasta patrones de comportamiento de (determinados) traductores o

incluso diferencias entre un idioma como lengua original y como lengua de traducción, en el caso de los corpus bidireccionales (en los que un idioma puede ser tanto lengua original como meta).

Respecto a la utilidad de los corpus paralelos en la enseñanza de la traducción, dicha utilidad ha quedado plenamente de manifiesto en Gallego Hernández (2020). Por otro lado, el uso generalizado de memorias de traducción entre los profesionales de la traducción da suficiente cuenta de la importancia de los corpus paralelos en el ejercicio de la profesión.

La presente comunicación tiene como objetivo mostrar las aplicaciones para los investigadores, los traductores y los docentes de traducción del corpus inglés-español/español-inglés PaEnS, un corpus paralelo bilingüe de acceso libre en Internet (www.corpuspaens.eu).

En primer lugar, se describirá brevemente el contenido del corpus:

— Por un lado, obras originales actuales y sus traducciones. Estas obras, pertenecientes a géneros de ficción (acción, literatura infantil y juvenil, etc.) y de no ficción (ensayo, divulgación científica, etc.), constituyen el núcleo del corpus.

— Por otro lado, colecciones de textos originales y sus traducciones provenientes de proyectos externos. Estas colecciones amplían la variedad lingüística del corpus. Hay tres disponibles actualmente, cuyas fuentes son el corpus Europarl, el corpus de las charlas TED, y el de la comunidad de voluntarios Global Voices.

Tras la descripción del contenido, se mostrará de forma escueta cómo realizar búsquedas de concordancia en el corpus. Las posibilidades de búsqueda son numerosas, lo que abre un amplio abanico de posibilidades de investigación, docentes y traductoras. Estas centrarán la comunicación.

Por citar algunos ejemplos, se puede buscar por fecha, obra, autor, texto original o meta, o una traducción concreta para un término específico (búsqueda bilingüe). La ventana de resultados indica cuál es el segmento original y el traducido, a diferencia de otros corpus de acceso libre (p. ej., Linguee).

El corpus se puede utilizar, por tanto, para estudiar traducciones de obras de un mismo autor, en el marco de estudios de recepción, o investigar sobre el estilo de un traductor. También puede ser explotado como fuente monolingüe, de inglés o español, para analizar, por ejemplo, extranjerismos. El hecho de que incluya literatura de ficción supone que ofrezca correspondencias de lenguaje coloquial, una ventaja para traductores literarios o audiovisuales, quienes pueden echar en falta estas correspondencias en diccionarios. Además, la variedad de tiempos verbales que incluye el corpus es superior a la de otros

corpus de lenguaje fundamentalmente técnico o administrativo, lo que lo hace especialmente útil para el estudiantado. Estos y otros muchos ejemplos serán comentados en la sesión al tiempo que se muestra en vivo cómo realizar las búsquedas pertinentes.

Referencias

Bernardini, S. (2004). Corpora in the classroom. En J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 15-36). John Benjamins.

Doval, I. (2018). Corpus paralelos en la enseñanza de lenguas extranjeras: un ejemplo de aplicación basado en el corpus PaGeS. *CLINA, 4*(2), 65-82.

Gallego Hernández, D. (2020). *Apuntes de traducción económica, comercial y financiera. Aproximación contrastiva (francés y español) por géneros textuales*. Universidad de Alicante.

Johansson, S. (2007). Using corpora: From learning to research. En E. Hidalgo, L. Quereda y J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 17-30). Rodopi.

Johansson, S. (2009). Some thoughts on corpora and second-language acquisition. En K. Aijmer (Ed.), *Corpora and language teaching* (pp. 33-44). John Benjamins.

Aportaciones de los corpus a la elaboración de un vocabulario fundamental del español

Isabel Sánchez López, University of Granada

El vocabulario fundamental debe aglutinar las palabras esenciales de un idioma, tan necesarias para la comunicación básica. Con esto queremos destacar la importancia y a su vez la necesidad de contar con un trabajo de estas características en nuestra lengua. Las palabras que lo componen deben conformar el núcleo esencial del léxico de una lengua y son indispensables para la construcción de oraciones y expresar ideas fundamentales.

Alcanzar un resultado final completo y coherente requiere una revisión de los recursos que los diferentes estudios han considerado como la base del mismo. En esta ocasión queremos detenernos en los corpus. Son colecciones sistemáticas y extensas de textos que representan el uso real de un idioma en contextos variados. Estos corpus pueden ser compilados de diferentes fuentes, como libros, artículos, conversaciones, internet y otros medios escritos y hablados. El análisis de los corpus léxicos es esencial para comprender cómo se utilizan las palabras en diferentes contextos y cómo evoluciona el lenguaje a lo largo del tiempo.

La relación entre el vocabulario fundamental y los corpus léxicos es un tema esencial en lingüística y lexicografía para entender la conexión entre las palabras esenciales de un idioma y las colecciones extensas de textos que representan su uso. Para comprender esta relación, es crucial examinar tanto el concepto de vocabulario fundamental como el papel de los corpus léxicos en el análisis lingüístico.

La relación más primigenia entre el vocabulario fundamental y los corpus léxicos radica en que el primero proporciona una base sólida y esencial, mientras que los segundos ofrecen una representación más amplia y dinámica del uso del lenguaje en la práctica. El vocabulario fundamental establece las palabras clave que son esenciales para la comunicación, pero los corpus léxicos permiten analizar cómo estas palabras se utilizan en contextos específicos, identificando variaciones en significado, cambios semánticos y nuevos usos que puedan surgir con el tiempo.

Los lexicógrafos y lingüistas utilizan los corpus léxicos para recopilar datos sobre el uso real de las palabras y para construir diccionarios y recursos lingüísticos más precisos y actualizados. Al examinar los corpus léxicos, pueden identificar patrones de uso, determinar la frecuencia de ciertas palabras y entender mejor las connotaciones y matices de significado asociados con ellas. Este enfoque basado en datos reales proporciona una visión más completa y precisa del lenguaje, en contraste con las definiciones estáticas que a menudo se encuentran en diccionarios tradicionales.

La relación entre el vocabulario fundamental y los corpus léxicos también destaca la naturaleza dinámica del lenguaje. Mientras que el vocabulario fundamental establece una base sólida, los corpus léxicos revelan cómo el lenguaje está constantemente cambiando y adaptándose a las necesidades de la sociedad y a las influencias culturales. Las nuevas palabras, expresiones y significados emergen a medida que el lenguaje evoluciona, y los corpus léxicos proporcionan una ventana única para observar estos cambios en tiempo real.

Así el objetivo de nuestro trabajo es intentar ahondar en esta relación para poder justificar su valor parcial dentro de un proyecto como el que traemos, la elaboración de un vocabulario fundamental del español.

Referencias

Bosque, I. (dir.) (2004). *Redes. Diccionario combinatorio del español contemporáneo*. SM.

Kilgarriff, A. (2005). Putting the Corpus into the Dictionary, *Proceedings MEANING Workshop*, Trento.

Mel'čuk, I.A. *et al*. (1984-1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques* I-IV. Les Presses de l'Université de Montréal.

Oster, U. (2009): La adquisición de vocabulario en una lengua extranjera: de la teoría a la aplicación didáctica, *Porta linguarum*, *11,* 33-50.

Richter, M., Quasthoff, U., Hallsteinsdóttir, E. y Biemann, C (2006). Exploiting the Leipzig Corpora Collection, en *Proceedings of the IS-LTC 2006.* Ljubljana, Slovenia.

Sánchez, A. (dir.) (2001). *Gran diccionario de uso del español basado en el corpus lingüístico CUMBRE*. Sociedad General Española de Librería.

Wanner, L. (2006). ¿El corpus como un Diccionario de colocaciones? En M. Alonso Ramos (Ed.), *Diccionarios y fraseología* (*Anexos de Revista de Lexicografía*, 3). Universidade da Coruña.

Żmigrodzki, P. (2005). Dictionary as a Text Corpus - Text Corpus as a Dictionary. *Perspectives of Scholarly Lexicography in Poland*.

Automatic Readability Analysis of Small-range Corpus Based on Native Chinese Textbooks for Junior High School

Lizhen Hao, Xiamen University

Reading is the main way for human beings to obtain information and understand the world. Accurate graded reading is conducive to the improvement of students' learning enthusiasm and efficiency. Because it takes time and effort to grade the difficulty of text manually, and most of the current readability automatic grading techniques are based on a large-range corpus, leading to the wide range of the grading results, in this paper the automatic text readability analysis method is applied to a small corpus of Chinese native language textbooks for junior high school, aiming to explore whether the impact of existing indicators on readability and their explanatory ability will change in the face of a small range of corpus and the more detailed readability grading. Based on the corpus of 300 texts in junior high school Chinese textbooks, selecting 16 text features as explanatory variables from lexical, syntactic and textual levels and the level of Chinese characters, and using the readability level 0 and 1 as the dependent variable, we build a binary logistic regression model to grade the readability of texts.

The result show that the average stroke number of characters has a significant negative impact on the difficulty of texts in the face of the small range of Chinese texts in junior high school, which is contrary to common sense. Meanwhile, the average paragraph length, the single sentence proportion, the conjunction proportion, the idiom proportion and the average sentence length all have significant positive effects on text difficulty, and the effect of single sentence proportion is also contrary to common sense. Compared with the large corpus, the importance of lexical level indicators to text difficulty grading decreases in the small-range corpus, while the importance of syntactic features increases. At the same time, non-text factors such as genre and style all play an important role in readability grading.

These prove that the narrowing of corpus will lead to the great fluctuation of the explanatory ability and the influence direction of linguistic features. In addition, we further found that the most important predictive variables change when the linguistic features are used to identify the texts of different grades. Finally, confusion matrix and ROC curves were used to evaluate the prediction effect of the model on the test set, and the prediction accuracy reached 61.7%, which proves the effectiveness of the model. Compared with the support vector machine model, our model is simpler and more economical. And compared with the linear regression model, the text grading mechanism and model evaluation method in this thesis are both more scientific and effective. Moreover, this study can provide some suggestions for the planning of graded Chinese reading in junior high school.

References

程勇, 徐德宽, 董军, (2020). 基于语文教材语料库的文本阅读难度分级关键因素分析与易读性公式研究, 语言文字应用, 01, 132-143.

刘苗苗, 李燕, 王欣萌, 甘琳琳, 李虹. (2021). 分级阅读初探:基于小学教材的汉语可读性公式研究, 语言文字应用, 02, 116-126.

吴思远, 蔡建永, 于东, 江新. (2018). 文本可读性的自动分析研究综述. 中文 信息学报, 12, 1-10.

吴思远, 于东, 江新. (2020). 汉语文本可读性特征体系构建和效度验证. 世界汉语教学, 01, 81-97.

Caylor J. S., et al. (1973). Methodologies for determining reading requirements of military occupational specialties. *Adult Literacy, 81.*

Dale E., Chall J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, *27*(1), 37-54.

Gunning R. (1952). *The technique of clear writing*. McGraw-Hill.

Graesser A. C., Mcnamara D. S., & Kulikowich J. M. (2015) Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234.

Heilman M., Collins-Thompson K., Eskenazi M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*.

Laughlin G. H. M. (1969). SMOG Grading-A new readability formula. *Journal of Reading, 12*(8), 639-646.

Sung Y. T., Chen J. L., Cha J. H., et al. (2015). Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods, 47*(2), 340-354.

Sung Y. T., Chang T. H., Lin W. C., et al. (2016). CRIE: An automated analyzer for Chinese texts. *Behavior Research Methods, 48*(4), 1238-1251

Change in motion: On the rise and development of mirative wind up

Mario Serrano Losada, Complutense University of Madrid

Evidential and mirative strategies are pervasive in English today, as speakers often resort to the expression of source of information and surprise, respectively. Such strategies include, for instance, the use use of specific intonation patterns, disjuncts like *apparently* or *surprisingly* or grammaticalized expressions like the mirative [*what a* NP] construction. Among these, a group of verbs stand out: evidential and mirative raising verbs (EMRVs). When used in certain constructions, EMRVs express a range of evidential (e.g. *seem*, *appear*) and mirative (e.g. *chance*, *happen, prove*) meanings. While the rise and devolopment of some EMRVs can be traced back to Middle English, other members of this category have developed in more recent times, as is the case of mirative *turn out*, which is first recorded in the eighteenth century, or mirative *end up*, a twentieth century creation (AUTHOR).

The present paper zooms in on the development of mirative *wind up*, exemplified in (1)-(2):

I went to have my tooth finished, **winding up** for tea at the Club. (Woolf, 1918, OED, s.v. *to wind up*, 4.e.)

Incredibly, the 1968 campaign **wound up** being the best not only of Gibson's career but also one of the best in MLB history. (COCA:MAG:2019)

In both (1)-(2), *wind up* is used to introduce surprising information, with overtones of unexpectedness and counterexpectation (DeLancey 1997, Aikhenvald 2012). Such instances can be traced back to earlier uses in which the verb was used with conclusive meaning, as illustrated in (3), where *wind up* does not refer to the unexpected nature of the event, but rather indicates that the said novel is coming to a conclusion:

I shall be better directed in what manner to **wind up** the Catastrophe of the pretty Novel. (Richardson, 1740, OED, s.v. *to wind up*, 4.b)

As in the case of verbs like *turn out* and *end up* (AUTHOR), mirative meaning seems to have developed from erstwhile culminative senses like the one illustrated in (3). The present paper investigates the emergence and development of the mirative uses of *wind up*. Early mirative readings can be found as early as the eighteenth century; however, these do not become established until the twentieth century, when the verb undergoes a considerable surge in frequency.

Moreover, recent evidence suggests that the verb might be on its way to becoming an EMRV, as examples (4)-(5) illustrate:

As it winds up, enlisting a cadre of female singers to essentially pay tribute to yourself […] pays off. (NOW:US:2015)

It was quite a relaxed start but it **wound up to be** quite stressful (NOW:GB:2018)

In (4), *wind up* is used in the parenthetical construction typical of EMRVs, [*as it* EMRV], while in (5), it is used in a raising construction proper. Such innovative uses, although scant, seem to indicate that mirative *wind up* is being analogically attracted to the class of English EMRVs. Data for the present paper has been drawn from a number of sources, including the OED, EEBO, CLMET, COHA, COCA and NOW.

Sources

CLMET = De Smet, Hendrik et al. *The Corpus of Late Modern English Texts*, *version 3.1*. https://perswww.kuleuven.be/~u0044428/.

COCA = Davies, Mark. 2008-. *Corpus of Contemporary American English*. https://www.english-corpora.org/coca/.

COHA = Davies, Mark. 2010-. *Corpus of Historical American English*. https://www.english-corpora.org/coha/.
EEBO = Davies, Mark. 2017. *Early English Books Online Corpus*. Available online at https://www.english-corpora.org/eebo/.
NOW = Davies, Mark. 2016-. *Corpus of News on the Web*: *3+ billion words from 20 countries*, *updated every day*. https://www.english-corpora.org/now/.
OED = Oxford English Dictionary Online. Oxford University Press. http://www.oEd.com/.

References

[Author]
Aikhenvald, A. Y. (2012). The essence of mirativity. *Linguistic Typology*, *16*(3), 435–485. DOI: https://DOI.org/10.1515/lity-2012-0017
DeLancey, S. (1997). Mirativity: The grammatical marking of unexpected information. *Linguistic Typology*, *1*(1), 33–52. DOI: https://DOI.org/10.1515/lity.1997.1.1.33

Citing others in the Coruña Corpus

Margarita Mele Marrero, University of La Laguna

The current citation network owes much to the “publish or perish” imperative born in the first half of the nineteenth century. As Csiszar (2017) explains, the slogan is directly connected with the Royal Society of London’s decision of counting the number of papers published in its Philosophical Transactions as a formula to evaluate scientific work. The suggestion for such a procedure was made in 1830 by one of its members, Charles Babbage. This method was much discussed by his colleagues, particularly taking into account that at the time “scientific reputations were built not on periodicals but on books and other proofs of genius” (Csiszar, 2017: 163). Such evaluation would be partially replaced by the Science Citation Index and Impact Factor that appeared in the second half of the twentieth century. To what extent scientists were originally conscious of the importance of citing others and the impairment caused by plagiarism is one of the concerns of this analysis. This paper focuses on early scientific writing in English, its purpose is to study when, what for and how authors acknowledged their sources or the work of their peers during the eighteenth and nineteenth centuries.

Corpus linguistics has provided the possibility to analyze structured discourse to attain representative data, otherwise difficult to access. However, large quantities of information may, for the sake of generalization, lead to the omission of relevant details. Validating results from least to most may avoid a simple quantitative, and sometimes even inaccurate, use of corpora. Therefore, in order to obtain results, I have extracted the information related to citations firstly from a subcorpus of the Coruña Corpus (CC), The Corpus of English Chemistry Texts (CECheT). A subcorpus on Chemistry is relevant as a sample of the beginning of scientific writing since it comprehends, in the different types of texts of the period, the study of the composition of materials, creation of new ones, discussion of previous procedures and instruments. On the other hand, CECheT’s size permits to offer not only a reliable quantitative analysis but also a qualitative one difficult to assume in a larger corpus. Numbers and percentages about earlier historical periods might be misleading if particular cases are not taken into account. Finally, of the CC subcorpora on hard sciences, CECheT is the one that offers more results for our objectives and concentrating first in this

selection will allow me to extrapolate its data, analyzed in more detail, to the other hard science subcorpora of CC. The evaluation of the results obtained will be carried out applying a pragmatic point of view and considering attitude markers (Hyland 2005) present in the citations.

Some of the conclusions reached show how citations were used both to acknowledge and to denounce plagiarism, that there is a pragmatic variation in the form of the citation from the eighteenth to the nineteenth centuries and that the number of citations does not seem to be connected with the authors' academic success. Whether present-day evaluation systems have changed this, is a question that needs to be solved.

References:

Csiszar, A. (2017). The Catalogue That Made Metrics, and Changed Science. Nature, 551(7679), 163–165. https://doi.org/10.1038/551163a

Hyland, K. (2005). Stance and Engagement: A Model of Interaction in Academic Discourse. Discourse Studies, 7(2), 173–192. https://doi.org/10.1177/1461445605050365

Moskowich, I., Lareo Martín, I., Camiña-Rioboó, G., & Crespo, B. (2012). Corpus of English Texts on Astronomy (CETA). A Coruña: Universidade da Coruña. https://doi.org/10.17979/spudc.9788497497084

Moskowich, I., Puente-Castelo, L., & Monaco, L. M. (2020). Corpus of English Life Sciences Texts (CELIST). A Coruña: Universidade da Coruña. https://doi.org/10.17979/spudc.9788497498388

Moskowich, I., Puente-Castelo, L., & Monaco, L. M. (2022). Corpus of English Chemistry Texts (CECheT). A Coruña: Universidade da Coruña. https://doi.org/10.17979/spudc.9788497497848

Climate change impact on global health

Stefania M. Maci, University of Bergamo

As WHO (2023) asserts, "[c]limate change is impacting human lives and health in various ways. It threatens the essential ingredients of good health - clean air, safe drinking water, nutritious food supply and safe shelter - and has the potential to undermine decades of progress in global health." Climate change can clearly have significant impacts on human health: Not only is warming the planet leading to more frequent and severe heat waves that can cause heat exhaustion, dehydration and even death, but it can also exacerbate preexisting health conditions such as cardiovascular disease, respiratory illness and allergies. In addition, climate events can lead to injury, displacement and mental health problems such as post-traumatic stress disorder (PTSD). Overall, this climate change poses a significant threat to global health and well-being. It therefore seems essential to address climate change through mitigation and adaptation measures, which is essential to protect public health (Abbas et al., 2022).

Since representations of climate change discourse "have enacted their own discursive formations, which people discuss and act upon at local, national and global scales" (Taylor,

2013, pp. 17), this study examines how digital discourse surrounding climate change undergoes discursive adaptation measures when the WHO discusses the impact of climate change on health with various audiences. Specifically, we analyze the WHO's six web platforms, each corresponding to a different continental region, using a Corpus Linguistics (CL) approach augmented by AI and Large Language Models. By leveraging generative pretrained transformers (GPTs) from the MedAlpaca collection (Han, 2023), a series of LLMs trained on medical data, we aim to identify medical discourse within the context of regional climate change discourse and evaluate its significance and objectives. Furthermore, our CL analysis will be accompanied by a multimodal analysis conducted with the machine learning library Scikit-learn (Hackeling, 2017), to identify image similarities and assess how clusters of similar images are used to achieve specific communicative objectives within both region-specific discourse and broader discourse across regions. We will thus determine how the same topic has been adapted discursively to reach diverse audiences residing in distinct geographical, social, and economic contexts, with the goal of encouraging proactive policy-making, raising public awareness, and promoting collective action.

References

Abbas, A., Ekowati, D., Suhariadi, F., & Fenitra, R.M. (2022). Health implications, leaders societies, and climate change: A global review. In U. Chatterjee, A. O. Akanwa, S. Kumar, S.K. Singh & A. Dutta Roy (Eds.), *Ecological footprints of climate change. Springer climate.* Springer. DOI: https://DOI.org/10.1007/978-3-031-15501-7_26

Taylor, C. (2013). The discourses of climate change. In T. Cadman (Ed.) *Climate Change and Global Policy Regimes. International Political Economy.* Palgrave Macmillan

WHO (2023). *Climate change.* https://www.who.int/health-topics/climatechange#tab=tab_1 [28 April 2023]

Huan, T., Adams, L. C., Papaioannou,J., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., & Bressem K. K. (2023). MedAlpaca: An open-source collection of medical conversational AI models and training data. arXiv:2304.08247

Hackeling, G. (2017). *Mastering machine learning with scikit-learn*. Packt Publishing

CoLaGe (Corpus for the Study of Language and Gender in Spanish): un corpus bidialectal del español oral

Andrea Carcelén Guerrero, University of Helsinki

Gloria Uclés Ramada, University of Alicante

Pekka Posio, University of Helsinki

Sven Kachel, University of Kaiserslautern-Lamdau

Esta comunicación nace con el propósito de presentar, por un lado, la metodología de recogida y elaboración del corpus oral CoLaGe (Corpus for the Study of Language and Gender in Spanish) desarrollado en el proyecto Gender, Society and Language Use: evidence from Mexico and Spain y, por otro, describir las características de su diseño y construcción atendiendo a las particularidades que definen a los corpus orales en las diferentes fases de construcción (Sinclair, 2004; McEnery, Xiao y Tono, 2006; Adolph y Knight, 2010; Reppen, 2010; Baker, 2014; Rojo, 2021; Egbert, Biber y Gray, 2022; Carcelén Guerrero *et al.*, 2024).

Entre los objetivos de investigación de este proyecto se encuentran 1) estudiar las interrelaciones entre el género de los hablantes y su uso de la lengua; 2) profundizar en la comprensión de la relación entre lengua, sociedad y género de forma comparada en dos zonas lingüísticas (en Valencia-España y en Guadalajara-México); y 3) averiguar si las diferencias de género en el uso de la lengua se explican mejor a través del establecimiento de un enfoque escalar en cuanto a la autoadscripción de género, a diferencia del enfoque tradicional que considera hombre/mujer como categorías binarias y, por tanto, mutuamente excluyentes (Bem, 1974, Thompson y Pleck, 1986; Kachel, Steffens y Niedlich, 2016).

Para ello, se han recogido dos subcorpus orales de base lingüística y sociopsicológica que compara las dos zonas dialectales mencionadas (CoLaGe -V y CoLaGe -G). Según investigaciones anteriores, a pesar de compartir idioma, México y España difieren en cuanto a normas y roles de género (Hausmann *et al.*, 2014; Social Watch, 2012; INEGI, 2006; Comisión Europea, 2017), lo que proporciona un fructífero punto de comparación. El corpus CoLaGe recopila cuatro tipos de materiales en ambas ciudades para garantizar la intercomparabilidad; entre ellos se incluye una entrevista sociolingüística dividida en dos partes (una centrada específicamente en el género), juegos de rol que simulan situaciones conflictivas, una tarea de descripción de imágenes para obtener datos fonéticos y encuestas psicológicas sociales. La muestra de hablantes ha sido seleccionada atendiendo a criterios de edad y nivel sociocultural, así como de género. Los materiales han sido codificados y alineados audio y texto en ELAN (Max Planck Institute for Psycholinguistics, 2021), siguiendo unas convenciones de transcripción que combinan las propuestas de Jefferson (2004) y de Briz y Grupo Val.Es.Co. (2002); además los datos han sido pseudonimizados para garantizar la privacidad de los participantes.

Con este corpus se pretenden desarrollar estudios comparativos para determinar la relación entre el género del hablante y el uso de la lengua, investigaciones ampliamente desarrolladas en el ámbito anglófono para el inglés (West y Zimmerman, 1987; Butler, 1990; Mulac *et al.*, 2001; Argamon *et al.*, 2003, Bucholtz y Hall, 2005; Newman *et al.*, 2008; Harrington *et al.*, 2008; Speer y Stokoe, 2011), pero escasamente representadas en el mundo hispánico para el español.

Referencias

Adolphs, S. y Knight, D. (2010). Building a spoken corpus. What are the basics? En A. O'Keeffe y M. J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics*, (pp. 38-52). Routledge.
Argamon, S., Koppel, M., Fine, J. y Shimoni, A. R. (2006). Gender, genre, and writing style in formal written texts. Text. *Interdisciplinary Journal for the Study of Discourse*, *23*, 321-346.
Autor (2024).
Baker, P. (2014). *Using corpora to analyze gender*. Bloomsbury Academic.
Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*(2), 155-162.
Briz Gómez, A. y Grupo Val.Es.Co. (2002). *Corpus de conversaciones coloquiales*. Anejos Oralia. Arco Libros.
Bucholtz, M. y Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7(4-5), 585–614.
Butler, J. (1990). *Gender trouble: Feminism and the subversion of identity*. Routledge.
Comisión Europea, (2017). *Gender equality, stereotypes, and women in politics*. Directorate-General for Communication, Luxembourg.
Egbert, J., Biber, D. y Gray, B. (2022). *Designing and evaluating language corpora. A practical framework for corpus representativeness*. Cambridge University Press.
Harrington, K., Litosseliti, L., Sauntson, H. y Sunderland, J. (2008). *Gender and language research methodologies*. Palgrave Macmillan.
Hausmann, R., Tyson, L. D., Bekhouche, Y. y Zahidi, S. (2014*). The Global Gender Gap Report 2014*. World Economic Forum, Geneva.
Instituto Nacional de Estadística y Geografía (INEGI). (2016). Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH).
Jefferson G. (2004). Glossary of transcript symbols with an introduction. En Lerner G, (Ed.). *Conversation Analysis: studies from the first generation* (pp. 13-23). John Benjamins.
Kachel, S., Steffens, M. C. y Niedlich, C. (2016). Traditional masculinity and femininity: Validation of a new scale assessing gender roles. *Frontiers in Psychology, 7.*
McEnery, T., Xiao, R. y Tono, Y. (2006). *Corpus-Based Language Studies*. Routledge.
Mulac, A., Bradac, J. y Gibbons, P. (2001). Empirical support for the gender-as-culture hypothesis An intercultural analysis of male/female language differences. *Human Communication Research, 27*(1), 121-152.
Newman et al., (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes, 45,* 211-236.
Reppen, R. (2010). Building a corpus: What are the key considerations? En A. O'Keeffe, y M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 31- 37). Routledge.
Rojo, G. (2021). *Introducción a la lingüística de corpus en español.* Routledge.
Sinclair, J. (2004). Corpus and text. Basic principles. En M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice*. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm
Social Watch. (2012). Gender equity index. En *Sustainable development: The right to a future*. http://www.socialwatch.org/report2012.
Speer, S. y Stokoe, E. (2011). *Conversation and gender*. Cambridge University Press.
Thompson, E. H. y Pleck, J. H. (1986). The structure of male role norms. *American Behavioral Scientist*, *29,* 531-543.
West, C. y Zimmerman, D. H. (1987). Doing gender. *Gender & Society, 1,* 125-151.

Corpus linguistics applied to the language of law: A corpus-driven analysis of French parliamentary debates on immigration

Nadia Makouar, Aston University

This paper on corpus linguistics and language of law addresses an analysis of linguistic and semantic features in French parliamentary debates concerning immigration and the integration of foreigners in France. The bill called “Controlling Immigration, Improving Integration” was proposed through an accelerated procedure in February 2023 and passed in January 2024, despite many articles having been rejected by the Constitutional Council.

Drawing on de Galembert's notion of the assembly debate as a ritualised moment (2010) through which “the legal transubstantiation of the agreement called for to become a norm is pursued and completed”, this paper discusses the semantic negotiation of terms specific to the debates, such as ‘control’, 'improvement', 'migrants', 'foreigners', and ‘asylum seekers’, as well as the variation and transformation of their meaning over time. The aim of this study is then to examine the meaning-making process within the law-making process regarding both the control and improvement of future conditions for foreigners, as well as the expression of their human rights and their future status in the debates.

This corpus-driven analysis focuses on the discourse and the variation in change between political parties in the legislative process during the debates in the National Assembly and in the Senate. The analysis will be carried out using a corpus of the discussions and public sessions of the French Senate's examining committee and debates within the National Assembly, held between February 2023 and December 2023. The corpus is analysed using qualitative and quantitative methods, combining top-down and bottom-up approaches. Two main parameters are used for understanding the variation and change of meaning until the point of stabilisation in the legislative text: contrastive analysis, which compares the terms and phraseology of different speakers, and diachronic analysis, which helps understand the evolution of terms and expressions over the course of the debates.

To achieve this, corpus methods for semantics (McEnery & Brezina, 2022; Rastier, 2015; Lecolle, 2018) are applied to identify linguistic and semantic features. Therefore, two combined methods of corpus analysis are used: the structural approach to identify the main themes of discussion (Hierarchical analysis, Loubère & Ratinaud, 2014), and the contrastive approach by using co-occurrences, keyness scores, and hypergeometric tests (Lexico 3–Lamalle et al., 2009). Indeed, changes in terminology and temporality can be captured by contrastive and structural analysis methods such as concordances, co-occurrences, and hierarchical classification. The paper ends by discussing the application of the global-local approach to legal corpora and the relationship between corpus linguistics, temporality, semantic change, and the semiotics of law.

References

De Galembert, C. (2011). Les trois corps du parlementaire. Un débat d'assemblée au prisme d'Alceste. Faire parler le parlement. Making Parliament Speak, Oct 2011, Paris, France. halshs00832729

Goźdź-Roszkowski, S. (2021). Corpus linguistics in legal discourse. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, *34*(5), 1515-1540.
Lamalle, C., Martinez, W., Fleury, S., Salem, A., Fracchiolla, B., Kuncova, A., & Maisondieu, A. (2003). Lexico3: Outils de statistique textuelle. Manuel d'utilisation. Lexi&co. http://lexi-co.com/ressources/manuel-3.41.pdf
McEnery, T., & Brezina, V. (2022). *Fundamental principles of corpus linguistics.* Cambridge University Press.
Lecolle, M., Veniard, M., & Guérin, O. (2018). Pour une sémantique discursive: propositions et illustrations*. Langages*, (210), 35-54.
Loubère, L., & Ratinaud, P. (2014). *Documentation IRaMuTeQ 0.6 alpha 3 version* 0.1.documentation_19_02_2014.pdf (iramuteq.org)
Van Dijk T. A. (2004). Text and Context of Parliamentary Debates. In P. Bayley (Ed.), *Cross-cultural perspectives on parliamentary discourse* (pp. 339-372). John Benjamins Publishing Company.
Rastier, F., & Riemer, N. (2015). Interpretative semantics. In *The routledge handbook of semantics* (pp. 491-506). Routledge.

Corpus-based language tools: The case of English for Specific Purposes in the wine and olive oil tasting domains

Lucía Sanz-Valdivieso, University of Valladolid

English is the *lingua franca* for global exchange in professional and scientific communication. It is in these environments where a general English command is not enough to meet professional communication expectations, making English for Specific Purposes (ESP) one of the main weak points of professionals around the world. To answer this need, ESP digital applications that promote the development of productive and receptive skills might be one of the main means professionals have to meet today's evolving market expectations (Jakobs & Perrin, 2014; Mahlow & Dale, 2014; Tarp 2020). This proposal presents a study where an imperative need for ESP tools is identified, particularly in the wine and olive oil industries in Spain. The aim of the study is to tackle this need through the development of a set of corpus-based ESP linguistic tools (Pérez Blanco & Izquierdo, 2021). These tools are focused on the terminology, phraseology and rhetorical aspects of the ESP of wine and olive oil tasting. A computer-aided methodology was followed to compile and analyze a specialized comparable corpus of English-Spanish wine and olive oil tasting notes (Piao et al., 2003; Kilgarriff et al., 2014; Biber & Conrad, 2019). Keywords, terms and phraseological units were identified, extracted and classified (Gläser, 1994; López Arroyo & Moreno Pérez, 2019) in order to build a bilingual glossary which is at the core of the toolset developEd. Besides, we identified the rhetorical structure of tasting notes and recurrent model lines—"prototypical lexico-grammatical structures" with gaps "to be filled in by user" (Ramón & Labrador, 2015, p. 243). These data were used to build a toolset with several functionalities, including consultation, production and pedagogical practice (Fuertes-Olivera & Tarp, 2020). First, L2 ESP users can simply consult a glossary including Spanish-English equivalents and examples of real use of thousands of terms and phraseological units. Second, users can write tasting notes through the writing assistant based on grammatical and rhetorical patterns found in the corpora, using the glossary to customize the text to a particular wine or olive oil. Finally, users can develop their ESP competences through more than a hundred interactive terminological and rhetorical activities. These tools are the result of corpus analysis, which ensure that users are learning and producing language that meets the expectations of the discourse community where these communicative needs arise. This, in turn, promotes successful communications and acceptance within the discourse community

(Swales, 1990, 2002), which impacts the international image and economic and touristic revenue of the wine and olive oil industries. These tools are useful to develop not only professionals' ESP productive skills, but also trainees and learners' productive and receptive skills. One thing is certain: corpus-based digital tools are a reliable and accessible means to support specialized and professional language users, besides of fostering lifelong learning that allow citizens of the 21st century to meet current professional demands.

References

Biber, D., & Conrad, S. (2019). *Register, genre and style*. Cambridge University Press.

Fuertes-Olivera, P., & Tarp, S. (2020). A window to the future: Proposal for a lexicography-assisted writing assistant. *Lexicographica*, *36*, 257–286. DOI: https://DOI.org/DOI/10.1515/lex-2020-0014

Gläser, R. (1994). Relations between phraseology and terminology with special reference to English. *ALFA, 7-8*, 41–60.

Jakobs, M., & Perrin, D. (2014). Introduction and research roadmap: Writing and text production. In E.-M. Jakobs & D. Perrin (Eds.), *Handbook of Writing and Text Production* (pp. 1–26). De Gruyter Mouton.

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. DOI: https://DOI.org/10.1007/s40607-014-0009-9

López Arroyo, B., & Moreno Pérez, L. (2019). Lexical chunks in English and Spanish sales contracts: A corpus-based study, *Terminology*, *25*(1), 32–59.

Mahlow, C., & Dale, R. (2014). Production media: Writing as using tools in media convergent environments. In E. M. Jakobs & D. Perrin (Eds.), *Handbook of writing and text production* (pp. 209–230). De Gruyter Mouton.

Pérez Blanco, M., & Izquierdo, M. (2021). Developing a corpus-informed tool for Spanish professionals writing specialised texts in English. In J. Lavid-López, C. Maíz-Arévalo and J. R. Zamorano-Mansilla (Eds.), *Corpora in translation and contrastive research in the digital age: Recent advances and explorations* (pp. 147–173). John Benjamins. DOI: https://DOI.org/10.1075/btl.158.06per

Piao, S. L., Rayson, P., Archer, D., Wilson, A., & McEnery, T. (2003). Extracting multiword expressions with a semantic tagger. In *Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, (pp. 49–56). Association for Computational Linguistics. https://www.aclweb.org/anthology/W03-1800

Ramón, N., & Labrador, B. (2015) The rhetorical structure of technical brochures: A proposal for technical writing. *32nd International Conference of the Spanish Association of Applied Linguistics (AESLA): Language Industries and Social Change, Sevilla, Spain* (pp. 241–245). Elsevier. https://DOI.org/10.1016/j.sbspro.2015.02.059

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. (2002). *Research genres: Exploration and applications*. Cambridge University Press.

Tarp, S. (2020). Integrated writing assistants and their possible consequences for foreign-language writing and learning. In I. A. Bocanegra Valle (Ed.), *Applied linguistics and knowledge transfer: Employability, internationalization and social challenges* (pp. 53–76). Peter Lang. https://DOI.org/10.3726/b16992

Creación del corpus de errores en lengua catalana del nivel c1: el celc

Francesca Romero Forteza, Polytechnic University of València

Este trabajo describe y explica una primera aproximación a la creación de un corpus de errores tomando como base metodológica la lingüística de corpus, el análisis de errores y las investigaciones sobre la interlengua. Se trata de una propuesta pionera al crear el primer corpus de errores en lengua catalana del nivel C1, el CELC-C1 (*Corpus d'Errors en Llengua Catalana del Nivell C1*). El objetivo principal de la investigación es detectar las deficiencias en la competencia comunicativa de la población valenciana en el nivel escrito de la lengua.

El corpus se ha elaborado a partir de las pruebas de acreditación de este nivel realizadas a través de la Comissió Interuniversitària d'Estandardització d'Acreditació de Coneixements de Valencià, conocida como CIEACOVA. Se trata de uno de los tres organismos oficiales acreditadores de los niveles de conocimiento de la lengua catalana en la Comunitat Valenciana. Organiza dos convocatorias anualmente. Para esta investigación se ha analizado una muestra aleatoria de los exámenes realizados en la convocatoria del mes junio de 2023.

Las pruebas de nivel de la CIEACOVA constan de diversos ejercicios, tanto orales como escritos. Para crear nuestro corpus, el CELC-C1, hemos compilado los errores de 126 exámenes (el 50% aptos y el 50% no aptos), concretamente de los dos ejercicios de redacción de que constan. El primero es un texto de 150 palabras, redactado después de escuchar un archivo de audio; y el segundo, un texto de 250 palabras sobre un tema a elegir entre dos opciones. En conjunto suman un total de 50.400 palabras.

Respecto a la metodología aplicada para elaborar el CELC-C1 se han seguido las siguientes fases: (1) selección del corpus, (2) identificación de los errores del corpus (manualmente) y (3) clasificación de los errores. Esta clasificación se ha llevado a cabo a partir de los criterios establecidos por la CIEACOVA, que los ordena en 6 categorías. Son las que enumeramos a continuación junto al porcentaje de errores encontrados: adecuación (7,9%), coherencia (5,3%), cohesión (34,7%), corrección gramatical (35,8%), acentuación (15,8%) y precisión léxica (0,4%).

Algunas de las principales conclusiones a las que nos ha permitido llegar el estudio son, por ejemplo, que los errores hallados afectan a todos los niveles lingüísticos. Asimismo, existen tanto errores interlingüísticos (sobre todo por la interferencia del español), como intralingüísticos (omisión de las reglas ortográficas, principalmente). Así pues, se comprueba que después de casi 40 años de enseñanza en valenciano y del valenciano, existe una fuerte interferencia con el español, además de una considerable falta de dominio del registro formal de la lengua. Consideramos, pues, que investigaciones como esta son muy relevantes de cara a mejorar la formación lingüística de las personas interesadas en acreditar el nivel de conocimiento de una de las dos lenguas oficiales de la Comunitat Valenciana. Finalmente, cabe añadir que en esta primera investigación el corpus está compuesto por errores producidos en la lengua escrita. En futuras investigaciones se prevé ampliarlo con los errores producidos en la lengua oral.

Referencias

Antolí Martínez, J. M. (2018). El procés de constitució del Corpus Informatitzat de la Gramàtica del Català Modern (CIGCMod). Objectius, criteris i avaluació.

Bach, C., Saurí Colomer, R., Vivaldi, J. y Cabré, M. T. (1997). El corpus de l'IULA: descripció.

Barbasán, I. (2017). El error y su consideración en el marco de la enseñanza de lenguas extranjeras. En *Cervantes y la universalización de la lengua y la cultura españolas: actas del LI Congreso Internacional de la Asociación Europea de Profesores de Español (AEPE): celebrado en Palencia (España) del 24 al 29 de julio de 2016* (pp. 112-121). Agilice Digital.

Barbasán, I. Errores léxicos en la interlengua de aprendientes de L2. Un estudio contrastivo italiano y español. *Cuadernos de Filología Italiana, 28,* 45-71. DOI: https://dx.DOI.org/10.5209/cfit.70732

Boledai, G., Botti, S., Pobleteii, B., Castilloii, C., Fuenmayorii, M. E., Badiai, T. y Lópezii, V. (2004). CuCWeb: Un corpus del català construït a partir de la web. *dins II Congrés Online de l'Observatori per a la Cibersocietat, Barcelona.* http://www.cibersociedad.net/congres2004/index ca. (26.10. 2005).

Cabré, M.T., Bach, C. y Vivaldi, J. (2006). *10 anys del corpus de l'IULA.* Institut Universitari de Lingüística Aplicada-Universitat Pompeu Fabra.

Carrió-Pastor, M. L. (2020). Conocer la lengua a través de los corpus: La herramienta METOOL, retos para el análisis de los marcadores discursivos. *Pragmalingüística, 28*(1), 255-274.

Carrió-Pastor, M. L. y Alonso-Almeida, F. (2022). Corpus paralelos español-inglés. En *Lingüística de corpus en español/The Routledge handbook of Spanish corpus linguistics* (pp. 89-103). Routledge.

Casas-Deseures, M. y Colomé, L. C. (2015). Detecció, correcció i justificació d'errors de normativa en l'alumnat universitari: La formació lingüística dels futurs mestres. *Treballs de sociolingüística catalana*, 175-197.

Cremades Cortiella, E. (2021). Anàlisi d'errors en l'expressió escrita del català com a llengua addicional: contrast entre l'alumnat serbi i l'anglòfon. *CLIL Journal of Innovation and Research in Plurilingual and Pluricultural Education, 4*(2), 21-34. DOI: https://DOI.org/10.5565/rev/clil.65

James, C. (1998*). Errors in language and use: Exploring error analysis.* Longman.

Llauradó, A., Martí, M.A. y Tolchinsky, L. (2012). Corpus CesCa: Compiling a corpus of written Catalan produced by school children. *International Journal of Corpus Linguistics, 17*(3) 428-441.

Martínez, J. M. A. El procés de constitució del Corpus Informatitzat de la Gramàtica del Català Modern (CIGCMod). Objectius, criteris i avaluació.

Martínes, V. y Sánchez-López, E. (2014). L'ISIC-IVITRA i el metacorpus CIMTAC. Noves aportacions a la lingüística de Corpus.

Montesinos, A. I. y Romero, F. (2020). New Linguistic Corpus in Catalan and the Teaching of Languages for Specific Purposes at University. En M. L. Carrió (Ed.). En *Corpus analysis in different genres: Academic discourse and learner corpora* (pp 269-281). Routledge.

Penadés, I. (2003) Las clasificaciones de errores lingüísticos en el marco del análisis de errores. *Linred: Lingüística en la Red, 1*, 1-29.

Rodríguez, C. (2020) Anàlisi d'errors en català com a llengua estrangera en l'expressió oral d'estudiants universitaris txecs. *Estudis Romànics, 42*(83-99).

Sánchez-López, E. (2018). Panorama històric de la constitució de corpus: orígens, consolidació i expansió. *Notandum, 48,* 87-109. DOI: http://dx.DOI.org/10.4025/notandum.48.6

CSR and (un)transparent communication of equality vs. equity: A mini-diachronic corpus-based analysis

Federico Zaupa, University of Modena and Reggio Emilia

The notions of *equality* and *equity* have increasingly become of particular relevance, together with those of *diversity*, *inclusion*, and *belonging*, for the promotion of social sustainability goals and practices in management and CSR communication (e.g., Mazzei/Ravazzani 2012; Oswick/Noon 2014; Utting 2007). Although they may sound similar, the two words differ significantly in their meanings (Bronfenbrenner 1973, 9), as suggested by studies from different disciplines, including law and economics. While the former is associated with equality of opportunities, irrespective of an individual's condition and any existing disparities, the latter acknowledges the root causes of any disparities and consists in diversifying the allocation of different resources or opportunities according to the different backgrounds or needs of workers and stakeholders, to create equal opportunities (Minow 2021, 180). Although there is a tendency to promote equity (rather than equality), and complain about inequalities (Klassen 2006, 70), the "fluid" conceptualization of these two terms has allowed scholars to share the view that the two words are commonly used interchangeably. Recent studies have analysed how diversity, equity and inclusion (DEI) is discursively constructed and communicated (e.g., Malavasi 2023), but, to the best of my knowledge, there is no linguistic research that has focused more specifically on the terms *equality* and *equity*, and whether they are linguistically promoted in a transparent way.

Drawing on this background, this paper addresses the following research questions:

1) How are *equality* and *equity* discursively constructed in CSR communication?

2) Which linguistic resources are employed in CSR reports to communicate *equality* and *equity*?

3) Do these linguistic resources convey transparency or vagueness in the communication of *equality* and *equity*?

These research questions are addressed through a diachronic corpus-based analysis of a pilot corpus including 2020-23 CSR/ESG reports from international companies. Firms under investigation are chosen among the most equitable in international indices such as the Refinitiv Global Equity Index 2023, and operate in various sectors, such as automotive, banking, hardware and software, retailing, and pharmaceutical and biotechnology. Data analysis focuses on the collocational and lexico-phraseological patterns (Sinclair 2004) of the words *equality* and *equity* in their context. Closer reading of their extended concordance lines also allows to shed light on whether the companies under investigation transparently or vaguely (e.g., Schnackenberg and Tomlison 2016; Jin 2022) communicate their commitment and approach to *equality* and/or *equity*.

The findings of this study reveal that, compared to the definitions provided in the studies previously mentioned, both words tend to be used interchangeably with lexical items

denoting aspects of social cohesion: *equal** with gender issues, and *equit** with income and racial issues. Most of the instances examined in the corpus also suggest a discursive construction of both equality and equity as key values the companies are committed to. This is signaled by commitment statements, and the use of futurity in relation to the companies' effort to achieve equality and foster equitable treatments. However, the abundance of expressions conveying vagueness – including vague items associated with quantity, quality assessment adjectives and verbs, and generic references to places and time – and the collective use of these words with *diversity* and *inclusion* suggest a lack of transparency with respect to the more concrete initiatives through which companies will show their commitment to these values.

References

Bronfenbrenner, M. (1973). Equality and equity. *The ANNALS of the American Academy of Political and Social Science*, *409*(1), 9-23.

Klassen, S. (2006). What is equity? In G. Kochendörfer-Lucius & B. Pleskovic (Eds.), *Equity and development*. World Bank Publications.

Malavasi, D. (2023). The discursive construction of equality, diversity and inclusion. Insights from an analysis of CSR reports in the USA, UK and Japan. *Lingue e Linguaggi, 58*(2023), 153-171.

Mazzei, A., & Ravazzani, S. (2012). Leveraging variety for creativity, dialogue and competition, *Journal of Communication Management, 16* (1), 59-76.

Minow, M. Equality vs. equity. (2001). *American Journal of Law and Equality, 1*(2021), 167-193. DOI: https://DOI.org/10.1162/ajle_a_00019

Oswick, C., & Noon M. (2014). Discourses of diversity, equality and inclusion: Trenchant formulations or transient fashions? *British Journal of Management, 25*, 23-39.

Schnackenberg, A. K., & Tomlinson E. C. (2016). Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships. *Journal of management, 42*(7), 1784-1810. DOI: https://DOI.org/10.1177/0149206314525202

Sinclair, J. M. (2004). *Trust the text. Language, corpus and discourse*. Routledge.

Utting, P. (2007). CSR and Equality. *Third World Quarterly, 28*(4), 697–712. http://www.jstor.org/stable/20454957.

Data-driven empirical translation studies: relating error annotations and metadata in a learner corpus/ ¿Nueva terminología en el campo de los videojuegos: un estudio de corpus?

Marlén Izquierdo, University of Basque Country

Empirical approaches to translation research have traditionally focused on productoriented studies that observe and describe real world translation phenomena from a corpus approach. Echoing Olohan's call for "contextualising translation by combining corpusbased investigations with other kinds of methodologies and analyses" (2003, p. 419), many scholars have underlined the need to integrate in such empirical observations more social, contextual and cognitive data (Sutter & Lefer, 2019). In this regard, for a few years now the convergence between corpus-based and process-oriented translation studies is shaping current empirical translation studies (Kotze, 2019), thus connecting the three branches of

Translation Studies, namely, product-, process-, and function-oriented research (Holmes, 1972). This effort requires new-generation corpora that are "more carefully designed to take consideration of translators' backgrounds and the circumstances of text production" (Kotze 2020, p. 356). The aim of this study is to give an example of this approach by describing data taken from an English-Spanish (EN-ES) learner translation corpus enriched with standardized metadata related to the source text, the translation task and the learners (Granger & Lefer, 2020). It is believed that such information, relating to the process of translation, would enrich the interpretation of the translation as product. Within this framework, the study aims to answer the following questions: i) what kind of problems abound in a multiple learner translation corpus? ii) do the same chunks trigger the same or different translational errors in each translation? and finally iii) considering the learners' metadata, do any patterns stand out as indicative of a process-related product effect? Data for the analysis was taken from an English-Spanish (EN-ES) sub-corpus of the Multilingual Student Translation (MUST) Corpus/Project (Granger & Lefer 2020). This sub-corpus is a multiple translation corpus (Espunya 2014) as it features several translations into ES of the same EN source text (ST). All the textual pairs were aligned at the paragraph and sentence level and annotated for errors as well as for good translation choices. The annotation was done using the first version of the MUST-developed Translation-oriented Annotation System (TAS 1.0) (Granger & Lefer 2020). I then juxtaposed the TAS annotations suggested for each translation with the metadata referring to the learners' linguistic background and their use of CAT tools to do the translation task. Generally speaking, the findings of the study reveal some chunks in the ST that pose recurrent problems in multiple translations. The error annotations point at an abundance of content-related problems, closely followed by language issues. While it is impossible to ascertain what each error is due to, different patterns emerge if every TAS level is contrasted with the two types of metadata, learner's multilingual background -or not, and learner's reliance on technological aids. Teaching implications based on the results obtained could involve the development of data-driven activities for error preventive purposes.

References

De Sutter, G., & Lefer, M. A. (2019). On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives*. DOI: https://DOI.org/10.1080/0907676X.2019.1611891

Espunya, A. (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation, 48,* 33-43. DOI: https://DOI.org/10.1007/s10579-013-9260-1

Granger, S., & Lefer, M. A. (2020). The multilingual student translation corpus: A resource for translation teaching and research. *Language Resources and Evaluation, 54,* 1183-1199. DOI: https://DOI.org/10.1007/s10579-020-09485-6

Holmes, J. (1972). The name and nature of translation studies. In L. Venuti (Ed.), *The Translation Studies Reader* (First Edition) (pp. 172–185). Routledge.

Kotze, H. (2019). Converging what and how to find out why: An outlook on empirical translation studies. In L. Vandevoorde et al. (Eds). *New Empirical Perspectives on Translation and Interpreting* (pp. 333-371). Routledge.

Olohan, M. (2002). Corpus linguistics and translation studies: Interaction and reaction. *Linguistica Antverpiensia*, New Series- Themes in Translation Studies. DOI: https://DOI.org/10.52034/lanstts.v1i.29

Deborah, Linguist yet Professor Michael. How British corpora reflect gender-relationthrough forms of address

Michael T.L. Pace-Sigge, University of Eastern Finland

This paper investigates the use of terms of address in relation to a number of female and male names. As research as early as the 1990s (see, amongst others, Acker, 1990; Lakoff and Lakoff, 1990; Tannen, 1994; Wodak, 1996) has shown, there is a clear link in the discourse between work roles and gender. More recently, large data-sets allow research into the actual, natural language usage that highlights in how far females are 'automatically' named second after males, e.g. Wright and colleagues (2005) who consider frequency distribution. There has been a particular focus on binominals – for instance Mollin (2012) who looked at the (ir-)reversibility of binominals in English, or the corpus-based study by Motschenbacher (2013) who looks at the conjunct order in personal binomials. Looking at more recent developments, (author) (2020), investigating dissertations in the BAWE corpus, found that there appears a small-scale deliberate attempt by some authors to invert the 'natural' order. This corpus-assisted research targest *British English* use as recorded in the BNC 1994 and BNC 2014 to provide a basis for a qualitative and diachronic look at the uses of such terms like *professor, director, minister* etc. as node words. This allows to observe whether there have been any changes over a twenty-year-period. This paper will focus on male or female first names most frequently appearing as collocates in L3-L1 position and R1-R3 in order to highlight the connections between profession and gender. Where it is not clear whether the name refers to a male or female, the actual person's details has been checked with the *DuckDuckGo* search engine.

The use of these corpora provides an empirical snapshot of the choice of address employed in Britain in the 1980s/1990s and then in the early 2000s – in particular in newsprint of the time. It also gives insight in the concrete nesting (cf. Hoey, 2005) of female names as compared to male names in mainstream discourse; this enables the construction of how the readership is psychologically primed to connect positions of responsibility and learning with the idea of 'maleness'. Crucially, even small differences - like, for example, whether a title or a name are given first – indicate subliminal differences. Whereas no single area of life (academia, business, politics) does give clear parity for females and males according to this data, there is, nevertheless, clear evidence of marked progress. Over the period of 20 years, the evident gap between women named in positions of power and influence has significantly narrowed, with the ratios being a lot lower in 2014 than 1994. However, while UK politics seems to have shown the greatest move towards parity, overall, the changes are uneven and there are, in fact, areas where fewer females seem to appear in 2014 compared to 1994.

References

British National Corpus (1994). Retrieved 21 October, 2023, from http://www.natcorp.ox.ac.uk/

British National Corpus (2014). *User Manual and Reference Guide, Version 1.1.* (BNC2014). Retrieved 21 October, 2023, from http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf
Hoey, M. (2005). *Lexical priming. A new theory of words and language.* Routledge
Lakoff, R. (2003). Language, gender, and politics: Putting "women" and "power" in the same sentence. In J. Holmes & M. Meyerhoff (Eds.), *The handbook of language and gender* (pp. 161-178). Blackwell Publishing Ltd.
Lakoff, R. T., & Lakoff, R. (1990). *Talking power*. Basic Books.
Mollin, S. (2012). Revisiting binomial order in English: Ordering constraints and reversibility. *English Language and Linguistics, 16*(1), 81–103.
Motschenbacher, H. (2013). Gentlemen before Ladies? A corpus-based study of conjunct order in personal binomials. *Journal of English Linguistics, 41*(3), 212–242.
Wright, S. K., Hay, J., & Bent, T. (2005). Ladies first? Phonology, frequency, and the naming conspiracy. *Linguistics, 43*(3), 531–561.
Tannen, D. (1994). *Talking from 9 to 5: How women's and men's conversational styles affect who gets heard, who gets credit, and what gets done at work*. William Morrow and Company.
Wodak, R. (1996) Power, discourse, and styles of female leadership in school committee meetings. In *Discourse and power in educational organizations* (pp. 31-54). Hampton Press.

Deconstructing news narratives of child sexual abuse: Unveiling the underlying boundaries

Nuria Lorenzo-Dus, Swansea University

Sergio Maruenda Bataller, University of València

In recent years, the prevalence of online grooming (OG) has become a cause for significant concern worldwide. In Spain, a 2022 report by the Spanish Criminality Statistics Portal for Cybercrime indexes 954 reported cases of online child sexual grooming and 628 cases of child pornography. 84% of these types of crime have children as targets. This alarming trend underscores the importance of critically examining how online grooming is represented in news discourse in Spain but also globally. Despite the growing prevalence of OG, research on media representations of this phenomenon remains limited and fragmentEd. This paucity of research is particularly concerning given the potential influence of media portrayals on public perceptions and policy responses. Extant research on media representations of OG has predominantly used quantitatively content analysis methodologies, often failing to delve into the discursive and linguistic aspects of these representations. By analysing the linguistic choices employed in media coverage, researchers can gain valuable insights into the underlying frameworks and biases that inform these representations. While a few studies have taken a more linguistic approach (e.g., Cheit, 2003; Cheit et al. 2010), these exceptions highlight the need for a more systematic and comprehensive examination of linguistic patterns in OG discourse. The application of news values theory (Bednarek & Caple 2017) has been largely overlooked in the context of OG research. Understanding the news values that drive OG coverage provides valuable insights into the motivations and biases that shape media portrayals, enabling a more critical evaluation of these representations.

Despite the limited research, there is growing evidence to suggest that media representations of OG can significantly impact public perceptions and policy decisions. Kitzinger & Skidmore (1995) and Nair (2019) have shown how media coverage can influence public attitudes towards child protection measures and the formulation of policy frameworks. Moreover, Carson et al. (2015) highlighted the role of media representations in raising awareness of OG and enabling the identification of potentially harmful online behaviours. Despite their potential influence, media representations of OG often fall short of ideal practices. Saewyc et al. (2013) identified concerns over child blaming and stereotyping in media coverage, potentially contributing to victim-blaming attitudes and hindering reporting of actual incidents. Cheit (2003) and Cheit et al. (2010) criticised the prevalence of 'stranger danger' narratives and sensationalised reporting, which can create undue fear and anxiety among the public while potentially diverting attention from more nuanced realities for OG. Moreover, Doring & Walter (2020) noted the neglect of prevention strategies and support services for victims in media coverage, failing to provide essential information for public awareness and protection.

Driven by our extensive experience in analysing OG discourse, in this paper we scrutinize media portrayals of this crime. Our aim is to critically appraise how the media frames and presents online grooming as another form of gendered violence, identifying potential biases and stereotypes that may hinder effective prevention and intervention efforts. We therefore shed light on the linguistic choices, naming practices and narrative structures employed by the media when covering this sensitive topic. This analysis enables us to pinpoint areas where media discursive representations can be improved to promote more accurate, responsible reporting. Our findings serve as a valuable resource for policymakers, educators and social advocates, providing insights into how media portrayals can influence public perceptions and inform strategies for combating online grooming.

References

Bednarek, M., & Caple H. (2017). *The discourse of news values: How news organizations create newsworthiness*. 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. DOI: 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. DOI: 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. DOI: 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: https://doi.org//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

Dependency Directions as a Stylometric Tool: Distinguishing Genres in Czech through Syntactic Analysis

Miroslav Kubát, University of Ostrava

Xinying Chen, University of Ostrava

Czech is a language characterized by its notably flexible word order, a feature stemming from its reliance on inflectional forms (declension and conjugation) to denote syntactic relationships. This linguistic structure affords speakers a multitude of message variations. However, the arrangement of words is far from random; it adheres strictly to logical relationships among words and additional principles (cf. Siewierska & Uhlířová, 1998). The choice of word order is also determined by the speaker's communicative intent and emotional disposition, a concept referred to as functional sentence perspective (cf. Firbas, 1992).

This study deals with word order in the Czech language across diverse text types. Employing the framework of dependency syntax, it examines the variations in dependency directions (percentages of head-initial and head-final) of basic syntactic functions (subject, object, attribute adverbial) across various genres. For example, in a sentence “I eat an apple” the subject “I” is dependent on its head predicate “eat”. In this case, the predicate is positioned behind the subject, so it is head-final. The object “apple” is dependent on the predicate “eat”. In this case, the head is positioned in front of the object, so it is head-initial. These variations could pave the way for utilizing dependency directions as a syntactic index, introducing an innovative method for differentiating among text types. This approach has the potential to enhance our understanding of written Czech's syntactic details and may serve as a stylometric tool for text classification.

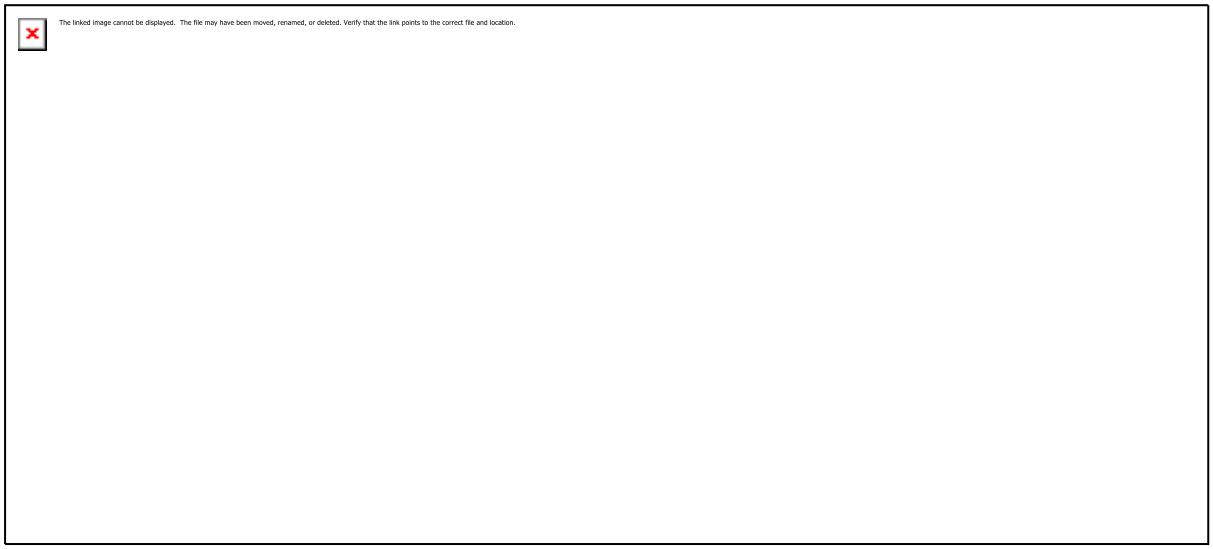
The dataset comes from the Czech National Corpus, namely the balanced corpus of contemporary written Czech SYN2020 (Křen et al., 2020). The corpus has 100 million words and covers texts mainly from 2015–2019. SYN2020 consists of three genre groups (fiction, non-fiction, newspapers and magazines) which are divided into genres such as novel, short story, poetry, drama, administrative text, scientific text, etc.). Besides lemmatization and morphological annotation, SYN2020 is also syntactically annotated (Jelínek et al., 2021). The syntactic annotation is based on the Prague Dependency Treebank (Bejček et al., 2012). SYN2020 therefore provides a rich landscape for linguistic investigation.

The results (see Table below) show that dependency directions of basic syntactic functions (subject, object, attribute adverbial) play an important role in distinguishing different genres. A distinct pattern emerges, effectively segregating the genres into fiction and non-fiction.

Notably, genres within fiction, such as poetry, novels, and drama, cluster together, contrasting with non-fiction genres like administrative and scientific texts. Meanwhile, journalistic texts exhibit considerable variation depending on the specific syntactic function

analyzEd. The analysis of attributes, in particular, yields the most promising insights. More specifically, for attributes, fiction displays a greater inclination towards a head-final structure compared to nonfiction. Journalistic texts are then in the middle between fiction and non-fiction. These observations tentatively affirm that dependency directions serve as a valuable syntactic index for stylometric applications. Extending this methodology to authorship attribution could provide intriguing insights into its efficiency in identifying authors. Furthermore, adapting and applying this approach to various languages could offer broader, cross-linguistic perspectives on its utility and adaptability.

Dependency directions (percentages of head-initial and head-final) of subject, object, attribute, adverbial in different genres.



References

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., & Žabokrtský, Z. (2012). Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012). Mumbai* (pp. 231-246).

Firbas, J. (1992). *Functional sentence perspective in written and spoken communication.* Cambridge University Press.

Jelínek, T., Křivan, J., Petkevič, V., Skoumalová, H., & Šindlerová, J. (2021): SYN2020: A new corpus of Czech with an innovated annotation. In K. Ekštein, F. Pártl & M. Konopík (Eds.), *Text, speech, and dialogue. TSD 2021. Lecture notes in computer science*, vol. 12848 (pp. 48-59). Springer.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., & Škrabal, M. (2020). *SYN2020: Representative corpus of contemporary written Czech. Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague*. http://www.korpus.cz

Siewierska, A., & Uhlířová, L. (1998). An overview of word order in Slavic languages. In A. Siewierska (Ed.), *1 Constituent Order in the Languages of Europe* (pp. 105-150). De Gruyter Mouton.

Designing a representative corpus of Maltese

Joseph Buttigieg, University of Malta

This paper aims to address the lack of representativeness in an existing corpus of Maltese by proposing an alternative design. Representativeness has been described as an elusive concept (Nelson, 2010, p. 60). It is considered an essential quality by some (Biber, 1993; Leech, 2007, p. 135) but difficult or impossible to achieve by others (Váradi, 2001, pp. 591–592; Stefanowitsch, 2020, p. 29; Hunston, 2022, pp. 33–34). For the purpose of this study, it is not understood in absolute terms but rather as "the extent to which a corpus permits accurate generalizations about the quantitative linguistic patterns that are typical in a target language or discourse domain" (Egbert et al., 2022, p. 11). Attaining the highest degree of representativeness possible would allow for the generalisation and extrapolation of results that otherwise would apply exclusively for the corpus.

Maltese is a language of Semitic origin with a Romance superstructure and mixed vocabulary due to its many contacts over 1000 years of history. As a lesser used language, it does not have a wealth of resources at its disposal, but in the last decade the situation has changed for the better. A recent important development has been the launching of Korpus Malti 4.0 (Micallef et al., 2022), consisting of 450 million tokens. Although this general corpus has addressed a huge gap in Maltese linguistic studies, it is not without its shortcomings, the main issue being lack of balance and representativeness.

In order to test the proposed design, I am building a snapshot corpus based on samples of written Maltese from 2022. The process entails (a) mapping and quantifying the written production (where possible), (b) operationalising it through establishing boundaries and strata, and (c) deciding on sample sizes and sampling methods. One of the challenges in mapping the production derives from the fact that English is an official language in Malta, and is widely used as a language of instruction, especially at post-secondary and tertiary levels. It is also the main language of communication in the industrial, financial and entertainment sectors. This negatively affects the amount of written data in Maltese available to the researcher.

The data shall be compiled according to three different designs to enable comparison in terms of lexical richness and other grammatical features. The results should indicate which corpus design permits accurate linguistic parameter estimates about written Maltese in 2022. This in turn would be a good point of departure for building a larger corpus.

References

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243–257.

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representativeness.* Cambridge University Press.
Hunston, S. (2022). *Corpora in applied linguistics* (Second edition). Cambridge University Press.
Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 133–149). Rodopi.
Micallef, K., Gatt, A., Tanti, M., van der Plas, L., & Borg, C. (2022). Pre-training data quality and quantity for a low-resource language: New corpus and BERT Models for Maltese. *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing* (pp. 90–101). DOI: https://DOI.org/10.18653/v1/2022.deeplo-1.10
Nelson, M. (2010). Building a written corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (First edition) (pp. 53–65). Routledge.
Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology.* Language Science Press.
Váradi, T. (2001). The linguistic relevance of Corpus Linguistics. In *Proceedings of the Corpus Linguistics 2001 Conference,* (pp. 587–593). https://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/papers/varadi.pdf

Detección de errores frecuentes en la escritura académica de estudiantes chilenos: tendencias para una plataforma inteligente

Anita Ferreira Cabrera, University of Concepción

Existe una brecha investigativa y educativa en torno a la delimitación de las reales competencias y precisión lingüísticas logradas en el uso de la lengua en la habilidad escrita por parte de los estudiantes de enseñanza media en el sistema educativo chileno. La producción escrita en español como lengua materna, en el ámbito educativo chileno, es un eje fundamental en las bases curriculares y programas. El objetivo principal es identificar los errores de precisión lingüística que se observan en estudiantes de segundo año de enseñanza media e identificarlos a través de la plataforma con técnicas de Inteligencia Artificial, *Mejora tu Escritura*. Se necesita contar con estudios que permitan recabar la información sobre los errores más relevantes y urgentes que requieren de apoyo por parte de los profesores y profesoras del sistema educacional chileno e incorporarlos en reconocimiento de plataformas inteligentes. Para ello, se ha diseñado un estudio descriptivo mixto cuantitativo y cualitativo con el fin de delimitar aquellos errores frecuentes que se sostienen en el tiempo. La investigación es novedosa dado que incorpora la metodología de lingüística de corpus de aprendientes de lenguas y un análisis de errores asistido por la tecnología (del inglés *Computer-aided-error-analysis (CEA*)) para la identificación de los errores.

Los corpus digitales de aprendientes se construyen a partir de textos auténticos producidos por estudiantes que están practicando la habilidad escrita. Los textos se recolectan de acuerdo a criterios específicos cumpliendo con condiciones de tipología textual, ámbito geográfico, temporalidad, etc., lo que permite que el corpus se constituya en una muestra representativa de una lengua particular o de un subconjunto de la misma. Los corpus

permiten que el investigador detecte las fortalezas y debilidades del grupo lingüístico estudiado, por lo tanto, los resultados se consideran una fuente valiosa de datos lingüísticos. Los estudios se han centrado en la identificación de frecuencias, sistematicidad y gravedad de los errores.

Para la recopilación del corpus, se desarrollaron tres actividades de escritura en diferentes instancias en el contexto de la asignatura de Lengua y Literatura. Los textos fueron etiquetados y procesados a través de UAM Corpus Tools. Los resultados arrojan las siguientes tendencias más relevantes: (1) la omisión de la coma en enunciados, (2) la omisión de la tilde en palabras agudas, (3) la omisión de la tilde en palabras esdrújulas, (4) la omisión de la tilde dierética y (5) diacrítica, (6) la alternancia de los grafemas s/c/z, (7) la omisión de mayúsculas, (8) la sustitución de conectores aditivos y (9) la omisión del grafema h. En conclusión: Se sugiere como orientación para el profesorado enfocarse en estos errores en sus revisiones textuales y así contribuir en los procesos de corrección y mejoramiento de la habilidad escrita en español como L1.

El proyecto de investigación en que se circunscribe este trabajo tiene como objetivo principal identificar los errores de precisión lingüística con el propósito de desarrollar una plataforma inteligente para detectar errores focalizados y apoyar con corrección metalingüística provista por los profesores. Se ha desarrollado una plataforma inteligente, *Mejora tu Escritura* basada en Lingüística de Corpus de aprendices y en técnicas de Inteligencia Artificial de procesamiento de lenguaje natural para reconocer los errores más frecuentes en los textos de los estudiantes y apoyar con retroalimentación metalingüística por parte de los profesores.

References

Ferreira Cabrera, A. (2022). Diseño e implementación del corpus de aprendientes de español como lengua extranjera (CAELE). *Círculo de Lingüística Aplicada a la Comunicación, 90,* 137-154.

Ferreira Cabrera, A. (2023). Dificultades en la precisión ortográfica de palabras agudas en la escritura académica de estudiantes del sistema educacional chileno. *Revista Sophia Austral, 29*(2023), 1–24.

Ferreira Cabrera, A. (2023). Carencias en el uso de la coma en la producción escrita de estudiantes chilenos de segundo año de educación media. *Boletín De Filología*, *58*(2), 290–318.

Granger, S. (2017). Learner corpora in foreign language education. En S. Thorne y S. May (Eds.). *Language and Technology. Encyclopedia of Language and Education* (Third edition) (pp. 427-440). Springer International Publishing.

Granger, S. (2012). How to use foreign and second language learner corpora. En A. Mackey y S. M. Gass. (Eds.). *Research methods in second language acquisition: A practical guide* (pp. 7-29). Blackwell Publishing.

O'Donnell, M. (2022). *UAM Corpus Tool, 6.2. Software de computador*. http://www.corpustool.com/index.html

Digital discourse on TripAdvisor: a genre analysis of negative hotel reviews written in Italian, French and Spanish

Irene Cenni, Ghent University

The advent of the Web 2.0 brought about a significant transformation in the tourist experience and in its modes of communication. One of the prevailing genres in today's digital tourism discourse is represented by online reviews, produced and consumed daily by millions of users on global platforms such as TripAdvisor, Booking.com or Airbnb (Vásquez, 2011; Mariottini & Hernández Toribio, 2017). Online tourist reviews are mainly used to share a personal (travel) experience with service providers and fellow travelers, at the same time they offer an evaluation of tourism services.

The aim of the present case study is to provide an in-depth investigation of how tourists are sharing their post-trip experiences online from a linguistic and pragmatic standpoint. Specifically, we focused on negative hotel reviews posted on TripAdvisor and we adopted a cross-linguistic perspective.

Building on previous work, we now focused on a different set of languages, considering parallel corpora of reviews written in Italian, French and Spanish (N100 for each language, for a total of N300) posted between 1st of January 2021 and 31st December 2022.

Ultimately, the aim was to verify whether users writing in different languages share their experiences adopting uniform linguistic norms and communicative habits or display different discursive preferences. In order to reach this goal, we analyzed not only which topics tourists include in their reviews (Ekiz et al.2012), but also which communicative moves they adopted to share their experiences and opinions, applying the genre analysis framework (Swales, 1990).

Results indicate a tendency towards a cross-linguistic similarity. This finding seems to point toward a potential standardization process of how travel experiences are shared online in different languages (Piccioni, 2014), within the highly popular genre of online reviews.

References

Ekiz, E., Khoo-Lattimore, C., & Memarzadeh, F. (2012). Air the anger: Investigating online complaints on luxury hotels. *Journal of Hospitality and Tourism Technology, 3,* 96–106.

Mariottini, L., & Hernández Toribio, M. I. (2017). La narración de experiencias en TripAdvisor. *Rilce*, *33*(1), 302-330.

Piccioni, S. (2014). Cortesía y lenguas de especialidad entre lo local y lo global: El caso de las reseñas de hoteles en español e inglés. *NORMAS. Revista de Estudios Lingüísticos Hispánicos, 4,* 93-116.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Vásquez, C. (2011). Complaints online: The case of Tripadvisor. *Journal of Pragmatics, 43,* 1707-1717.

Diseño y elaboración de un corpus textual divulgativo de reproducción asistida para evaluar el nivel de percepción de las pacientes

Ana Reyes Herrero, University of Alicante

Diversos estudios (Blanco Pérez y Gutiérrez Couto, 2002; Porras-Garzón y Estopà, 2020; Martínez Sánchez, 2022) que han medido el grado de legibilidad de los textos médicos han demostrado que no son lo suficientemente comprensibles para el público general que consulta información sobre salud. En el caso de la reproducción asistida (RA), nos encontramos ante un ámbito que, al igual que otras especialidades biomédicas, está cargado de lenguaje especializado como por ejemplo epónimos, abreviaciones, extranjerismos, neologismos (Aleixandre-Benavent et al., 2017: 24), que son conocidos para los profesionales, pero no para el público general. Por lo tanto, cuando una paciente se está realizando un tratamiento de reproducción asistida se enfrenta a una gran carga física, psicológica, económica y social, a lo que hay que añadir la gran cantidad de información especializada que recibe: tratamientos, test adicionales y analíticas, pautas de medicación, consentimientos informados, etc. Nos encontramos ante una paciente que se enfrenta, en la mayoría de los casos, a una información nueva y desconocida, lo que puede incrementar su nivel de estrés al no comprender correctamente una información que, cuando se está realizando un tratamiento, es esencial.

De este modo y en el marco la tesis doctoral “Análisis del discurso biomédico: la difusión oral y escrita de textos sobre Reproducción asistida a través de Internet y redes sociales” (I-PI 108-22) en colaboración con la clínica de RA IVF-Life, el objetivo principal es crear un corpus en español de los textos divulgativos publicados por la clínica en su página web y en su canal de YouTube con el fin de medir su grado de legibilidad y evaluar el nivel de percepción de la terminología especializada de la RA por parte de las pacientes que están sometiéndose a un tratamiento de fertilidad en esta clínica.

La medicina y la lingüística son dos disciplinas que están estrechamente relacionadas (Estopà y Lorente, 2022) y cada cambio que experimenta la primera, supone un cambio lingüístico, pues es necesario denominar y representar las nuevas realidades que surgen (Cabré, 1993). Dentro del ámbito tan amplio y complejo que es la medicina y las ciencias de la salud, nos encontramos con muchas especialidades, como por ejemplo la reproducción asistida (RA). La reproducción es una cuestión que nos ha acompañado desde el inicio de los tiempos y es un sector especialmente intenso en España, sobre todo en la Comunidad

Valenciana, una de las Comunidades pioneras en la implantación de tratamientos de RA (Martínez-Martínez y Bote, 2019, p. 591-592).

Como resultado, hemos obtenido un corpus monolingüe formado por dos subcorpus: (1) un subcorpus de textos escritos extraídos de tres secciones (*Tratamientos de Fertilidad*, *Pruebas de Fertilidad* y *Blog)* de la página web de la clínica y (2) un subcorpus de textos orales compuesto por una serie de vídeos del canal de YouTube de la clínica, que posteriormente se han pasado a un formato escrito con el programa de reconocimiento de voz *Amazon Transcribe.*

Referencias

Aleixandre-Benavent, R., Bueno Cañigral, F. J. y Castelló Cogollos, L. (2017). Características del lenguaje médico actual en los artículos científicos. *Edición Médica, 18*(2), 23-29.

Blanco Pérez, A. y Gutiérrez Couto, U. (2002). Legibilidad de las páginas web sobre salud dirigidas a pacientes y lectores de la población general. *Revista Española de Salud Pública*, *76*(4), 321-371. DOI: https://DOI.org/10.1590/S1135-57272002000400007

Cabré Castellví, M. T. y Sager J. C. (1993) *La terminología: Teoría, metodología, aplicaciones*. Empuries.

Estopà, R. y Mercè, L. (Eds.) (2022), *La terminología, espejo de la evolución del conocimiento científico: El caso de la reproducción asistida*. Barcelona: Institut de Lingüística Aplicada de la Universitat Pompeu Fabra & Documenta Universitària, Sèrie Monografies 15.

Martínez-Martínez, A, L. y Bote Díaz, M. A. (2019). Concilia o revienta: Determinantes socioeconómicos y demográficos del uso de técnicas de reproducción humana asistida en perspectiva territorial. *Política y Sociedad, 56*(3) 583-601. DOI: https://DOI.org/10.5209/poso.60510

Martínez Sánchez, O. (2022). *La alfabetización en salud: Un análisis del discurso de la reproducción asistida en páginas web* [Trabajo final de máster, Universidad de Alicante] Repertorio Institucional de la Universidad de Alicante. http://hdl.handle.net/10045/133850

Porras-Garzón, J. M. y Estopà, R. (2020). Escalas de legibilidad aplicadas a informes médicos: Límites de un análisis cuantitativo formal. *Círculo de Lingüística Aplicada a la Comunicación*, *83*, 205-216. https://DOI.org/10.5209/clac.70574

El acusativo preposicional en catalán al final de la Edad Moderna (1833-1903). Un estudio de corpus

Josep E. Ribera i Condomina, University of València

Este trabajo aborda el fenómeno del marcaje diferencial de objecto (MDO) en el catalán del siglo XIX a partir del *Corpus textual informatitzat de la llengua catalana* (CTILC) y discute algunos de los problemas de la extracción automatizada de datos para el análisis de esta construcción, situada en la interfaz entre la sintaxis, la semántica, el discurso y la pragmática.

Los estudios sobre el MDO en una variedad de lenguas, que incluyen el español, el rumano, el portugués y, más recientemente, el catalán (Pineda, 2021, 2023), se han centrado en dos tipos de restricciones jerarquizadas en tres escalas: la animacidad (ia) y la referencialidad, relacionada con la categoría léxica (ib), y el grado de afectación impuesto por el verbo sobre el objeto (von Heusinger y Kaiser, 2011), relacionado con la escala de transitividad (Tsunoda, 1985) (ii).

(i) a. humanos > animados > inanimados

b. pronombre personal > nombre propio > SN definido > SN específico indefinido > SN indefinido no específico > no argumental

(ii) acción efectiva (*matar*) > percepción (*veure*) > actividad encaminada a una meta (*cercar*) > conocimiento (*conèixer*) > sentimiento (*témer*) > relación (*tenir*) > capacidad (*ser capaç de*)

En catalán de la Edad Moderna, el MDO solo se produce con objetos directos animados. Los objetos situados en posiciones altas de las escalas son más proclives al MDO (1) que los situados en las posiciones más bajas (2).

(1) un marit va matà *a la seva dona* (ü animado, ü SN definido, ü *matar*)

(2) Mas tu tem *Déu* (ü animado, ü nombre propio, X *témer*)

Este estudio analiza cualitativa y cuantitativamente los factores de la animacidad, la categoría léxica y el grado de afectación del objeto directo (variables explicativas) para condicionar el MDO (variable dependiente). A partir del análisis de las ocurrencias del MDO con verbos representativos de los cinco primeros grados de la escala de transitividad o afectación —los que son susceptibles de recibir el MDO— nos proponemos responder a las cuestiones siguientes:

a) ¿En qué medida la frecuencia del MDO se puede explicar a partir de las escalas de animacidad y de referencialidad?

b) ¿En qué medida la frecuencia del MDO responde al condicionamiento previsto en la escala de afectación?

c) ¿Cómo interaccionan los tres factores?

Puesto que la animacidad y la referencialidad no están lematizadas en el corpus, este trabajo combina la obtención automatizada de los datos, a partir de la información léxica lematizada, con la discriminación manual de los objetos directos léxicos animados. Así, los contextos potenciales con MDO se han restringido a los objetos postverbales adyacentes al verbo, preposicionales (lema V + *A*) o no preposicionales (lemas V + SUBSTANTIVO y V + ARTÍCULO). Ello permite reducir los miles de ocurrencias que proporciona el corpus y optimizar el análisis manual.

Los resultados muestran que la categoría léxica del objeto constituye el principal factor que condiciona el MDO. El factor del grado de afectación del objeto previsto por la escala de Tsunoda (1985) resulta también significativo, sobre todo en los casos de SN definidos y SN indefinidos. Sin embargo, la jerarquía de la escala de transitividad se ve alterada por otros factores como la agentividad potencial del objeto.

Referencias

Pineda, A. (2021). The development of DOM in the diachrony of Catalan: (Dis)similarities with respect to Spanish". En J. Kabatek, P. Obrist y A. Wall (Eds.), *Differential Object Marking in Romance: The Third Wave* (pp. 243-277). De Gruyter. DOI: https://doi.org//10.1515/9783110716207-009

Pineda, A. (2023). L'acusatiu preposicional en català: d'on venim i cap a on anem? *Caplletra, 74*, 149-182. DOI: https://doi.org//10.7203/Caplletra.74.26040

Tsunoda, T. (1985). Remarks on transitivity. *Journal of Linguistics, 21,* 385–396.

von Heusinger, K. y Kaiser G. A. (2011): Affectedness and differential object marking in Spanish, *Morphology, 21,* 593–617. DOI: https://doi.org//10.1007/s11525-010-9177-y

El aspecto léxico de pacientes con Alzheimer: un análisis de corpus desde la Gramática del Papel y la Referencia

Alejandro Suárez Rodríguez, University of Las Palmas de Gran Canaria

En esta comunicación, ofrecemos un estudio estadístico de la distribución y la frecuencia relativas del *Aktionsart* o aspecto léxico en los predicados verbales producidos por pacientes españoles con Alzheimer, por medio de las transcripciones mostradas en el corpus de Peraita y Grasso (2010). El alzhéimer suele estar dividido en tres etapas (temprana, intermedia y avanzada) que dependen del grado de deterioro cognitivo y lingüístico. Con esto, los objetivos de esta investigación son, por un lado, comprobar el tipo de verbo más frecuente en los pacientes españoles con Alzheimer y cómo se distribuyen estos *Aktionsarten*. Por otro lado, nos proponemos comparar la frecuencia y distribución en las distintas etapas, así como comprobar si el aspecto léxico puede ayudar a identificar la etapa en la que se encuentran los pacientes.

Dividimos el método de estudio en dos partes. Primero, hemos seleccionado el corpus compilado por Peraita y Grasso (2010) como parte de una investigación neuropsicológica. En ella, analizan el deterior léxico-semántico de pacientes españoles y argentinos con Alzheimer gracias a seis categorías semánticas: perro, pino y manzana (seres vivos) y coche, pantalón y silla (seres no vivos; Peraita y Grasso, 2010: 204). El corpus contiene las transcripciones de 211 españoles y argentinos, si bien nos hemos concentrado en aquellos pacientes españoles que aparezcan en las tres etapas; sin embargo, en el corpus solo existe registro de pacientes españoles en las etapas temprana e intermedia. Al acotar la muestra de pacientes, encontramos que estos emitieron 1459 predicados verbales en conjunto, de los que analizaremos una muestra de 285 verbos de la etapa temprana y 189 de la intermedia (con un intervalo de confianza del 95%; López-Roldán y Fachelli, 2015).

En segundo lugar, usamos la Gramática del Papel y la Referencia (GPR; Van Valin y LaPolla, 1997; Van Valin, 2005) para determinar el tipo de *Aktionsarten*, pues es una teoría funcionalista que intenta explicar la relación entre la sintaxis, la semántica y la pragmática.

En concreto, la representación semántica de la GPR parte del predicado verbal y adapta la clasificación del aspecto léxico o *Aktionsart* de Vendler (1967). Para determinar qué tipo de *Aktionsart* debemos asignar a los verbos, la GPR ofrece ocho pruebas (Van Valin, 2005; González Vergara, 2006; Cortés Rodríguez, González Vergara y Jiménez Briones, 2012; Van Valin, 2018) que, aunque no son infalibles, si las aplicamos secuencialmente, nos permiten identificar el *Aktionsart* de cada predicado verbal. Por último, aplicamos estas pruebas a la muestra del corpus de Peraita y Grasso (2010), además de que calculamos la frecuencia y la distribución de los diferentes *Aktionsarten* mediante estadísticos básicos: media, mediana, moda, rango, desviación típica, varianza y coeficiente de variación.

Al analizar estas muestras por separado y en conjunto, observamos que los estados son el tipo de *Aktionsart* más frecuente en las dos etapas, seguidos de las actividades y las realizaciones activas. La distribución nos muestra que los pacientes usan sistemáticamente los estados y las actividades, en detrimento del resto de verbos. Asimismo, observamos cómo los estados y los verbos causativos decrecen cuanto mayor es el deterioro cognitivo. Aunque se necesita mayor investigación, los resultados de estas muestras indican una mejor identificación de las etapas a partir del aspecto léxico.

Referencias

Cortés Rodríguez, F., González Vergara, C. y Jiménez Briones, R. (2012). Las clases léxicas. Revisión de la tipología de predicados verbales. En R. Mairal Usón, L. Guerrero and C. González Vergara, (Eds.), *El funcionalismo en la teoría lingüística: La gramática del papel y la referencia* (pp. 59-84). Ediciones Akal.

González Vergara, C. (2006). La gramática del papel y la referencia: Una aproximación al modelo. *Onomázein*, *14*(2), 101-140.

López-Roldán, P. y Fachelli, S. (2015). *Metodología de la investigación social cuantitativa*. Universidad Autónoma de Barcelona.

Van Valin, R. D. y LaPolla, R. (1997). *Syntax: Structure, meaning and function*. Cambridge University Press.

Van Valin, R. D. (2005). *Exploring the syntax-semantics interface*. Cambridge University Press.

Van Valin, R. D. (2018). Some issues regarding (active) accomplishments. En R. Kailuweit, L. Künkel y E. Staudinger, (Eds.), *Applying and expanding role and reference grammar* (pp. 71-94). Freiburg Institute for Advanced Studies, Albert-Ludwigs-Universität Freiburg.

Vendler, Z. (1967). *Linguistics in philosophy*. Cornell University Press.

El discurso biosanitario en torno a la salud de la mujer: un análisis asistido por corpus de los recursos metadiscursivos y evaluativos

Giovanni Garofalo, University of Bergamo Studies

Luisa Chierichetti, University of Bergamo Studies

Esta propuesta de comunicación se sitúa en la línea de estudios anteriores sobre la construcción del discurso biosanitario (Bordelois, 2009; Gutiérrez Rodilla 1998, 2005;

Merayo Pérez, 2014, Rodríguez Arce, 2008) y sobre la relación entre discurso y salud (Bañón Hernández 2018) y salud y género (Esteban Galarza, 2006, Cardozo Rufo, 2022), para explorar los recursos metadiscursivos y valorativos característicos de la comunicación institucional sobre la salud femenina.

En concreto, nos centramos en un corpus de documentos institucionales (de casi un millón de palabras) recabados del portal del Ministerio de Sanidad de España. El conjunto de estos documentos –primer núcleo del subcorpus de documentos institucionales del proyecto DISBIOCOM[1]– constituye una colonia de géneros (Bhatia 2002, 2004) producidos en contextos socio-retóricos parecidos e interrelacionados (informes, estudios cualitativos, guías, folletos informativos, estándares y recomendaciones, recopilaciones de buenas prácticas, etc.), destinados ya sea al público femenino más amplio o a los profesionales del ámbito bio-sanitario. En su conjunto, estos documentos manifiestan la postura oficial del Ministerio de Salud con respecto a los temas tratados y se prestan para indagar tanto en el posicionamiento discursivo (*stance-taking*) de este actor público como en los recursos evaluativos (*appraisal*) privilegiados por este en su 'canal oficial' en la red.

De hecho, se consideran ya superadas las perspectivas anteriores que veían estas clases de textos como manifestación de un discurso esencialmente 'proposicional y expositivo', a saber, como mero vehículo de 'contenidos científicos' comunicados de forma aséptica. Al hilo de la literatura existente sobre valoración (Martin y White, 2005, Alba-Juez y Thompson, 2014, Mackenzie y Alba-Juez 2019), creemos que existen muy pocos discursos –o tal vez ninguno– que puedan considerarse totalmente vacíos de rasgos evaluativos y los textos del corpus de estudio no son una excepción. Para describir los significados evaluativos del emisor institucional, su presencia en el discurso y su intento por implicar al destinatario, se adopta una metodología basada en corpus y guiada por él y se propone, por un lado, un análisis cuali-cuantitativo de los recursos previstos por el modelo interpersonal de metadiscurso (Hyland, 2004; 2005; Hyland y Tse, 2004) y, por otro, un rastreo de los rasgos léxico-gramaticales adscribibles al dominio semántico de la 'implicación' (*engagement*) postulado por la Teoría de la Valoración (Martin y White, 2005). La combinación de estos dos enfoques analíticos ayuda a subsanar algunas carencias del modelo de metadiscurso de Hyland, que contempla solo la evidencialidad citativa y no incluye la directa/indirecta (atestiguada, reproducida o inferida, Dendale y Tasmowski, 2001) como categoría analítica del metadiscurso interpersonal.

Los primeros resultados de esta investigación en ciernes parecen indicar que el metadiscurso observable en el corpus se relaciona con el propósito indirectamente persuasivo del discurso biosanitario institucional, que apunta a fomentar conductas virtuosas, máxime en la esfera de la salud sexual. Por otra parte, una marcada prevalencia de estructuras de contracción dialógica (principalmente la negación y los enlaces concesivos, asociados a los valores de 'refutación') dejan entrever un posicionamiento dialogístico del emisor que busca confutar creencias arraigadas, sobre todo aquellas relacionadas con comportamientos de la esfera sexual y reproductiva, erróneamente considerados como *habitus* o 'módo dóxico' (Bourdieu, 2000) de vivir la sexualidad.

Referencias

Alba-Juez, L. y Thompson, G. (Eds.) (2014). *Evaluation in context.* John Benjamins.
Bañón Hernández, A. M. (2018). *Discurso y salud. Análisis de un debate social.* Ediciones Universidad de Navarra.
Bhatia, V. K. (2002). Applied genre analysis: A multi-perspective model. *Ibérica, 4*, 3-19.
Bhatia, V. K. (2004). *World of written discourse. A genre-based view.* Continuum.
Bordelois, Y. (2009). *A la escucha del cuerpo. Puentes entre la salud y las palabras.* Libros del Zorzal.
Bourdieu, P. (2000). *La dominación masculina.* Anagrama.
Cadorzo Rufo, V. (2022). Género y salud. Análisis de la incorporación de la perspectiva de género en el Sistema Nacional Integrado de Salud en Uruguay (2005-2020). Avances y desafíos pendientes. *Rev Méd Urug, 38*(1), e38112. http://www.scielo.edu.uy/pdf/rmu/v38n1/1688-0390-rmu-38-01-e912.pdf
Dendale, P. y Tasmowski, L. (2001). Introduction: evidentiality and related notions. *Journal of Pragmatics, 33*(3), 339-348.
Esteban Galarza, M. L. (2006). El estudio de la salud y el género: Las ventajas de un enfoque antropológico y feminista. *Salud Colectiva, 2*(1), 9-20.
Gutiérrez Rodilla, B. (1998). *La ciencia empieza en la palabra. Análisis e historia del lenguaje científico.* Península.
Gutiérrez Rodilla, B. (2005). *El lenguaje de las ciencias.* Gredos.
Hyland, K. (2004). Disciplinary interactions: Metadiscourse in L2 postgraduate writing. Journal of *Second Language Writing, 13*, 133-151.
Hyland, K. (2005). *Metadiscourse.* Continuum
Hyland, K. y Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics, 25*(2), 156-177.
Mackenzie, J. L. y Alba-Juez, L. (Eds.) (2019). *Emotion in discourse.* John Benjamins.
Martin, J. R. y White, P. R. R. (2005). *The language of evaluation.* Palgrave/Macmillan.
Merayo Pérez, A. (2014). *La comunicación con el paciente.* Elsevier.
Ministerio de Sanidad (s.f.) https://www.sanidad.gob.es/ciudadanos/enfLesiones/enfTransmisibles/sida/prevencion/home.htm
(12/12/2023) https://www.sanidad.gob.es/organizacion/sns/planCalidadSNS/equidad/saludGenero/saludSexualReproduccion/home.htm(12/12/2023) https://www.sanidad.gob.es/organizacion/sns/planCalidadSNS/equidad/saludGenero/home.htm (12/12/2023)
Rodríguez Arce, M. A. (2008). *Relación médico-paciente.* Ed. Ciencias Médicas.

El efecto de definitud en español: nuevos datos dialectales y viejos enfoques teóricos

Jorge Agulló, University of Vienna

1. INTRODUCCIÓN: ÁMBITO Y OBJETIVOS Las construcciones existenciales con haber en español son sensibles al denominado efecto de definitud o de cuantificación de Milsark (1974, 1977), que restringe la presencia en la posición de pivote (en cursiva en los ejemplos) de pronombres personales (1a), nombres propios (1b) y, en general,

constituyentes definidos (1c) —siguiendo la Jerarquía de Definitud de Aissen (2003) (cfr. Farkas (2000, 2002))—, así como constituyentes cuantificados (1d):

(1) a. *Hay él en la habitación.

b. *Hay Juan en la habitación.

c. *Hay el niño en la habitación.

d. *Hay cada libro en la habitación.

El efecto de definitud, según nos es conocido, despliega una amplia variación interlingüística (cfr. Leonetti 2008; La Fauci y Loporcaro 1997; Bentley 2013, 2015; McNally 1997, 2016): en tanto que el español (cfr. los datos de (1)) o el francés (cfr. Bouchard 1997 o Paykin y Van de Velde 2021) observan una versión fuerte del efecto, el catalán (cfr. Rigau 1988 y Brucart y Rigau 2006) o el italiano se adhieren a una versión más o menos laxa o, para algunos autores, inexistente. Los datos de variedades dialectales de contacto entre el español y el catalán, sin embargo, permanecen aún sin explorar. Datos como los de (2), que muestran pivotes definidos y específicos, han eludido toda descripción:

(2) a. Te obligaban a hablar en castellano, cuando había Franco (COSER, La Serra [Barcelona]).

b. Hay los balcones abiertos (COSER, La Serra [Barcelona]).

c. Hasta cuando había mi hermano vivo (COSER, Torregrossa [Lleida]).

d. Si no hubiera habido mi madre (COSER, Torregrossa [Lleida]).

Así, el objetivo de este estudio es (i) aducir nuevos datos de las variedades del español en contacto con catalán, que servirán para sustentar la hipótesis del debilitamiento u obsolescencia del efecto de definitud en tales variedades de contacto, y (ii) trazar el área dialectal que corresponde a estos usos.

2. METODOLOGÍA El trabajo se basa en la observación de un corpus significativo (N = 4996) de construcciones existenciales obtenidas a partir del Corpus Oral y Sonoro del Español Rural (COSER) (FernándezOrdóñez 2005-), que me ha permitido trazar una dialectología de las construcciones existenciales en el español europeo y clasificarlas en relación con la definitud del pivote. La medición de las frecuencias por medio de R programming language (R Core Team) y su cartografiado por medio de QGIS (QGIS Association) suceden a la constitución del corpus.

3. RESULTADOS Y DESARROLLOS TEÓRICOS Los resultados de la investigación sustentan un trazado nítido del área dialectal que corresponde a los usos de (2), que, en la

cartografía, se muestran arrinconados en Cataluña, el archipiélago balear, zonas colindantes de Aragón y Levante, donde tanto menos frecuente es el fenómeno cuanto más meridional es el enclave. La influencia del catalán es, por ende, evidente, lo cual aconseja desarrollar una explicación en el ámbito del contacto de lenguas.

La hipótesis que sustentaré es que el debilitamiento del efecto de definitud (i. e. las variantes de (2)) en las variedades de contacto se debe al surgimiento de las denominadas interdialect variants o 'variantes interdialectales' (e. g. Trudgill 1989a, 1989b, 1992; Tuten 2001, 2006). Las variantes interdialectales, como las formas de (2), reorganizan o reordenan formas y rasgos de las lenguas o dialectos que contribuyen a la variedad bajo estudio. Las variedades de contacto entre el español, que respeta el efecto de definitud (cfr. los datos de (1)), y el catalán, que permite sustantivos definidos y específicos, nombres propios e incluso pronombres personales (cfr. Villalba (2016)), vulneran o sortean el efecto de definitud y muestran, en consecuencia, la obsolescencia de la restricción gramatical en situaciones de contacto.

Referencias

Aissen, J. (2003). Differential object marking: Iconicity vs. Economy. *Natural Language & Linguistic Theory,* 435-483.

Bentley, D. (2013). Subject canonicality and definiteness effects in Romance there-sentences. *Language,* 675-712.

Bentley, D. (2015). Definiteness effects and linking. En D. Bentley, F. M. Ciconte y S. Cruschina (Eds.), *Existentials and locatives in Romance dialects of Italy* (pp. 161-216). Oxford University Press.

Bouchard, D. (1997). L'effet existentiel. En J. Auger & Y. Rose (Eds.), *Explorations du lexique* (pp. 31- 45). CIRAL.

Brucart, J. M. y Rigau, G. (2006). La quantificació. En J. Solà, M. R. Lloret, J. Mascaró y M. Pérez Saldanya (Eds.), *Gramàtica del català contemporani: Vol. 2. Sintaxi* (pp. 1517-1589). Editorial Empuries.

Farkas, D. F. (2000). Varieties of definites. https://people.ucsc.edu/~farkas/papers/definites.pdf

Farkas, D. F. (2002). Varieties of indefinites. En B. Jackson (Ed.), *Semantics and linguistic theory (SALT)* (Vol.12) (pp. 59-83). Cornell University.

La Fauci, N. y Loporcaro, M. (1997). Outline of a theory of existentials on evidence from Romance. *Studi Italiani di Linguistica Teorica e Applicata, 26*, 5-55.

Leonetti, M. (2008). Definiteness effects and the role of the coda in existential constructions. En Müller, H. Høeg y A. Klinge (Eds.*), Essays on nominal determination. From morphology to discourse management* (pp. 131-162). John Benjamins Publishing Company.

McNally, L. (1997). *A semantics for the English existential construction.* Routledge.

McNally, L. (2016). Existential sentences crosslinguistically: Variations in form and meaning. *Annual Review of Linguistics, 2*, 211-231.

Milsark, G. L. (1974). *Existential Sentences in English* [Massachussetts Institute of Technology]. https://dspace.mit.edu/bitstream/handle/1721.1/13021/26114819-MIT.pdf?sequence=2

Milsark, G. L. (1977). Towards an explanation of certain peculiarities of the existential construction in English. *Linguistic Analysis, 3,* 1-30.

Paykin, K. y Van de Velde, D. (2021). La possession inaliénable et le verbe avoir existentiel. En P. Lauwers, K. Paykin, M. Ilioaia, M. Meulleman y P. Hadermann (Eds.), *Quand le*

syntagme nominal prend ses marques. Du prédicat à l'argument (pp. 173-188). Éditions et presses universitaires de Reims.
Rigau, G. (1988). Strong pronouns. *Linguistic Inquiry, 19*(3), 503-511.
Trudgill, P. (1989a). Contact and isolation in linguistic change. En *Language change: Contributions to the study of its causes* (pp. 227-237). Mouton de Gruyter.
Trudgill, P. (1989b). Language contact and simplification. *Nordlyd, 15,* 115-121.
Trudgill, P. (1992). Dialect typology and social structure. En E. H. Jahr (Ed.), *Language contact: Theoretical and empirical studies* (pp. 195-211). De Gruyter.
Tuten, D. N. (2001). Modeling koineization. En L. J. Brinton y D. Lundström (Eds.), *Historical Linguistics* 1999*. Selected papers from the 14th International Conference on Historical Linguistics, Vancouver, 9-13 August 1999* (pp. 325-336). John Benjamins Publishing Company.
Tuten, D. N. (2006). Koineization. En C. Llamas, L. Mullany y P. Stockwell (Eds.), *The Routledge companion to sociolinguistics* (pp. 185-191). Routledge.
Villalba, X. (2016). Definiteness effect, pronouns and information structure in Catalan existentials. En S. Fischer, E. Rinke y T. Kupisch (Eds.), *Definiteness effects: Bilingual, typological and diachronic variation* (pp. 172-209). Cambridge Scholars Publishing.

El potencial de los corpus en el aula de chino: estado actual y uso complementario con ChatGPT

Lingzhi Nie, Autonomous University of Barcelona

El desarrollo de los corpus no sólo aporta nuevos métodos para la investigación lingüística, sino que también proporciona nuevas herramientas para la enseñanza de lenguas extranjeras. Ante las necesidades innovadoras de la enseñanza internacional del chino y las necesidades de la enseñanza del chino con fines específicos, es imperativo resumir la aplicación práctica de los corpus en la enseñanza del chino para fomentar su uso. Basándose en la investigación existente sobre la enseñanza de idiomas con corpus y teniendo en cuenta las características de los corpus chinos de uso frecuente, este artículo explora la aplicación actual de diferentes tipos de corpus en la enseñanza internacional de chino a alumnos de diferentes niveles y orígenes culturales. Dado que el desarrollo de Internet y de la inteligencia artificial ofrece nuevas oportunidades para su aplicación pedagógica, tras resumir las ventajas de la integración de corpus en el aula de chino como lengua extranjera con la tendencia de desarrollo de corpus en el contexto de big data, presentamos nuestras ideas y sugerencias para la futura construcción y aplicación de corpus en la práctica docente.

English travels by train: a corpus linguistic analysis of an Italian on-board magazine

Valentina Di Francesco, University of Ferrara

The use of English in non-English tourism communication may be considered an empirical indicator of the status of English as a global language. Tourism discourse can count on two different macro-systems in which, on the one hand, tourism is seen as promotional language, and on the other hand is seen as a language for specialised purposes(Gotti 2003, 2006; Maci 2020).As with other types of discourses, tourism discourse can be found ina large varietyof genres such as brochures, tourist guides, advertisement, travel agencies and hotels communications, travel websites and the press. The promotion of a country can also 'travel' through on-board magazines like inflight magazines which travellers can read for free on the airplanes or in the airport lounge.

Partially inspired by Thurlow and Jaworski's work (2003) on inflight magazines, I will propose an analysis of the language which is used in *La Freccia*, the on-board magazine of the Italian Railway company *Trenitalia* (FS). As reported on the English version of the FS website "over time, the monthly news magazine has developed a strong identity, becoming a highly-recognisable brand, also thanks to its ability to offer inspiration for trips to take in tune with the desires and tastes of a curious travelling public attracted by cultural, artistic and musical stimuli"(FS Group, 2024).The available digitalized material, which covers the years from 2010 and 2024, is collected in a small corpus which is built for the purpose(approximately 200 units; each magazine consists of 100 pages, approximately).Although the articles are mainly written in Italian, I observed a frequent use of Anglicisms. Moreover, in some editions of the magazine some articles are written in Italian and English. It may mean that English is used for specific reasons such as to give the magazine an international flavour, or to be in line with the idea that the use of English words can be perceived as trendy and/or stylish, etc.

Drawing from Macy's (2020) remarks on inflight materials and on Sinclair's (1991) principles, I use corpus linguistics to detect the Anglicisms in the texts and in which context they are inserted; additionally, I observe in which cases and for which topics the articles are provided in English. In both cases, the use of a concordancing programme gives evidence of the distribution of English in Italian texts diachronically. Some results show an increasing trend in the use of Anglicisms over time. Moreover, data can highlight to what extent and in which way English is seen to be used as a language to promote art, culture, music, food and touristic places in an Italian magazine for travellers on train. This may also open to further insights into the relationship between this particular Italian text-type in which English is frequently used and the potential identification of the reader of this on-board magazine.

References

Gotti, M. (2003). *Specialized discourse: Linguistic features and changing conventions.* Peter Lang.

Gotti, M. (2006). The language of tourism as specialized discourse. In O. Palusci & S. Francesconi (Eds.), *Translating tourism: Linguistic / cultural representations* (pp. 15-34). Università di Trento: Dipartimento di Studi Letterari, Linguistici e Filologici.

Maci, S. (2020). *English tourism discourse: Insights into the professional, promotional and digital language of tourism.* Hoepli.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford University Press.

Thurlow, C., & Jaworski, A. (2003). Communicating a global reach: Inflight magazines as a globalizing genre in tourism. *Journal of Sociolinguistics, 7*(4), 579-606.
Ferrovie dello Stato (FS) Group. Retrieved 31, January, from https://www.fsitaliane.it/content/fsitaliane/en/media/la-freccia-magazine.html

Ensayos clínicos y traducción automática neuronal: clasificación de errores según MQM-DQF

Alicia Picazo Izquierdo, Alicante University

Adelina Gómez González-Jover, Alicante University

Las aplicaciones informáticas han sacudido el mundo de la lingüística y de la traducción en las últimas décadas. Tanto así, el mercado de la traducción profesional se ha visto afectado por la inclusión de la traducción automática neuronal, basada en inteligencia artificial y en redes neuronales artificiales, que se puede considerar el paradigma actual (Forcada, 2017), y se estima que en los próximos años la demanda sea cada vez mayor (Mordor Intelligence, 2024; Acumen, 2023). Empresas tecnológicas que publican a ritmos vertiginosos, comercios electrónicos que desean llegar a una audiencia más amplia con los menores gastos posibles o incluso organismos internacionales que han desarrollado sus propios sistemas de TA a fin de mejorar su productividad son algunos ejemplos de la demanda actual. La cuestión principal radica en la calidad y la naturalidad de los textos traducidos. Diversos estudios tienen como objeto la evaluación de la calidad en la traducción automática. Se han desarrollado varias tipologías de errores (Nord, 1997; Baker, 1992; House, 2015; TAUS, 2017, entre otros) y métodos de análisis de corpus paralelos, y se ha tratado de estandarizar el concepto de calidad en la traducción (Hatim y Mason, 1988; Albir, 2017, entre otros). El punto de partida para obtener una traducción de calidad es la ausencia de errores.

Los objetivos de este estudio descriptivo son: a) comparar el corpus paralelo; b) detectar los errores generados por el motor de traducción automática; y, c) clasificarlos según la tipología de errores MQM-DQF. Para ello, se sigue una metodología analítica mediante corpus, pues se compila un conjunto de ocho textos clasificados con un alto nivel de especialización para la observación del comportamiento lingüístico. Consta de protocolos de ensayos clínicos en inglés con distinta temática y extensión, recuperados de la página oficial de la Biblioteca Nacional de Medicina de EE.UU. El lenguaje médico especializado se caracteriza por la multitud de términos médicos abstractos, nombres de medicamentos y de métodos, siglas o abreviaturas, y estructuras sintácticas propias del lenguaje técnico. Estas características hacen a este tipo textual proclive a la traducción automática, pues la creatividad, el cuidado del lenguaje y la riqueza lingüística tienen una importancia menor. Se traduce el corpus al español a través de la herramienta de traducción automática neuronal DeepL Pro. Después, se alinea el corpus paralelo con el software LF Aligner. Finalmente, se analizan manualmente los errores en los que incurre la TA. Los resultados del análisis arrojan más de doscientos errores de traducción en un total de treinta y cinco mil palabras aproximadamente que contiene el corpus. Estos errores se han clasificado según su nivel de importancia (*critical, major, minor, neutral)*. Los errores más frecuentes pertenecen al

estilo de redacción, y están causados por la adherencia al texto origen. También cabe destacar los errores terminológicos causados por variación conceptual o denominativa y los errores gramaticales relacionados con la flexión.

Referencias

Acumen. (2023). *Machine translation market Size - Global industry share, analysis, trends and forecast 2022-2030.* Acumen Research and Consulting.

Albir, A. H. (2017). *Traducción y traductología.* Cátedra.

Baker, M. (1992). *In other words.* Routledge.

Forcada, M. (2017). Making sense of neural machine translation. *Translation Spaces, 6*(2), 291-309.

House, J. (2015). *Translation quality assessment. Past and present.* Routledge.

Hatim y Mason. (1988). *A textbook of translation.* Longman.

Mordor Intellingence. (2024). *Machine translation market size & share analysis - Growth trends & forecasts (2024-2029).* Mordor Intelligence.

Nord, C. (1997). *Translation as a purposeful activity: Functionalist approaches explained.* St. Jerome.

TAUS. (2017). *Quality evaluation using an error typology approach.* TAUS BV.

Estrategia combinada para la definición del vocabulario arquitectónico contemporáneo francés: un enfoque basado en corpus y tecnología IA

Zaida Bartolomé-Díaz, University of Las Palmas de Gran Canaria

La arquitectura evoluciona constantemente, generando conceptos inéditos y por ende una nueva terminología.

En nuestra investigación nos hemos interesado por esta terminología moderna de la arquitectura y para poder estudiarla hemos generado un corpus específico creado gracias a Sketch Engine compuesto por una selección de textos contemporáneos de arquitectura en francés disponibles en Internet. Este conjunto de datos, con textos fechados en las últimas dos décadas, captura las tendencias y avances en el ámbito de la arquitectura actual y recoge el léxico reciente de la disciplina.

El hecho de emplear nuestro propio corpus está justificado por un lado por la dificultad de encontrar uno que se ajuste a nuestras necesidades específicas y, por otro lado, por la importancia, tal y como señalan Crosthwaite y Baisa (2023), de comprender con precisión el contexto de los textos. En efecto, en un entorno en el que la inteligencia artificial ya está presente de manera habitual resulta complejo obtener un nivel de detalle contextual preciso y real utilizando únicamente modelos de lenguaje basados en la generación automática de textos mediante procedimientos estadísticos.

Así, a partir de nuestro corpus y gracias a la función de extracción terminológica de Sketch Engine identificamos y recopilamos los términos más relevantes de la arquitectura contemporánea en francés.

Posteriormente, nos centramos en definir con precisión estos términos, puesto que muchos de ellos aún no están recogidos en ninguna obra lexicográfica especializada.

Para este proceso, nos valimos de ChatGPT, siguiendo las pautas sugeridas por Schryver (2023) y Barrett (2023)[1], y exploramos cómo al proporcionar una lista de términos clave, la IA puede generar automáticamente entradas lexicográficas básicas. Igualmente, mostramos cómo se puede iniciar la elaboración de listas de significados, ejemplos y palabras asociadas a estos términos.

Este proceso nos plantea algunos desafíos, como los señalados por Jakubíček y Rundell (2023), quienes apuntan a que una dificultad en las definiciones generadas por ChatGPT podría ser que este se inspira en diccionarios antiguos. Mostramos nuestros intentos para prevenir este problema mediante instrucciones más precisas que orienten al sistema y así evitar dichos enfoques.

Igualmente, debido a las limitaciones temporales de la base de datos de ChatGPT, la cual llega hasta septiembre de 2021, algunos de los términos más recientes no están incluidos en su conocimiento. Mostramos en estos casos cómo ChatGPT es capaz de intuir el significado de muchos de estos términos contemporáneos gracias a indicaciones concretas o a contexto real aportado manualmente.

Por último, en un esfuerzo por estructurar estas definiciones de manera formal y estándar, solicitamos a ChatGPT reorganizar la información lexicográfica utilizando el formato TEI-LMF (Text Encoding Initiative-Lexical Markup Framework). Este formato, basado en XML, proporciona pautas para describir léxico, como diccionarios y tesauros, de manera estructurada.

Con nuestro trabajo, presentamos así una metodología para generar una herramienta lexicográfica básica relativa a la arquitectura contemporánea en la que la IA, particularmente modelos de lenguaje avanzados como ChatGPT, puede complementar y enriquecer nuestro conocimiento, comprensión y representación de un campo terminológico para el que no existen recursos lexicográficos recientes. Tal y como sugiere Barrett (2023), podemos empezar a ver ChatGPT como una herramienta lexicográfica. No obstante, esta requiere aún aportes humanos al inicio, para alcanzar resultados óptimos y poder llegar al *prompt* adecuado mediante ensayo y error, y al final para pulir la salida de nuestra consulta.

Referencias

Barrett, G. (2023). Defin-O-Bots: Challenging A.I. to create usable dictionary content. *24th Biennial Conference of the Dictionary Society of North America, 31 May – 3 June 2023, CO, EE.UU.*

Crosthwaite, P. y Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics, 3*(3), 100066. DOI: https://DOI.org/10.1016/j.acorp.2023.100066

Jakubíček, M. y Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? *Proceedings of the eLex 2023 conference*, 518-533. https://elex.link/elex2023/wp-content/uploads/102.pdf

Schryver, G. M. de. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography, 36*. DOI: https://DOI.org/10.1093/ijl/ecad021

Estrategias de atenuación en terapias de pareja

Josefa Contreras-Fernández, Polytechnic University of València

Existen diversos tipos de terapias para tratar parejas en conflicto. Entre ellas está la *emotionally focused couple therapy* o EFT (Wiebe y Johnson, 2016; Power, 2020). En la EFT se destaca la importancia de las emociones y se menciona cómo han de proceder los terapeutas con el fin de que las parejas puedan tener nuevas experiencias emocionales y solucionar así sus dificultades. En ese sentido, es necesario que exista afiliación terapéutica con el propósito de fomentar no solo la participación en las sesiones de terapia y la confianza en el terapeuta, sino también la autoexploración y autoreflexión y, de este modo, poder llegar a una posible solución del conflicto.

Al visualizar un corpus de terapias, cedido por el Dr. Rodríguez González del proyecto EFFECTs, se aprecia el recurso de estrategias pragmáticas para cuidar las imágenes de las personas implicadas al gestionar la sensibilidad de los temas tratados y el conflicto entre la pareja. Entre estas estrategias, destaca el uso de atenuantes, que parece ser un instrumento tanto de reducción de la fricción social como de retórica para gestionar la relación entre el trabajo de la terapeuta y la pareja (Albelda, Briz, Cestero, Kotwica y Villaba, 2014; Albelda, 2016; Figueras, 2018; Autor, 2020). Asimismo, también los atenuantes cumplen una función afiliadora (Uclés 2020).

El objetivo de este trabajo es, por una parte, analizar qué estrategias particulares utilizan los terapeutas en sesiones de terapia de pareja y averiguar con qué función las utilizan, y, por otra parte, analizar las intervenciones de la pareja para determinar las estrategias que usan y su finalidad, es decir, de autoprotección de la imagen, prevención de una posible amenaza a la imagen del otro o reparación de un daño hecho a la imagen del interlocutor. Asimismo, se pretende observar si a lo largo de las sesiones se produce un cambio en las estrategias utilizadas por los pacientes o en la finalidad con la que la usan.

Para llevar a cabo esta investigación se han visionado veinte sesiones de terapias con diferentes terapeutas en las que se han seleccionado los fragmentos donde los terapeutas empleaban dichas estrategias, anotando las reacciones de los interlocutores. Se pudo observar, por una parte, que los terapeutas utilizan diferentes mecanismos lingüísticos, sobre todo, partículas fático-discursivas, como estrategias de atenuación. Asimismo, hacen uso de estrategias de afiliación con la intención de proteger las imágenes y empatizar con los pacientes con el fin de sintonizar con ellos, ayudarles en sus reflexiones y de intentar solucionar, de este modo, el conflicto. Por otra parte, el análisis mostró que la pareja utiliza la atenuación principalmente para proteger su propia imagen y crear una alianza con el terapeuta, con mayor frecuencia en las sesiones iniciales de la terapia.

References

Albelda, M. (2016). Sobre la incidencia de la imagen en la atenuación pragmática. *Revista Internacional de Lingüística Iberoamericana, 27,* 19-32.

Albelda, M., Briz A., Cestero A. M., Kotwica D. y Villalba C. (2014). Ficha metodológica para el análisis pragmático de la atenuación en corpus discursivos del español. ES.POR.ATENUACIÓN. *Oralia, 17,* 7-62.

Albelda, M. y Estellés, M. (2021). Mitigation revisited. An operative and integrated definition of the pragmatic concept, its strategic values, and its linguistic expression. *Journal of Pragmatics, 183,* 71-86.

Bravo, D. (1993). *La atenuación de las divergencias mediante la risa en negociaciones españolas y suecas.* Universidad de Estocolmo.

Figueras, C. (2018). Atenuación, género discursivo e imagen. *Spanish in Context, 15*(2), 258-280.

Figueras, C. (2021). Mitigation in discourse: Social, cognitive and affective motivations when exchanging advice. *Journal of Pragmatics, 173,* 119-133.

Hernández-Flores, N. (2008). Politeness and other types of facework: Communicative and social meaning in a television panel discussion. *Pragmatics, 18*(4), 577–603.

Lindström, A. y Sorjonen, M.-L. (2013). Affiliation in conversation. *The handbook of conversation analysis*, 350-369.

Peräkylä, A., Henttonen, P., Voutilainen, L., Kahri, M., Stevanovic, M., Sams, M. y Ravaja, N. (2015). Sharing the emotional load: Recipient affiliation calms down the storyteller. *Social Psychology Quarterly*, *78*(4), 301-323.

Power, A. (2020). My questions about emotionally focused couple therapy (EFT) and a few answers. *Attachment: New Directions in Psychotherapy and Relational Psychoanalysis*, *14*(1), 23-41.

Uclés, G. (2020). Las funciones interactivas del marcador español '¿no?' Las fronteras entre la atenuación y la protección de la imagen. *Signos*, *53*(104), 790-814.

Steensig, J. (2012). Conversation analysis and affiliation and alignment. En C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics. Oxford.* Blackwell Publishing. https://DOI.org/10.1002/9781405198431.wbeal0196

Wiebe, S. y Johnson, S. (2016). A review of the research in emotionally focused therapy for couples. *Family Process*, *55*(3), 390-407.

Evaluating Large Language Models (LLMs) in Annotating Specialized Opinion Texts: A GPT-Fueled Analysis of Functional Discourse Units (FDUs)

Javier Fernández-Cruz, University of Málaga

Carla Fernández-Melendres, University of Málaga

Irina Muñoz Toala, University of Málaga

This presentation aims to evaluate the performance of Large Language Models (LLMs) when classifying the textual structure of specialized opinion texts, particularly tourism reviews and broadsheet newspaper economic columns. Additionally, this presentation explores the particularities of the textual structure to improve ongoing Sentiment Analysis efforts framed in a larger project.

Several approaches to the annotation of functional units that compose a text have been proposed in literature. The Rhetorical Structure Theory (RST) by Mann & Thompson (1988) and its automatic parser by Marcu (1997, 1999) aim to discern the hierarchical structure of a text by identifying Elementary Discourse Units and their interrelated functions, contributing to overall text coherence. More recently, the Functional Discourse Units (FDU) proposed by Egbert et al. (2021) are utilized for segmenting conversational texts based on communicative purposes, particularly in identifying crucial discourse segments for accurate assignment of text polarity by discerning the opinion holder's expressions.

The proliferation of Language Model architectures underwent a significant transformation in 2023, exemplified by advancements like GPT-4 or Llama (e.g., Moreno-Ortiz, 2024). This expansion has significantly expanded the horizons of our text classification endeavors, yielding unprecedented efficacy and outcomes. Despite these advancements, the potential of these tools remains largely unexplored and unevaluated. Notably, the recent introduction of ChatGPT has garnered considerable attention within the Natural Language Processing (NLP) community, showcasing its ability to generate high-quality responses to human input and self-correct errors based on subsequent conversations. Moreover, the nuances and implications of zero-shot versus few-shot learning strategies in this context merit further investigation.

In this context, Chang et al. (2023) provide a comprehensive and ongoing survey on the evaluation of LLMs. This paper conducts the most extensive evaluation to date of ChatGPT's performance across 140 diverse NLP tasks, including question-answering, text summarization, and bias detection. Despite its versatility and impressive results in some benchmarks, ChatGPT still falls short of reliably solving many challenging tasks (Qin et al., 2023; Ortega-Marín et al., 2023). This acknowledgement highlights the need for a nuanced understanding of LLMs capabilities, reinforcing the importance of ongoing evaluations.

Our corpus consists of ~200 manually annotated opinion texts from specialized domains (currently, tourism and economy news). The annotation framework consists of five layers: (1) polarity, (2) discursive functions, (3) aspect, (4) entity and (5) opinion holder. We analyzed manually-annotated FDU patterns with LLM-generated patterns with zero-shot learning

strategies. Preliminary evaluation metrics demonstrated a high level of agreement. This research showcases the potential for automated discourse analysis tools to efficiently categorize and analyze FDUs in specialized texts, providing valuable insights for linguistic and discourse studies.

References

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., & Xie, X. (2023). A survey on evaluation of large language models. DOI: https://DOI.org/10.48550/arXiv.2307.03109

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. DOI: https://DOI.org/10.1515/text-2020-0053

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243-281. DOI: https://DOI.org/10.1515/text.1.1988.8.3.243

Marcu, D. (1997). The rhetorical parsing of unrestricted natural language texts. *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, 96–103. DOI: https://DOI.org/10.3115/976909.979630

Marcu, D. (1999). The automatic construction of large-scale corpora for summarization research. *Proceedings of the 22nd annual international ACM SIGIRconference on Research and development in information retrieval,* 137–144. DOI: https://DOI.org/10.1145/312624.312668

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

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a general-purpose natural language processing task solver? https://DOI.org/10.48550/arXiv.2302.06476

Examining the potential of AI in the annotation of corpus examples for language learning

Iztok Kosem, University of Ljubljana

Tanara Zingano Kuhn, University of Coimbra

Špela Arhar Holdt, University of Ljubljana

Kristina Koppel, Institute of the Estonian Language

Carole Tiberius, Dutch Language Institute

Rina Zviel Girshin, Ruppin Academic Center

In language learning authentic examples play a crucial role. However, identifying good examples for educational purposes in language corpora is not an easy task. Examples may contain a combination of inappropriate or offensive language, spelling and grammatical mistakes, and insufficient context to be informative enough. There are automatic methods available to address this time-consuming task, however, they have shortcomings and human verification of the results is often still required (XXX, 2021).

We have attempted to tackle this problem with a combination of crowdsourcing and gamification, by developing a game called XXX (Crowdsourcing for Language Learning). XXX is a multi-language digital game devised to crowdsource the labelling of examples with problematic content for pedagogical purposes. The game currently supports Dutch, Estonian, Slovenian, and Brazilian Portuguese. In this game, players identify problematic examples automatically extracted from existing corpora, categorize the problems, and point out the constituent part of the sentence that is problematic.

Game development involved creating manually annotated “seed” corpora of 10,000 examples per language (XXX, 2022). They serve as a gold standard for the game, i.e. to help identify the reliability of the crowdsourcers, and can also be used for machine training, language learning and other purposes. The corpora were prepared by expert linguists as part of the CLARIN Resource Families project and are available, together with annotation guidelines, on PORTULAN CLARIN.

As the arrival of large data models, in particular ChatGPT, has taken the world by storm, we wanted to explore whether the preparation of seed corpora could be performed with artificial intelligence tools. To this end, we devised a task which involved asking ChatGPT to identify any problematic contents in the examples provided and provide comments. The same English prompt, which was identified to work better than a prompt translated into each language, was used for all four languages in the game.

The initial tests indicate that overall, ChatGPT performs very well when it comes to identifying vulgar, offensive or sensitive language. There were occasional problems linked to translation errors, e.g., the word zamorec in Slovenian (‘nigger’) was not identified as problematic due to incorrect translation. ChatGPT exhibits some difficulties with identifying spelling and grammar problems; in some languages, it does not detect them, whereas in others it simply corrects them without even pointing out that the example is problematic. Lack of context appears to be a category where the ChatGPT output often differs the most from the annotators’ decisions. In several cases annotated as problematic by annotators, ChatGPT takes a more forgiving path, even pointing out the value of the examples for language learning.

The ongoing analysis of 40,000 examples using ChatGPT will provide comprehensive insights into its performance. Our presentation at the conference will detail the results, offering language-specific observations and highlighting any consistent patterns across languages. This research aims to inform the integration of AI tools in the compilation of corpora for language learning, emphasizing their strengths and areas for improvement.

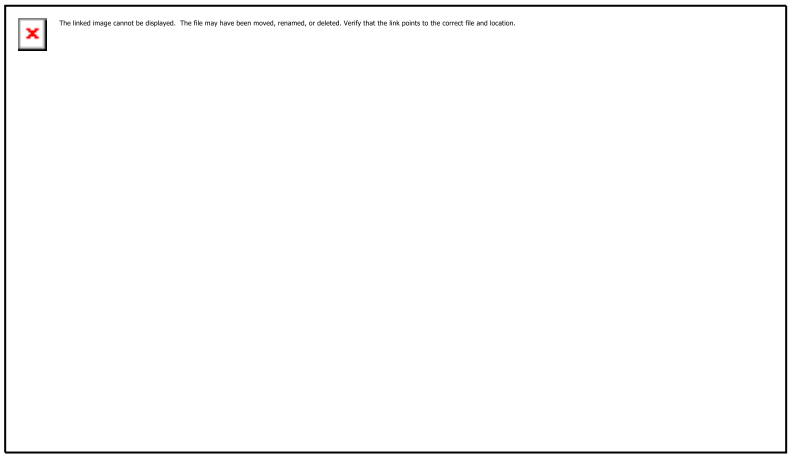
Experimenting a constructivist approach to annotation

Edmond Cane, Beijing International Studies University

Construction Grammar (CG) has been developed in the recent decades as a new research framework, hosting several approaches and models within, based on and sharing the main CG tenets, particularly the usage-based and emergentist concept regarding the development of linguistic structures, the constructional pairing of form/content as the main building block, containing all available information so packed, standing clearly as the other pole to all structuralist, rule-based approaches (Langacker 1987, Goldberg 1995, 2006, Croft & Cruse 2004, etc.). The existing corpora have their annotation models (POS tagging, syntactic parser) developed on rulebased frameworks and employ mainly probabilistic techniques. There is one corpus framework relying on the constructivist approaches - FCG (https://www.fcg-net.org/ see Steels 2011). The model presented here is a CG-based one, with features and concepts not very similar to the FCG.

The CG-based annotation coding here has been developed for Albanian, which is highly inflectional: the nouns for instance have a series of fixed order slot of affixes for number, (in)definiteness, case. In the CG-based FCG, these are processed as features to the noun and make the form side of the constructional pairing, based on Croft & Cruse (2004, p. 258). In this model here, the gender is inside the noun lemma, and the latter is assigned with gender without a marker, i.e., the speakers can recognize the m/f gender directly. The singular/plural distinction has been developed based on the concept of constructional pairing, i.e., only plural is marked, and this has been treated as a construction of its own, affixed to the noun, in the fixed order slot. Hence the relationship of definiteness, number and case to the noun lemma is a relation between constructions, not a relation between a lemma noun and its features within one construction (see Fig 1). This solution unblocks the problem of Albanian hybrid nouns for instance. There is a large number of masculine nouns, which display a feminine pattern and alignment in plural. This is due to the kind of plural marker (author 2021). Hence this model builds the alignment between the plural marker and the pronoun/adjective that complement the noun, which becomes a feminine alignment, apart from the noun lemma, with its own internal masculine gender. This solution makes the model straight, without patches or ambiguities.

Fig 1. The architecture of a complex NP showing the various patterns of relationships



The annotation code represents the speakers' knowledge and all the knowledge has been encoded to the constructions. The challenge is how the information is packaged. The relation of noun to the constructions for number, definiteness and case is governed by higher level schemas which validate the alternatives if there are two or more interpretations. The higher level schemas (or constructions) do the reading (recognize and annotate) even when there is a single interpretation. The validation is done only by the grammatical schemas, so the annotation code generates grammaticality (only grammatical strings), and the information recognizing and reading the text is provided from the available schemas directly – this principle underlies the FCG model too. At the level of NP, there are the bottom constructions as well as the alignments, which are fixed schemas. Both do the reading and the annotation of NPs.

There are three dimensions of processing and output. The first one recognizes/reads the bottom level constructions, which includes the noun or verb lemma only. The second governs the schema that completes the extended noun or verb, based on a template with slots (number, definiteness), and the third one reads and annotates the larger constructions at the level of AP, NP, VP (constituents).

It can be seen in Fig 1 that prepositions (and cases) operate outside the extended noun (2nd dimension), as they establish the link at NP level – so they are coded here.

This abstract presents only the difference in concept and technical patterns. The implications regarding the efficiency and use of corpora designed on this model could be vast. This model provides an automatic tagger without the long list of POS tags and it will not need the probabilistic trainer. It can achieve high accuracy directly. The code packages are open to adjustment along the explicit grammar knowledge, without employing the non-linguistic tools. In the presentation the audience will be shown packages of code (it is on GitHub), processing and output.

References

Croft, W., & Cruse, D. A. (2004). *Cognitive linguistics.* Cambridge University Press.
Garside R., Leech G., & McEnery A. (Eds.) (1997) *Corpus annotation.* Longman.
Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure.* University of Chicago Press.
Goldberg, A. E. (2006). *Constructions at work: The nature of generalizations in language*. Oxford University Press.
Langacker, R. W. (1987). Foundations of cognitive grammar: Volume II: Descriptive application. Leech, G., Garside, R., & Bryant, M. (1994, August). CLAWS4: The tagging of the British National Corpus. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics.*
Steels, L. (2011). Design patterns in fluid construction grammar. *Design Patterns in Fluid Construction Grammar*, 1-344.
POS Tagging for Albanian – models, designs:
Arkhangelskiy, T., Belyaev, O., & Vydrin, A. (2012). The creation of large-scale annotated corpora of minority languages using UniParser and EANC platform. In *Proceedings of COLING 2012* (pp. 83–92).
Hasanaj, B. (2012). *A part of speech tagging model for Albanian.* Lambert Academic Publishing.
Kabashi, B., & Proisl, T. (2018, May). Albanian POS tagging: Gold standard and evaluation. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018).*
Trommer, J., & Kallulli, D. (2004). A morphological tagger for standard Albanian. In *Proceedings of LREC* (pp. 1-8).

Existing Albanian corpora
Albanian National Corpus (http://albanian.web-corpora.net/)
SqTenTen on Sketch Engine https://www.sketchengine.eu/

Exploring Colloquialization in English as a Lingua Franca (ELF): Multivariate Analysis of Necessity Modals in Spoken and Written ELF

Chunyuan Nie, University of Eastern Finland

This study focuses on the change and variation in the use of necessity modals in spoken and written English use a lingua franca (ELF). The core modal must coexists with the semi-modals *have to, need to*, and *(have) got to* to express obligation and necessity and are interchangeable in certain contexts. The consensus from earlier studies reveals a broad

quantitative trend in both ENL (English as native language) and ESL (English as a second language) varieties in which *must* is declining in frequency, while the uses of *have to* and *need to* are increasing in expressions of obligation and necessity (Krug 2000; Collins 2009; Leech 2013). Moreover, a broadly observed pattern of frequency shift – colloquialization towards spoken conventions —is reflected in the increase of semi-modals and decrease of *must* (Leech 2003, p. 236). Empirical evidence of colloquialization in the use of necessity modals is presented in both ENL and ESL varieties with both descriptive and inferential statistics (Collins 2009; Collins & Yao 2012). Collins and Yao (2012) show that, overall, ESL varieties exhibit lower frequencies of semi-modals in spoken language and higher frequencies of corresponding modal verbs in written language. This leads to their conclusion that these varieties are less advanced than ENL varieties in this linguistic change. Nevertheless, Hansen (2018, p. 46) argues that the lower frequencies of semi-modals in spoken ESL varieties could stem from a preference for formal terms in some ESL varieties. In their respective multivariate analyses, Flach, Cappelle, & Hilpert (2023) investigate American English (AmE) and find a preference for *must* in more formal contexts and *have to* in informal ones. Similarly, Hansen (2018) observes this tendency in his study of Hong Kong English, Indian English, and Singapore English. However, the extent of this preference varies across these different varieties. ´

This study contributes to the limited existing comprehensive quantitative analysis of ELF, which transcends geographic borders and remains unindigenized in any region or country. Earlier empirical studies have provided empirical evidence with descriptive and inferential statistics, suggesting that ELF speakers not only respond to the linguistic change found in ENL and ESL varieties but also show divergence arise in the multilingual context of ELF (Deshors; Laitinen 2020; Laitinen &Lundberg). This study investigates two ELF corpora: The Vienna-Oxford International Corpus of English (VOICE.2013) and Written English as a Lingua Franca in Academic Settings (WrELFA 2015). Both corpora consist of approximately one million words, with VOICE representing spoken language and WrELFA representing written communication. The first part of the empirical analysis involves a comparative analysis of necessity modals' frequencies in spoken and written ELF corpora. The aim is to investigate whether the phenomenon of colloquialization is observable in ELF. This inquiry is especially pertinent in the context of ELF, where the use of English transcends geographical boundaries and involves speakers with diversified linguistic and cultural backgrounds. The second part of the empirical analysis delves into the alternation between *must* and *have to* within the ELF context, employing Bayesian regression analysis. Earlier multivariate analyses in ENL and ESL varieties have demonstrated the parallel influence of a set of linguistic and social factors on the choice between necessity modals (Tagliamonte 2004; Tagliamonte &D'Arcy 2007; Hansen 2018). This multivariate analysis aims to investigate the potential differences on the influence of these linguistic and social factors between the spoken and written language modes in the multilingual context of ELF.

References

Collins, P. (2009). Modals and quasi-modals in World Englishes. *World Englishes, 28*(3), 281-292.

Collins, P., & Yao, X. (2012). Modals and quasi-modals in New Englishes. In M. Hundt & U. Gut (Eds.), *Mapping unity and diversity world-wide. Corpus-based studies of new Englishes* (pp. 35-54). John Benjamins.
Deshors, S. C. (2020). English as a lingua franca: A random forests approach to particle placement in multi-speaker interactions. *International Journal of Applied Linguistics, 30*(2), 214-231.
Flach, S., Cappelle, B., & Hilpert, M. (2023). You must/have to choose: Experimenting with choices between near-synonymous modals. In S. M. Fitzmaurice & B. Kortmann (Eds.), *Models of modals: From pragmatics and corpus linguistics to machine learning* (pp. 149-176). De Gruyter.
Hansen, B. (2018). *Corpus linguistics and sociolinguistics: A study of variation and change in the modal systems of world Englishes.* Brill.
Krug, M. G. (2000). *Emerging English modals: A corpus-based study of grammaticalization.* De Gruyter.
Laitinen, M. (2020). Empirical perspectives on English as a lingua franca (ELF) grammar. *World Englishes, 39*(3), 427-442.
Laitinen, M., & Lundberg, J. (2020). ELF, language change and social networks: Evidence from real-time social media data. In A. Mauranen & S. Vetchinnikova (Eds.), *Language Change: The Impact of English as a Lingua Franca* (pp. 179-204). Cambridge University Press.
Leech, G. (2003). Modality on the move: The English modal auxiliaries 1961–1992. In R. Facchinetti, M. Krug & F. Palmer (Eds.), *Modality in Contemporary English* (pp. S. 223-240). De Gruyter.
Tagliamonte, S. A. (2004). Have to, gotta, must. In C. Mair & H. Lindquist (Eds.), *Corpus Approaches to Grammaticalization in English* (pp. 33-55). John Benjamins.
Tagliamonte, S. A., & D'Arcy, A. (2007). The modals of obligation/necessity in Canadian perspective. *English World-Wide, 28*(1), 47-87.
VOICE (2013). The Vienna-Oxford International Corpus of English (version 2.0 online). https://voice2.acdh.oeaw.ac.at/index.xql (last accessed 31 October 2023).
WrELFA (2015). The Corpus of Written English as a Lingua Franca in Academic Settings. http://www.helsinki.fi/elfa (last accessed January 30, 2024).

Fake news and propaganda in Russian news on social media

Elizaveta Kibisova, University of Oslo

Silje Susanne Alvestad, University of Oslo

Despite the increasing focus on fact-checking within the realm of social media, it remains a salient concern that such platforms continue to allow claims of questionable veracity to be spread unfiltered and read by thousands in a short time. Combined with people's resistance

to change their minds when proven wrong (Mosleh et al. 2021), fake news poses a threat to modern societies. On Runet (the Russian-speaking segment of the Internet), fake news is a particularly important subject, strongly intertwined with the political and social climate in Russia and neighbouring countries. Recently, the impact of fake news on the Russian-speaking audience became particularly visible during the COVID-19 pandemic and the invasion of Ukraine.

This study of propaganda techniques in Russian news reports on social media is conducted within the framework of a larger research project. In this broader project, we take a corpus linguistics approach to finding grammatical and stylistic features of the language of fake news. As part of the project, we have manually collected a corpus of Russian social media-based news articles of varying veracity. The levels of (non-)veracity, ranging from genuine to entirely fictitious, are established and labelled by independent professional fact-checkers; the procedures for their work and the results of their investigations are freely available online and based on open data. The social media platforms in question are the ones that are especially popular in Russia – namely, Telegram and VK, but also two of the most popular worldwide: Facebook and Twitter. The corpus includes over 350 news entries, readily available for linguistic research.

In this particular study, we follow Tandoc et al. (2018) in defining propaganda as a subcategory of fake news: based on facts, it also includes bias that promotes a particular side or perspective. The aims of our study are to investigate and describe the linguistic representation of manipulation techniques in news articles in Russian that can be categorised as propaganda. To achieve this goal, we adapt and apply Da San Martino *et al.*'s (2019) framework for propaganda detection and annotation in English to a subcorpus of our Russian social media-based news corpora. The subcorpus includes news entries labelled by fact-checkers as *'half-true'*, *'mostly true'*, or *'partially true'*. Unlike entirely fictitious *'false'* or *'fake'* entries, these categories of news articles tend to be mostly, but not entirely based on facts, thus providing a set of articles that align with our adopted definition of propaganda. By annotating instances of propaganda techniques used within each article, we detect the prevalent techniques, their possible combinations within one article, and the specific linguistic strategies that are used to convey these techniques. Preliminarily, analyses of a pilot sample show that the most frequent techniques in the subcorpus are attack on reputation (*Americans are a canonical society of consumers*), manipulative language, such as irony, and justifications by appealing to values (*fight against European values to protect traditions*). Our results in the form of quantitative and qualitative data can potentially aid manual and automated propaganda detection on social media.

References

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018) Defining "Fake News": A typology of scholarly definitions. *Digital Journalism, 6*(2), 137-153. DOI: https://doi.org//10.1080/21670811.2017.1360143

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news article. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint*

Conference on Natural Language Processing. 5640-5650. DOI: https://doi.org//10.18653/v1/D19-1565
Hindman, E. B. (2005). Jayson Blair, The New York Times, and paradigm repair. *Journal of Communication, 55*(2), 225–41. DOI: https://doi.org//10.1093/joc/55.2.225

From Ulyanov to Lenin: a corpus-based discourse analysis of Vladimir Lenin's works

Mikhail Mikhailov, Tampere University

Vladimir Lenin died 100 years ago, in 1924, and still remains one of the most controversial figures of the past century: some destroy his monuments while others believe him to be a prophet of the new era. In spite of numerous books devoted to this man, we know surprisingly little about him. Very few non-propagandist research of Lenin's language was performed (see Eihenbaum,1924; Tynjanov, 1924; Kruchenyh, 1928; Filin, 1974) and it seems that there is none or very few corpus-based research of Lenin's works, nor discourse analysis on such data.

The purpose of my research is to perform a corpus-based study of the evolution of the marxist discourse in Russia of 1880-1920-ies. The methodology used is corpus-based discourse analysis (Baker, 2006; Gillings et al., 2023). Two corpora were compiled for the purpose: Lencor, a corpus of Lenin's works (the Complete Works in 55 volumes, about 5.5 M running words) and Rudire, a corpus of Russian political discourse of the time of the Revolutions (books, articles, speeches, essays by most prominent figures in the Russian society of that time: statesmen, politicians, scholars, writers, educators etc, currently about 1.5 M running words). Lencor will be used as a research corpus and Rudire as a reference corpus. The texts were automatically parsed by Universal Dependency Grammar parser and uploaded to an instance of NoSketch Engine installed at the University server.

The research starts with keyword analysis. To obtain the most important words of Lenin's discourse the simple maths keyness score was used (Kilgarriff, 2009). This measure is built into the Sketch Engine software to obtain lists of keywords. My task was to find the high-frequency lexemes that occur in the research data significantly more frequently than in the reference data. For this reason the smoothing parameter N=1000 was used to filter out low-frequency words. The keyword search runs extremely well and produces reliably looking results. The central lexemes for the whole Lencor are *rabochij* 'worker', *partija* 'party', *bor'ba* 'fight N', *CK* 'Central Committee'. Lencor was split into 9 subcorpora by chronological periods, and the keywords for each period show the changes that occurred in the discourse. For example, the works of 1893-1901 are concentrated in economy and agriculture (*rabochij* 'worker', *proizvodstvo* 'production', *hozjajstvo* 'economy'), while the works of 1914-1917 are

all about the World War I and the political discussion around it (*vojna* 'war', *rabochij* 'worker', *imperializm* 'imperialism').

The next stage is the collocation analysis of these central lexemes. To give one short example, the lexeme *rabochij* 'worker' which is the most important in the Lenin's discourse, occurs in the corpus 26,121 times (4,605.90 per million). The collocation search finds the most important contexts for the word: *rabochij i krest'janin* 'worker and peasant', *soznatel'nyj rabochij* 'conscious worker', *naemnyj rabochij* 'hired worker', *rabochij deputat* 'worker deputy' etc. For the most interesting cases concordances are compilEd. The objective of research is to find how Lenin adapted to the political situation. The current findings demonstrate that Lenin's discourse has a tendency to simplification, it becomes more straightforward and more reader-orientEd. The most straightforward are the keywords of the last period (1921-23) with the following top five: *my* 'we', *kapitalizm* 'capitalism', *nado* 'must (be done)', *rabora* 'work, N', *ekonomicheskij* 'economical'.

References

Baker, P. (2006). *Using corpora in discourse analysis.* Continuum.

Eihenbaum, B. (1924). Osnovnye stilevye tendencii v rechi Lenina [In Russian: The main stylistic tendences in Lenin's speech]. *LEF, 1*(5), 57-71.

Filin, F. (1974). O slovare jazyka Lenina [In Russian: On a dictionary of Lenin's language]. *Voprosy jazykoznanija, 6*, 3-10.

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

Kilgarriff, A. (2009). Simple maths for keywords. In Mahlberg, M., González-Díaz, V. & Smith, C. (Eds.) *Proceedings of Corpus Linguistics Conference CL2009, University of Liverpool, UK.* https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf

Kruchenyh, A. (1928). *Priemy leninskoj rechi.* Moskva.

Tynjanov, J. (1924). Slovar' Lenina-polemista [In Russian: The vocabulary of Lenin as polemist], *LEF,1*(5), 81-111.

Gender, True Crime and Journalism: A Comparative Study of Media Portrayals in the Yorkshire Ripper Case (1975-1985)

Elena Castellano-Ortolà, University of València

This study conducts a comparative analysis of media coverage on the notorious 'Yorkshire Ripper' case (1975-1980) in both British and Spanish press. The portrayal of serial feminicides by Peter William Sutcliffe reveals the influence of social and literary conventions, particularly in the logics of law enforcement and the profit-driven press. Challenging the

treatment of the case has led to changes in British police protocol, but the paper suggests further contributions can be made by addressing discourse-generated biases in criminal profiling.

Utilizing Tabbert's Critical Stylistics, the research examines a bilingual corpus of 432 British articles from the Newsgen database and 141 Spanish articles from Hemeroteca Digital between 1975 and 1985. The focus is on understanding how press-driven tropes surrounding perpetrator and victim profiles cross cultural boundaries, especially in gender-based crime reporting. The analysis also considers the transnational nature of journalistic genres, going beyond traditional categories like chronicles, to explore how they either reinforce or challenge cultural ideologies, particularly in the context of gender-inflicted notions of criminal responsibility and victimhood.

References

Baldwin, C. (2001). The development of rhetorical privilege in the news reporting of violent crime. *Race, Gender & Class*, 8-19.

Barak, G (2019). Mass-mediated regimes of truth: Race, gender, and class in crime “news” thematics 1. *Race, Gender, and Class in Criminology the Intersections*, 105-123.

Gregoriou, C. (Ed.). (2012). *Constructing crime: Discourse and cultural representations of crime and 'deviance'.* Springer.

Lazar, M. (Ed.). (2005). *Feminist critical discourse analysis: Gender, power and ideology in discourse.* Springer.

Mandolini, N. (2021). *Representations of lethal gender-based violence in Italy between journalism and literature: Femminicidio narratives*. Routledge.

Mills, S. (2002). *Feminist stylistics.* Routledge.

Mislán, C. (2018). 26. Journalism, Gender, and Race. In Vos T. P. (Ed.), *Journalism* (pp. 511-530). De Gruyter Mouton.

Montoro, R. (2017). Feminist stylistics. In Burke, M. (Ed.), *The Routledge handbook of stylistics* (pp. 364-379). Routledge.

Monzó, E (2002). La traducción jurídica a través de los géneros: El transgénero y la socialización del traductor en los procesos de enseñanza/aprendizaje. *Discursos: Estudos de Tradução,* 21-36.

Smith, J. (2013). *Misogynies: Reflections on myths and malice.* Saqi.

Stuart A., Branston G., & Carter C. (1998). *News, gender, and power.* Psychology Press.

Tabbert, U. (2015). *Crime and corpus: The linguistic representation of crime in the press* (Vol. 20). John Benjamins Publishing Company.

Hacia la construcción de un corpus de informes médicos en español: superando barreras lingüísticas en la salud de la mujer

Ovidia Martínez Sánchez, University of Alicante

En el contexto actual, la comunicación eficaz en el ámbito médico es un elemento esencial para empoderar a los pacientes (Toledo Chavarri, *et al.*, 2016) y garantizar una toma de

decisiones informada sobre su salud. Sin embargo, en el plano textual, los informes médicos a menudo se redactan en un lenguaje técnico y complejo, lo que obstaculiza su comprensión por parte del público general (López Fuentes A., 2022). Esta tesis doctoral se centra en superar estas barreras lingüísticas mediante la creación de un corpus de informes médicos en español, específicamente orientado hacia la salud de la mujer, necesario para aplicar análisis basados en la Lingüística de Corpus (Parodi, G., 2022) y el Procesamiento de Lenguaje Natural (PLN) (Jurafsky, D. y Martin, H. J., 2023) en el ámbito médico (Quevedo Marcos, 2020). El propósito de esta tesis es diseñar y anotar un corpus o *dataset* de informes médicos en español para mejorar la comprensión de estos documentos por parte de las pacientes, mediante el uso de técnicas de PLN, específicamente la Simplificación Textual (ST) o *Text Simplification* (TS) (Espinosa-Zaragoza *et al.*, 2023), el lenguaje claro, la estandarización terminológica y evaluación de adaptaciones lingüísticas.

El estudio se sustenta en antecedentes que resaltan la falta de corpus anotados en español y la complejidad en la comprensión de informes médicos (López Fuentes A., 2022), dado que en su mayoría se encuentran en inglés (Joseph, et al., 2023). A diferencia de investigaciones anteriores, este proyecto recopila grandes conjuntos de datos, considera la perspectiva de género, y se alinea con los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas. Para alcanzar el objetivo planteado, presentamos una propuesta metodológica sobre la conformación y delimitación de un corpus médico representativo, tanto en términos cualitativos y cuantitativos, que refleje este conocimiento especializado. Se detallan los pasos del proceso de diseño, recopilación de textos y creación de una base de metadatos con información relevante, como tipo de informe, especialidad médica, cantidad de palabras, entre otros aspectos.

Desde una fase explotaría, se identifican y recopilan informes médicos procedentes de distintas zonas geográficas de España, tales como Andalucía, Comunidad Valenciana y Castilla-La Mancha, respetando las regulaciones de protección de datos (RGPD). Los criterios de selección incluyen que los informes estén redactados en español, sean de mujeres entre 18 y 64 años, y provengan de especialidades como Ginecología, Obstetricia o Atención Primaria. Además, se abordan los criterios para el tratamiento y conversión de los textos en formatos procesables, incluyendo la digitalización de informes, la limpieza textual, anonimización y almacenamiento. La recopilación y tratamiento de estos textos que conforman el corpus constituyen un proceso complejo y meticuloso, previo a su anotación y explotación.

Referencias

Espinosa-Zaragoza, I., Abreu-Salas, J., Lloret, E., Moreda, P. y Palomar, M. (2023). A review of research-based automatic text simplification tools. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria* (pp. 321–330).

Joseph, S. A., Kazanas, K., Reina, K., Ramanathan, V. J., Xu, W., Wallace, B. Li, J. J. (2023). *Multilingual Simplification of Medical Texts.* ArXiv, abs/2305.12532.

Jurafsky, D. y Martin, H. J. (2023). *Speech and language processing* (Third Edition draft) https://web.stanford.edu/~jurafsky/slp3/

López Fuentes, A. C. (2022). *Adecuación automática de términos biomédicos para personas no expertas: El caso de los informes médicos* [Tesis Doctoral: Universitat Pompeu Fabra]. En TDX (Tesis Doctorals en Xarxa). https://www.tdx.cat/handle/10803/674580
Parodi, G., Cantos-Gómez P., Howe C., Lacorte M., Muñoz-Basol J. y Muñoz-Basol J. (2022). *Lingüística de corpus en español*. Routledge, Routledge Handbooks Online.
Naciones Unidas (s.f.) Objetivos de Desarrollo (ODS). https://www.un.org/sustainabledevelopment/es/objetivos de desarrollo sostenible/
Quevedo, M. B. (2020). *Análisis de las herramientas de procesamiento de lenguaje natural para estructurar textos médicos.* [Trabajo fin de máster: Universidad de Navarra]. https://hdl.handle.net/10171/60003
Toledo Chavarri, A. et al. (2016). El papel de la documentación escita en el empoderamiento en salud: un estudio cualitativo. *Panacea*, *17*(44), 115 122. http://www.tremedica.org/panacea/IndiceGeneral/n44_tribunaAToledoChavarriEtAl.pdf

Hedging in academic writing. A corpus-based analysis of Clinical and Experimental Medicine MA and PhD theses in the MoreThesis Corpus

Marina Bondi, Univeristy of Modena and Reggio Emilia

Silvia Cavalieri, Univeristy of Modena and Reggio Emilia

Academic texts have come to be recognized as socially produced rhetorical artifacts that attempt to negotiate claims and persuade the reader, rather than as objective explanations of factual information. As a consequence, they are rich in hedged propositions (Musa 2014, p. 1) that enable writers to convey their uncertainty about the veracity of their claims by expressing different degrees of confidence. Pragmatically, hedges also show respect for readers by indicating that the claim is negotiable (Hyland 1994, p. 239), allowing authors to open a discursive space and mark statements as provisional, while conveying respect for colleagues' views (Hyland & Jiang, 2016). Although the phenomenon of hedging in academic writing has been thoroughly studied in various genres (e.g., abstracts, RAs, textbooks) (see Varttala, 2001; Hyland, 1998; Salager-Meyer, 1997), its use in MA and PhD theses has still been little considered by the literature. Moreover, very little attention has been paid to the evolution of employing hedging devices by writers with different degrees of expertise in their academic careers, even though it can be an important indication not only of increased writing proficiency but also of disciplinary enculturation (Abdollahzadeh, 2019; Wu & Paltridge, 2021).

Starting from these premises, the present study aims to investigate the role of hedging in academic texts dealing with MA and PhD theses in Clinical and Experimental Medicine written by Italian students in English. More specifically, the analysis adopts a corpus-based approach to focus on the quantitative and qualitative "developmental" changes in the use of hedging strategies in the two different stages of academic written production.

Data for the investigation are two sub-corpora of the MoreThesis Corpus currently under development at the Dep. of Studi Linguistici e Culturali of the University of Modena and Reggio Emilia and consist of a) the English MA theses (MA_MoreThesis) and b) the PhD

dissertations (PhD_MoreThesis) submitted by Clinical and Experimental Medicine students between 2014 and 2020. To have a complete overview, we will also compare our dataset to a reference corpus of scientific research papers freely available on SketchEngine, namely the Corpus of Elsevier Open Access Journals (collected from 2014 to 2020), choosing the sub-section 'medicine'. As for the methodology, we adopted a mixed-method approach, and we relied on the definition of hedges in Hyland's metadiscourse model (1998, 2018; Hyland & Jiang 2016; Hyland & Zou, 2021) to identify potential hedging items to search in our corpora as well as in the reference one. Keeping in mind the ongoing debate on the definition of modality and evidentiality (e.g. Alonso Almeida 2015a and b), we look at different functions of the wide range of items identifiable as hedges. We also look at how the different functions are deployed in different sections of the text.

Preliminary results show that, compared with their MA dissertations, students at the PhD stage demonstrated modest progress in using hedges, showing an increased ability to mitigate their positions. Moreover, at the doctoral level, students seem to convey their opinions about propositions by selecting from a larger range of linguistic hedging resources that are more in line with what is observable in the RAs reference corpus.

References

Alonso-Almeida, F. (2015a). On the mitigating function of modality and evidentiality. Evidence from English and Spanish medical research papers. *Intercultural Pragmatics, 12*(1), 33-57.

Alonso-Almeida, F. (2015b). The functions of *seem* and *parecer* in early medical writing. *Discourse Studies, 17*(2), 121-140.

Abdollahzadeh, E. (2019). A cross-cultural study of hedging in discussion sections by junior and senior academic writers. *Ibérica*, (38), 177-202.

Musa, A. (2014). Hedging in academic writing: A pragmatic analysis of English and Chemistry masters' theses in a Ghanaian university. *English for Specific Purposes, 42*(15), 1-26.

Hyland, K. (2018). *Metadiscourse: Exploring interaction in writing.* Bloomsbury Publishing.

Hyland, K. (1998). Boosting, hedging and the negotiation of academic knowledge. *Text & Talk, 18*(3), 349-382.

Hyland, K. (1994). Hedging in academic writing and EAF textbooks. English for specific purposes, *13*(3), 239-256.

Hyland K., & Jiang F. (2016). Change of attitude? A diachronic study of stance. *Written Communication, 33*(3), 251-274

Hyland, K., & Zou, H. J. (2021). "I believe the findings are fascinating": Stance in three-minute theses. *Journal of English for Academic Purposes, 50*, 100973.

Salager-Meyer, F. (1997). I think that perhaps you should: A study of hedges in written scientific discourse. In Miller, T. (Ed.), *Functional Approaches to Written Text: Classroom Applications* (pp 105-118). Washington, D.C., USA: English Language Programs-United States Information Agency.

Wu, B., & Paltridge, B. (2021). Stance expressions in academic writing: A corpus-based comparison of Chinese students' MA dissertations and PhD theses. *Lingua, 253*, 103071.

Varttala, T. (2001). *Hedging in scientifically oriented discourse. Exploring variation according to discipline and intended audience.* Tampere University Press

Inocuidad alimentaria y traducción: evidencias terminológicas a partir de un corpus paralelo inglés-español

Jorge Leiva Rojo, University of Málaga

Este trabajo tiene como objeto de estudio la traducción de alertas alimentarias, un tipo de texto que en los últimos años ha adquirido relevancia en ámbitos no especializados y cuyo último caso más notorio en España fue la alerta sanitaria relacionada con el brote de botulismo originado por el consumo de tortillas de patatas en julio de 2023.

En el ámbito de la traducción, aunque son numerosos los estudios que han ido surgiendo que abordan la traducción en la industria alimentaria —Rodríguez Rodríguez (2013), Chiaro y Rossato (2015), Moreno Paz y Rodríguez Tapia (2015), y Vidal Claramonte & Faber (2017)—, no se han localizado hasta el presente estudios que se centren en la traducción de este tipo de texto en concreto. Esta ausencia resulta llamativa no solo por la importancia de la información que proporcionan las alertas alimentarias, sino también por lo amplísimo de su público potencial: según los Centers for Disease Control and Prevention de los Estados Unidos (2020), cada año «an estimated 1 in 6 Americans (or 48 million people) get sick, 128,000 are hospitalized, and 3,000 die of foodborne diseases», por lo que la traducción de estas alertas es altamente relevante (Hallman y Cuite, 2009, p. 10).

Para este trabajo se recurrirá a un corpus de textos, paralelo y alineado en el nivel de la oración —originales escritos en inglés y traducidos al español—, con alertas alimentarias del *Food Safety and Inspection Service* de los Estados Unidos (FSIS) emitidas entre los años 2006 y 2020. El corpus, para cuya compilación se siguieron las recomendaciones sobre textos de la industria alimentaria de Li (2019), cuenta con 1279 bitextos y más de 350 000 y 475 000 palabras para los subcorpus inglés y español, respectivamente. Adicionalmente, para esta ocasión se ampliará el corpus con las 78 nuevas alertas alimentarias que el FSIS ha emitido y traducido al español entre 2020 y diciembre de 2023, con lo que el volumen final será mayor del que se informa aquí. Igualmente, se recurrirá a un corpus comparable, monolingüe (solo en lengua española), conformado por alertas alimentarias y diversos textos (informativos, normativos y divulgativos) sobre inocuidad alimentaria de organismos oficiales de México, Colombia y España, con objeto de poder validar las opciones de traducción encontradas en las alertas alimentarias traducidas al español en los Estados Unidos.

Con ayuda del estudio del corpus, se caracterizará cómo es la traducción de las alertas alimentarias traducidas al español por el FSIS, en primer lugar haciendo una descripción de algunos elementos originales (presencia de calcos, traducción de topónimos, tratamiento de unidades de medida, entre otros), para posteriormente hacer un estudio más detallados de la traducción de algunos elementos terminológicos, tales como la traducción de *soy* (de gran relevancia por ser un alérgeno y, por lo tanto, con una nutrida presencia en el corpus, con 226 casos en el corpus actual). También se estudiarán elementos terminológicos menos

frecuentes. Así sucede con el fruto seco *peanut*, que cuenta con 47 apariciones en el corpus, pero en este caso lo que resulta más problemático es dar con una traducción que sea entendible por el mayor número de hispanohablantes, ya que cada una de las tres opciones que se localizan, *maní, cacahuate* y *cacahuete* es la opción preferida, de acuerdo con el corpus comparable, en, respectivamente, Colombia, México y España.

Referencias

Centers for Disease Control and Prevention. (2020). *Key facts about food poisoning.* https://www.cdc.gov/foodsafety/food-poisoning.html

Chiaro, D. y Rossato, L. (2015). Introduction: Food and translation, translation and food. *The Translator, 21*(3), 237–243. DOI: https://DOI.org/10.1080/13556509.2015.1110934

Hallman, W. K. y Cuite, C. L. (2009). *Food recalls and the American public: Improving communications.* DOI: https://DOI.org/10.7282/T3639RQX

Li, S. (2019). A corpus-based multimodal approach to the translation of restaurant menus. *Perspectives, 27*(1), 1–19. DOI: https://DOI.org/10.1080/0907676x.2018.1483408

Moreno Paz, M. del C. y Rodríguez Tapia, S. (2015). La situación de la traducción agroalimentaria en la investigación y la formación en España. *Skopos. Revista Internacional de Traducción e Interpretación, 6,* 135–154. https://www.uco.es/ucopress/ojs/index.php/skopos/article/view/5658

Rodríguez Rodríguez, F. (2013). La traducción en el sector agroalimentario: Una simbiosis en auge. *Skopos. Revista Internacional de Traducción e Interpretación, 2*, 155–172. https://www.uco.es/ucopress/ojs/index.php/skopos/article/view/4423/4189

Vidal Claramonte, M. Á. y Faber, P. (2017). Translation and food: The case of mestizo writers. *Journal of Multicultural Discourses, 12*(3), 189–204. https://DOI.org/10.1080/17447143.2017.1339352

La anotación morfosintáctica de los corpus desde la perspectiva del usuario

Eva María Domínguez Noya, Galician Language Institute/ Ramón Piñeiro Center for Humanities Research

María Paula Santalla del Río, University of Santiago de Compostela

Siguiendo el modelo de De Benito (2019), que adopta una perspectiva general, y de otros trabajos que adoptan una más restringida (Hidalgo-Ternero, 2021; René, 2022), en esta contribución se lleva a cabo una revisión desde el punto de vista del usuario de lo que en concreto tiene que ver con la anotación morfosintáctica como elemento de acceso a los datos en corpus generales o específicos, principalmente, de español y gallego, pero también recurriendo a la comparación con corpus en otras lenguas cuando ello pueda ser

de utilidad. La evaluación de cómo la anotación morfosintáctica rentabiliza y facilita la explotación de un corpus por parte del usuario se basa en esta propuesta en la observación de los siguientes parámetros: a) permite o no la interfaz de consulta del corpus distinguir entre búsquedas de formas ortográficas, elementos gramaticales (*escribiéndome* es una forma ortográfica pero dos elementos gramaticales *escribiéndo/escribiendo* y *me*), lemas e hiperlemas (*cinc* y *zinc* son dos lemas, pero un hiperlema); b) si se salva, y cómo desde el punto de vista del usuario, la diferencia entre, por ejemplo, *escribiéndo/escribiendo*; c) qué concepto de lema se maneja (implica rigurosamente la clase de palabra o se hacen al respecto ciertas concesiones, acercándolo al de entrada lexicográfica); d) puede, y con qué visualizaciones, el usuario acceder al conjunto de etiquetas aplicadas al corpus; e) qué especificaciones morfosintácticas puede el usuario utilizar en sus búsquedas (lo que depende de la riqueza del sistema de etiquetas aplicado al corpus y de si se permite al usuario rentabilizarlo exhaustivamente o no); f) ¿tiene el usuario que conocer las etiquetas y los símbolos que las integran o puede manejarlas de manera transparente y amigable por medio de alguna facilidad integrada en la interfaz?; g) qué nivel de esfuerzo por parte del usuario exige, si la precisa, la discriminación de valores de categorías gramaticales (para adjetivos, por ejemplo, como *inteligente*), o de lemas (*música*, posiblemente relacionado con *músico/a/os/as* o con *música/s*) y h) ¿qué nivel de información se proporciona al usuario sobre cómo se han aplicado al corpus las distinciones morfosintácticas recogidas por la anotación? De todos estos parámetros se dará cuenta en esta presentación, entre otros, en los corpus CORPES XXI (Rojo 2010), CDH, ESLORA (Barcala *et al*., 2018; *Domínguez et al*., 2020b; Vázquez *et al.,* 2020), NOW (Davis, 2002; Rojo, 2010), CORGA (Domínguez *et al.,* 2020a), CAES (Palacios *et al.,* 2019), PRESEEA (Moreno, 2005) y Sketch Engine (Kilgarrif *et al*., 2004, Kilgarrif *et al*., 2014). Por ejemplo, con respecto al corpus de propósito general para el gallego CORGA se explicará a) que su interfaz de consulta distingue entre búsquedas de formas ortográficas, elementos gramaticales, lemas e hiperlemas; b) que los elementos gramaticales implicados en formas amalgamadas se reconstruyen (*escribindo* y *me*); c) que el lema se asocia siempre con la clase de palabra; d) que se facilita el etiquetario de modo condensado y a través de ejemplos en contexto; e) que posee un rico sistema de etiquetas que el usuario puede rentabilizar y f) manejar sin necesidad de conocerlas g) o en el que puede discriminar los valores de categorías gramaticales en algunos casos, aunque en otros la presencia de hipervalores obligará a una desambiguación manual; y h) que ofrece documentación acerca de cómo se ha anotado el corpus y de la metodología utilizada.

Corpus

CAES: Corpus de aprendices de español, https://galvan.usc.es/caes
CdE (NOW): Corpus del Español, https://www.corpusdelespanol.org/
CDH: Corpus del diccionario histórico de la lengua española, https://apps.rae.es/CNDHE/view/inicioExterno.view
CORGA: Corpus de Referencia do Galego Actual, https://corpus.cirp.gal/corga/
CORPES: Corpus del Español del Siglo XXI, https://www.rae.es/corpes/
ESLORA: Corpus para el estudio del español oral, https://eslora.usc.es/
PRESEEA: Corpus del Proyecto para el estudio sociolingüístico del español de España y de América, https://preseea.uah.es/corpus-preseea
Sketch Engine: https://www.sketchengine.eu/

Referencias

Barcala, M., Domínguez E., Fernández A., Rivas R., Santalla M. P., Vázquez V. y Villapol R. (2018). El corpus ESLORA de español oral: diseño, desarrollo y explotación. *CHIMERA: Romance Corpora and Linguistic Studies,* 5(2), 217-237. DOI: http://dx.DOI.org/10.15366/chimera2018.5.2.003

Civit Torruella, M. (2003). Criterios de etiquetación y desambiguación morfosintáctica de corpus en español. *Procesamiento del Lenguaje Natural*, Monografía 3. http://www.sepln.org/monografiasSEPLN/monografiaCivit.pdf

Davis, M. (2002). Un corpus anotado de 100.000.000 palabras del español histórico y moderno. *Procesamiento del lenguaje natural*, *29*, 21-27.

De Benito, C. (2019). Los corpus del español desde la perspectiva del usuario lingüista. *Scriptum digital, 8,* 1-21. https://www.raco.cat/index.php/scriptumdigital/article/view/361052/455867

Domínguez Noya, E. M., López Martínez M. S. y Barcala Rodríguez F. M. (2020a). Corpus de Referencia do Galego Actual (CORGA): Composición, codificación, etiquetaxe e explotación. En M. Blanco, H. Olbertz y V. Vázquez Rozas (Eds.), *Corpus y construcciones. Perspectivas hispánicas* (pp. 179-218). Servizo de Publicacións e Intercambio Científico da USC. DOI: https://dx.DOI.org/10.15304/9788417595876

Domínguez Noya, E. M., Rivas Cabanelas R., Santalla del Río M. P. y Villapol Baltar R. (2020b). Problemas afrontados en la etiquetación morfosintáctica del corpus ESLORA. En M. Blanco, H. Olbertz y V.Vázquez Rozas, *Corpus y construcciones. Perspectivas hispánicas* (pp. 243-271). Servizo de Publicacións e Intercambio Científico da USC. DOI: https://dx.DOI.org/10.15304/9788417595876

Hidalgo-Ternero, C. M. y Corpas Pastor G. (2021). La variación fraseológica: análisis del rendimiento de los corpus monolingües como recursos de traducción. *Étude Romanes de Brno, 42.* DOI: https://DOI.org/10.5817/ERB2021-1-1

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

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

Moreno Fernández, F. (2005), Corpus para el estudio del español en su variación geográfica y social. El corpus "PRESEEA". *Oralia, 8,* 123-139.

Palacios Martínez, I., Barcala Rodríguez F. M. y Rojo G. (2019). El corpus de aprendices de español (CAES) y sus aplicaciones para la enseñanza/aprendizaje del español como lengua extranjera. En M. Blanco, H. Olbertz y V. Vázquez Rozas (Eds.), *Corpus y construcciones: perspectivas hispánicas* (pp. 273-301). Servizo de Publicacións e Intercambio Científico da USC. https://dx.DOI.org/10.15304/9788417595876

Venegas R., Viviana Bosio I. y Cerda-Canales C. (2022). Los corpus sincrónicos del español. Descripción y potencialidades para la investigación teórica y aplicada de la lengua. *Journal of Spanish Language Teaching, 9*(2), 116-133. DOI: https://DOI.org/10.1080/23247797.2022.2157080

Rojo, G. (2010). Sobre codificación y explotación de corpus textuales: Otra comparación del corpus del español con el CORDE y el CREA. *Lingüística, 24,* 11-50.

Vázquez Rozas, V., Barcala M., Domínguez Noya E., Fernández Sanmartín A., Rojo G. y Santalla M. P. (2020). Codificación y anotación del habla en un contexto bilingüe: El corpus ESLORA de español de Galicia. En Á. J. Gallego y F. R. Urgell (Eds.), *Dialectología digital del español* (pp. 191-226). Universidade de Santiago de Compostela: Servizo de Publicacións e Intercambio Científico da USC.

Andrés Ortega Garrido, University of Bergamo Studies

La crítica divide la obra literaria de Ramón María del Valle-Inclán (1866-1936) en tres etapas creativas, en correspondencia con su evolución estética. De los inicios modernistas se pasa a una etapa de transición que desemboca en la estética del esperpento, original y propia, señal reconocible de un estilo personal e inimitable que, además, no tuvo epígonos. En efecto, la evolución de su obra coincide con el desarrollo de su estilo, tal como ha señalado la crítica (Montolío Durán, 1992a, 1992b; Abad Nebot, Peces Gómez, 1995), de modo que el papel desempeñado por la propia lengua y el uso que se hace de ella adquiere una importancia difícil de observar en otros autores españoles contemporáneos.

En este contexto, el estudio de la lengua de Valle-Inclán a través de la metodología de la lingüística de corpus se revela altamente significativo para conocer más en detalle la ingeniería verbal que caracteriza la prosa del autor gallego. En el presente trabajo nos concentraremos en la obra narrativa de la tercera etapa, que incluye la novela *Tirano Banderas* (1926) y las tres que componen *El ruedo ibérico*, es decir, *La corte de los milagros* (1927), *Viva mi dueño* (1928) y *Baza de espadas* (1932), así como algunas narraciones breves que constituyen fragmentos de futuros tomos de esta serie narrativa que el autor no llegó a concluir, como *Fin de un revolucionario* (1928) y *El trueno dorado* (1936). El objetivo de investigación consiste en el análisis del léxico empleado en la descripción de los personajes, quienes en la prosa del esperpento sufren un tratamiento que los acerca a lo caricaturesco, con rasgos incluso de animalización. Todo ello se realiza en buena medida a través de la selección léxica, analizable desde el punto de vista de las frecuencias de uso en torno a determinadas categorías gramaticales como sustantivos, adjetivos y verbos. En este sentido, Valle-Inclán prescinde de modo sistemático de los verbos *dicendi* habituales, a favor de otros más cercanos al ámbito animal o al de la hipérbole (en vez de “decir”, los personajes pueden “cacarear” o “gemir”, por ejemplo).

Nuestro marco de referencia teórico y metodológico corresponde a la lingüística de corpus, concretamente a la estilística de corpus (Semino & Short 2004; Stubbs, 2005; O’Halloran, 2007; Mahlberg 2013, 2014, 2016), con notables resultados también en el ámbito de la lengua española (Piccioni, 2015; Nieto Caballero, 2018; Nieto Caballero, Ruano San Segundo, 2020; Chierichetti, 2022). Nuestro corpus de estudio se compone de los textos digitalizados de las obras referidas, en las últimas versiones publicadas en vida de Valle-Inclán, y para el análisis empleamos la herramienta informática Sketch Engine. Mediante el estudio de n-gramas, palabras clave y paquetes léxicos (Scott, 1997; Biber, 2005; Biber, Conrad y Cortes, 2004), podemos observar con precisión algunas de las peculiaridades de la prosa valleinclaniana de esta última época, como la adjetivación recurrente y el empleo sistemático de una serie de verbos concretos. Además, proponemos una comparativa entre las modalidades de descripción del personaje en esta etapa creativa del autor y en su etapa modernista, haciendo patente la radical diferencia en cuanto a selección léxica se refiere.

Referencias

Abad Nebot, F. y Peces Gómez, M. L. (1995). La lengua literaria y el pensamiento lingüístico de Valle-Inclán: estado de la cuestión. En M. Aznar Soler, J. Rodríguez Rodríguez (Eds.), *Valle-Inclán y su obra. Actas del Primer Congreso Internacional sobre Valle-Inclán* (Bellaterra, del 16 al 20 de noviembre de 1992), Cop d'idees. Taller d'investigacions valleinclanianes, 79-86.
Baker, P. (2004). Querying keywords: Questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics, 32*(4), 346-359.
Baker, P. (2018). Keywords: Signposts to objectivity? En A. Čermáková y M. Mahlberg (Eds.), *The Corpus Linguistics Discourse: In honour of Wolfgang Teubert* (pp. 77-94) *(Studies in Corpus Linguistics*; vol. 87), John Benjamins Publishing Company.
Biber, D. (2005). Paquetes léxicos en textos de estudio universitario: Variación entre disciplinas académicas. *Revista Signos*, *38* (57), 19-29
Biber, D., Conrad, S. y Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405
Chierichetti, L. (2022). *"Caminando con la niña que fui". Algunas calas en la obra de Elvira Lindo desde la óptica de la estilística de corpus*. Comares.
Mahlberg, M. (2013). *Corpus stylistics and Dicken's fiction*. Routledge.
Mahlberg, M. (2014). Corpus stylistics. En M. Burke (Ed.), *The Routledge handbook of stylistics* (pp. 387-392), Routledge.
Mahlberg, M. (2016). Corpus stylistics. En V. Sotirova (Ed.), *The Bloomsbury companion to stylistics* (pp. 139-156). Bloomsbury.
Montolío Durán, E. (1992a). *Gramática en la caracterización de Valle-Inclán. Análisis sintáctico, pragmático y textual de algunos mecanismos de caracterización*. Promociones y Publicaciones Universitarias.
Montolío Durán, E. (1992b). La conciencia lingüística de Valle Inclán: La voluntad de renovar la lengua literaria. *Actas del II Congreso Internacional de Historia de la Lengua Española. Tomo II, Pabellón de España,* 777-786.
Nieto Caballero, G. (2018). Metodologías de corpus en el análisis de textos literarios en lengua española: el ejemplo de Pérez Galdós. *Estudios Humanísticos. Filología, 40,* 371-389.
Nieto Caballero, G. y Ruano San Segundo P. (2020). *Estilística de corpus: Nuevos enfoques en el análisis de textos literarios*. Peter Lang.
O'Halloran, K. (2007). Corpus-assisted literary evaluation. *Corpora, 2*(1), 33-63.
Piccioni, S. (2015). *Lingüística de corpus y literatura. Aproximaciones cuantitativas al análisis del estilo*. Editorial Académica Española.
Scott, M. (1997). PC Analysis of Key Words – And Key Words. *System, 25*(2), 233-245.
Semino, E. y Short, M. (2004). *Corpus Stylistics: Speech, writing and thought presentation in a corpus of English narratives*. Routledge.
Stubbs, M. (2005). Conrad in the computer: Examples of quantitative stylistics methods. *Language and Literature, 14*(1), 5-24.

Linguistic remix – Mapping the intertextual relationship of poetic texts with an n-gram approach

Emese K. Molnár, National Laboratory for Digital Heritage/Eötvös Lorand University, Doctoral School of Linguistics

Andrea Dömötör, National Laboratory for Digital Heritage

The aim of the project is developing an NLP tool that helps the quantitative analysis of intertextual relations between Hungarian poetic texts, and makes possible the corpus-based comparison of texts in the context of the theoretical framework of Remix Studies (RS). To investigate the connection between texts, RS offers a new, complex perspective by interpreting linguistic practices. RS links different disciplines with the aim of understanding the creativity of the new media era that has emerged from the technological changes of the 20th century and characterized by participation, collaboration and reuse (Navas et al., 2021). The concept of remix culture (Lessig, 2008) questioned and renegotiated not only the term of authorship, but also the understanding of creativity and culture. The concept of remix can be adapted successfully to the examination of cultural practices because – as Lessig points out – while the phenomenon of remixing may seem novel, its core mechanism has long been a part of human culture: remixing with (digital) media is identical to the fundamental process of language use. Evoking and incorporating the words of others into written works or conversations is so natural that we do not even notice the borrowing. Popular text generation tools which are based on the Large Language Models (LLMs) owe their success to the fact that they exploit this fundamental linguistic mechanism. These tools learn from large amount of textual data sets to determine the probabilistic values of linguistic patterns for generating textual content. In other words, they remix by recognising and regenerating existing linguistic patterns.

Taking this into account, the main aim of the research is better understanding of the creative potential of text production, i.e. to explore the networks of texts by adapting the concept of remix to linguistics research. This is achieved by extending the functions of the ELTE Poetry Corpus and ELTE Folk Song Corpus of the National Laboratory of Digital Heritage (Horváth 2020a, 2020b) and the ongoing Lyrical Poetry Corpus project of ELTE DiAGram Research Centre for Functional Linguistics (Domonkosi et al., 2018; Horváth et al., 2021). By adding new functions to these corpora they will be connected to international projects that focus on the creative reuse of texts, intertextual relations, recurrent patterns such as Chicago Homer (Kahane & Mueller, 2001), Tesserae (Coffee et al., 2013) or Commonplace Cultures (Gladstone & Cooney, 2020). Following these models the tool is based on an n-gram method to extract patterns for comparing the texts. The extracted trigrams are combinations of words, lemmas and POS tags, thus they not only help to find classic, literal cases of intertextuality, but are useful for exploring more abstract structural similarities of the texts as well. For example in the poem titled White table… by Gyula Juhász there is a line *„Áldott legyen! Nagy útról jöttem én”* (’Blessed be! I came from a great journey’) and an other poem titled I am the son of Góg and Magóg by Endre Ady has a line *„Verecke híres útján jöttem én”* (’I came along Verecke's famous path’). These two lines share two trigrams: ADJ/POS/ + *út* (’journey/path’)/lemma/ + *jöttem* (’came’)/word/ and *út* (’journey/path’)/lemma/ + *jöttem* (’came’)/word/ + *én* (’I’)/word/. The lines are not identical, but they have a similar structure. By finding these structural similarities to reveal the patterns of text production the research contributes to the understanding of linguistic creativity.

References

Coffee, N., Koenig, J-P., Shakti, P., Ossewaarde, R., Forstall, C., & Jacobson, S. (2013). The Tesserae Project: Intertextual analysis of Latin poetry. *Literary and Linguistic Computing, 28*, 221–228. DOI: https://DOI.org/10.1093/llc/fqs033

Gladstone, C., & Cooney, C. M. (2020). Opening new paths for scholarship: Algorithms to track text reuse in ECCO. In S. Burrows & G. Roe (Eds.), *Digitizing enlightenment: Digital humanities and the transformation of eighteenth-century studies* (pp. 353–374). Voltaire Foundation in association with Liverpool University Press.

Domonkosi, Á., Kuna, Á., Simon, G., Tátrai, Sz., & Tolcsvai Nagy, G. (2018). Poétikai mintázatok korpuszalapú kognitív stilisztikai kutatása: A Stíluskutató csoport kutatási terve. In Á., Domonkosi & G., Simon (Eds.), *Nyelv, poétika, kogníció. Elmélet és módszer a poétikai kutatásban* (pp. 211–222). Líceum Kiadó.

Horváth, P. (2020a). A vershangzás jellemzőinek automatikus feltárása József Attila verseiben. *Digitális Bölcsészet, 3*, 3–27. DOI: https://DOI.org/10.31400/dh-hun.2020.3.422

Horváth, P. (2020b). Az ELTE Verskorpusz automatikus annotációs eljárásai révén nyerhető kvantitatív adattípusok. In G., Simon & G., Tolcsvai Nagy (Eds.), *Nyelvtan, diskurzus, megismerés* (pp. 313–332). ELTE Eötvös Kiadó.

Horváth, P., Simon, G., & Tátrai, Sz. (2021). A lírai személyjelölés konstrukcióinak annotálási elveiről. In K., Laczkó & Sz., Tátrai (Eds.), *Líra, Poétika, Diskurzus* (pp. 133–166). ELTE Eötvös Collegium.

Kahane, A., & Mueller, M. (2001). *The Chicago Homer* (Web publication/site). University of Chicago Press/Northwestern University Library. http://digital.library.northwestern.edu/homer/

Lessig, L. (2008). *Remix. Making art and commerce thrive in the hybrid economy.* Penguin Press. DOI: https://DOI.org/10.5040/9781849662505

Navas, E., Gallagher, O., & Burrough, x., (Eds.). (2021). *The Routledge handbook of remix studies and digital humanities.* Routledge. https://DOI.org/10.4324/9780429355875

Logical markers in Academic Writing

Virginia Mattioli, University of Cantabria

This contribution aims to deep in the study of metadiscourse in Academic Writing through a corpus-driven analysis of logical markers (LM) in a set of English academic texts. Metadiscourse is “one of the main ways in which interaction is studied in Academic Writing” (Hyland, 2017) and its main forms of manifestation are connectives, modal expressions and personal pronouns (Luukka, 1994). This paper focuses on logical markers, a type of connectives which express the semantic relation between arguments and ideas. LM play an outstanding role in Academic Writing and can be divided into three classes according to their function: additive (e.g., and, furthermore, in addition, etc.), contrastive (e.g., rather, however, in contrast, etc.) and consecutive (e.g., thus, hence, therefore, etc.) (Dueñas, 2007; Hyland, 2017).

Previous authors (Dueñas, 2007; Hyland, 1998, 2017) highlight the specificity of the metadiscoursal features, including LM, to different academic disciplines. However,

considering their distribution into three different functional classes, it could be reasonable to think that LM present a stricter relationship with the purpose of the text than with the academic disciplines within they are usEd. In this sense, contrastive LM should be more frequent than LM of other classes in persuasive texts and contrastive studies, regardless the academic discipline; consecutive LM seem more likely to characterize those texts focused on cause/effect relationships of any field; despite their presence in any type of text, additive LM could be expected to be used in a greater extent in explicative and descriptive texts.

From here, the main objective of the present paper is determining whether the use and distribution of the LM is more related to the academic disciplines or to the purpose of the text. To do that, this study analyses and compares the distribution of the LM of each functional class in a set of corpora of English academic texts representing five different academic disciplines trying to answer to the following research question: is the distribution of the LM across the three functional classes the same in different academic disciplines? An affirmative answer would corroborate the main hypothesis of the study.

The material analysed includes five corpora compounded of English research articles, textbooks and PhD dissertation representing five different academic disciplines, respectively. Each corpus is further divided in subcorpora representing texts with different purposes (descriptive, contrastive, persuasive, etc.). The study follows a corpus-driven methodology: once adopted an exhaustive list of potential LM classified according to their function (additive, contrastive and consecutive), each one of them is searched for in the concordance list; then, the results are manually reviewed in order to observe them in context and discard those cases in which the items are used with a unique syntactic function. Each corpus is analysed separately and then the results are compared.

The results show the distribution and the specificity of the LM of each functional class in relation to the purpose of the text in each analysed academic discipline, revealing that both variables have an impact on the use of LM and assessing their influence on the selection of LM of each class. Beyond their contribution to metadiscourse studies, the results will benefit register and discourse studies about AW by describing the correlation between the choice of specific classes of connectors, the field and the textual purpose, being an interesting point of departure for further studies.

References

Dueñas, P. M. (2007). Same genre, same discipline; however, there are differences: A cross-cultural analysis of logical markers in academic writing. In C. Williams & D. Milizia (Eds.), *ESP across cultures, 4* (pp. 37-53). Edizioni B.A. Graphis.

Hyland, K. (1998). Persuasion and context: The pragmatics of academic metadiscourse. *Journal of pragmatics, 30*(4), 437-455.

Hyland, K. (2017). Metadiscourse: What is it and where is it going? *Journal of pragmatics, 113*, 16-29.

Luukka, M. R. (1994). Metadiscourse in academic texts. *Text and talk in professional context*, 77-88.

Mapping pre-digital tourism communication of the town of Ferrara and its province between the 1960s and the 2000s: a quantitative and qualitative analysis

Eleonora Federici, University of Ferrara

The present study aims to provide a diachronic overview of print tourism promotional materials published between the 1960s and the 2000s by Italian tourism institutions, specifically focusing on the town of Ferrara and its province. Employing discourse analysis and corpus-based methodologies I will conduct a critical analysis to examine the key linguistic features characterizing a corpus of promotional resources utilized by institutional operators within the tourism industry. (Dann, 1996).

The research seeks to identify and highlight the most effective communicative strategies employed in this sector over time, shedding light on the evolution of tourism discourse in Italy (Mocini, 2013; Hunston and Thompson, 2000) and outlining the changes occurred to a lesser known area of a major tourist destination. By undertaking an in-depth examination of these promotional materials (brochures, catalogues, magazines), my aim is to gain valuable insights into the language choices, rhetorical techniques techniques, and discursive practices employed by *ENIT* (*National Tourism Board*) and public operators (Province and Regional Tourist boards). I will try to answer the following research questions:

1. What are the discursive strategies employed to promote tourism in these materials?

2. How are the town of Ferrara and other locations portrayed and evaluated in the promotional

discourse?

3. Which are the key themes utilized to narrate the territory and make it more appealing?

4. How are visual and verbal strategies related?

In answering these questions I will take into account both linguistic and the extra-linguistic features (including cultural and ideological factors) with the aim of verifying the extent to which such touristic material contributes to the creation of a “postcard effect”, that is a standardised representation of Italian tourist destinations within the English-speaking world. Following Dann (1996), the resulting image of this area as the cradle of the Renaissance and a new ‘must destination’ for sustainable tourism will be discussed as a powerful means of attracting international tourists and upholding the country’s reputation abroad.

By addressing these questions, I will gain a deeper understanding of the specific towns being promoted, the consistency or variation in promotional trends over time, the discursive

and visual techniques adopted to attract tourists, and the ways in which these destinations are presented and narrated for international tourists.

References

Dann, G. (1996). *The language of tourism.* CAB International.
Hunston, S., & Thompson G. (Eds.) (2000). *Evaluation in text: Authorial stance and the construction of discourse.* Oxford University Press.
Mocini, Renzo. The promotional functionality of evaluative language in tourism discourse. *Lingue e Linguaggi, 9*(2013), 157-172.

Mean Dependency Distance in Contemporary Czech Language - a Genre Analysis

Xinying Chen, University of Ostrava

Miroslav Kubát, University of Ostrava

The objective of this study is to examine the syntactic complexity present in a range of text types within contemporary written Czech. To achieve this, we employ the Mean Dependency Distance (MDD) index, as introduced by Liu in 2008. We aim to discover how MDD values vary across different genres and explore the effectiveness of MDD in the field of stylometry.

Mean Dependency Distance measures the average distance between words and their syntactic dependents in sentences. Essentially, it quantifies how far apart words that are directly related in a sentence's grammatical structure tend to be. A lower MDD indicates that related words are generally closer together, suggesting a more compact sentence structure, while a higher MDD signifies that related words are more spread out, indicating a more complex or convoluted sentence structure. This metric can be crucial in comparing and understanding the syntactic characteristics of different texts, as well as in studying the cognitive aspects of language processing, as it relates to how information is organized and accessed in our minds during communication.

The analysis is based on the data from the Czech National Corpus. Specifically, we use corpus SYN2020 (Křen et al., 2020). It is a big balanced corpus of contemporary written Czech with a size of 100 million words. The texts are from the period of 2015–2019. There are three main style groups: fiction, non-fiction, and newspapers and magazines. Each is further divided into detailed subcategories like novels, poetry, drama, administrative texts, etc., allowing more detailed genre analysis. SYN2020 besides lemmatization and morphological annotation has also syntactic annotation (Jelínek et al., 2021). This annotation is based on the Prague Dependency Treebank (Bejček, 2012). This annotation marks dependency relations between words and assigns syntactic functions.

The results of the analysis (see Table below) reveal a range of MDD values across different genres, from the lowest MDD of 2.15 in drama and screenplays (SCR) to the highest MDD of 2.70 in scientific literature (SCI). These findings suggest a clear variation in syntactic complexity between genres, with more creative forms such as drama and screenplays (SCR) and poetry (VER) tending towards lower MDD values. This could indicate a preference for a more direct and compact sentence structure, potentially due to the need for brevity and clarity in spoken dialogue and the rhythmic constraints of poetry. Scientific literature (SCI) stands out with the highest MDD value of 2.70, which is not surprising given the typically complex sentence structures that convey detailed and precise information in academic writing. This could reflect the cognitive demand and the high level of syntactic organization required to articulate scientific concepts effectively.

In summary, these results suggest that genre is an important factor in the syntactic structuring of Czech texts. These insights underscore the utility of MDD as a metric in the field of stylometry and offer a quantitative foundation for understanding syntactic variation across genres.

References

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., & Žabokrtský, Z. (2012). Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)* (pp. 231-246). Mumbai.

Jelínek, T., Křivan, J., Petkevič, V., Skoumalová, H., Šindlerová, J. (2021): SYN2020: A new corpus of Czech with an innovated annotation. In K. Ekštein, F. Pártl & M. Konopík (Eds.), *Text, Speech, and Dialogue. TSD 2021. Lecture Notes in Computer Science* (Vol. 12848) (pp. 48-59). Springer.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., & Škrabal, M. (2020). *SYN2020: Representative corpus of contemporary written Czech.* Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. http://www.korpus.cz.

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

Metáforas conceptuales y marcos semánticos: análisis de la percepción de los movimientos sociales franceses en el corpus de prensa

Estéfano Rodríguez-Peláez, University of Granada/ University of Nice

El periodismo narrativo de los medios de comunicación, mediante la creación de relatos y narrativas que se presentan de manera persuasiva, tiene un impacto profundo en nuestra percepción de la realidad, dado que dichos relatos pueden moldear y dar forma a la manera en que comprendemos y experimentamos el complejo entramado de acontecimientos,

fenómenos y dinámicas que caracterizan el mundo contemporáneo (Zizek, 2012). Esta influencia, que se ejerce a través de la selección de historias y eventos para su presentación, así como a través de la construcción de discursos y enfoques narrativos, va más allá de la simple transmisión de información. Juega un papel fundamental en la configuración de nuestra perspectiva y en nuestra capacidad para tomar decisiones informadas y efectuar acciones en la sociedad en la que estamos inmersos. Proponemos una comunicación que analiza la percepción del lector español en relación con los movimientos sociales franceses más influyentes de la última década: *Nuit debout* (La Noche en Pie) y los *Gilets jaunes* (los Chalecos amarillos). Exploramos cómo las metáforas conceptuales (Lakoff y Turner, 1991; Lakoff y Johnson, 2002) utilizadas en la prensa española y los marcos semánticos (Buendía-Castro y Sánchez-Cárdenas, 2012; 2016; Faber, 2015; Faber y Mairal-Usón, 2017) identificados a través del proyecto FrameNet (Fillmore, 2008; Baker, 2009) pueden influir en la interpretación de estos acontecimientos. Nuestro enfoque se centra en cuatro destacados periódicos españoles: *El País*, *El Mundo*, *ABC* y *La Razón*. Compilamos manualmente un corpus *ad hoc* disponible en línea enriquecido con datos de la base de datos FACTIVA. A través de un análisis minucioso con dos herramientas de análisis de corpus, Atlas.TI (Friese, 2019) y Sketch Engine (Kilgarriff *et alii*., 2004; Kilgarriff *et alii.,* 2010; Kilgarriff *et alii.,* 2014), hemos identificado metáforas conceptuales recurrentes que se utilizan para describir y dar sentido a estos movimientos. Observamos que los periódicos españoles a menudo recurren a metáforas que comparan estos movimientos a catástrofes naturales o elementos de la naturaleza, lo que parece destinado a facilitar la comprensión del lector. Por ejemplo, se han encontrado expresiones como "la ola de Nuit debout barriendo las calles de París" o "los Chalecos amarillos desatando tormentas de protesta". Estas metáforas conceptualizan los movimientos sociales como fuerzas incontrolables de la naturaleza que irrumpen en la sociedad y después se extinguen, lo que puede influir en la percepción del lector sobre la intensidad y efímera duración de estos eventos. Además, para comprender mejor cómo estas metáforas influyen en la percepción del lector, relacionamos estas metáforas conceptuales con los marcos conceptuales que se activan en la prensa según el proyecto FrameNet. Comprobamos que los marcos semánticos Manifestación (*To protest*), Violencia (*Violence*) y Arresto (*Arrest*) se activan de manera prioritaria. El objetivo de esta investigación es arrojar luz sobre la influencia de las metáforas conceptuales y los marcos semánticos en la percepción del lector en relación con el corpus de la prensa sobre movimientos sociales. Comprender cómo la prensa configura la percepción de los eventos es fundamental en un mundo mediático donde la información desempeña un papel crucial. Los resultados de nuestro estudio ofrecen una perspectiva valiosa sobre cómo los medios pueden moldear la comprensión de eventos sociales y políticos, lo que tiene implicaciones tanto para la comunicación como para la construcción del discurso en la sociedad actual.

Referencias

Baker, C. F. (2009). La sémantique des cadres et le projet FrameNet: Une approche différente de la notion de «valence». *Langages*, (4), 32-49.

Buendía-Castro, M. y Sánchez-Cárdenas, B. (2012). Linguistic knowledge for specialized text production. In *LREC* (pp. 622-626).

Buendía-Castro, M. y Sánchez-Cárdenas, B. (2016). Using argument structure to disambiguate verb meaning. In *Proceedings of the XVII EURALEX international congress* (pp. 482-490).

Faber, P. y Mairal Usón, R. (2017). The functional lexematic model: Past, present and future. En J.A.H.C. Cutillas Espinosa, R. Manchón Ruiz y F. Mena Martínez (Eds.), *Estudios de Filología Inglesa* (pp. 315–340). Editum.
Faber, P. (2015). Frames as a framework for terminology. *Handbook of Terminology, 1*(14), 14-33.
Fillmore, C. J. (2008). Border conflicts: FrameNet meets construction grammar. En *Proceedings of the XIII EURALEX international congress* (Vol. 4968). Universitat Pompeu Fabra.
Friese, S. (2019). Qualitative data analysis with ATLAS.ti. Qualitative data analysis with ATLAS.ti, 1-344.
Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, R. y Suchomel. V. (2014). The sketch engine: Ten years on. *Lexicography, 1*, 7-36.
Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I. y Tiberius, C. (2010). A quantitative evaluation of word sketches. En *Proceedings of the XIV Euralex international Congress* (pp. 372-379).
Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D., Williams, G. y Vessier, S. (2004). En *Proceedings of the 11th EURALEX International Congress* (pp. 105-115).
Lakoff, G. y Turner, M. (1991). More than cool reason: A field guide to poetic metaphor. *Language, 67*(2), 320. DOI: https://DOI.org/10.2307/415109
Lakoff, G. y Johnson, M. (1980/2002). *Metaphors we live by* (Second Edition). University of Chicago Press.

Morphosyntactic and pragmatic variation in if/si-constructions: A corpus-based analysis of English and Spanish newspaper discourse

Cristina Lastres-López, University of Seville

In this paper, I explore morphosyntactic and pragmatic variation in constructions introduced by if in English and their Spanish equivalents introduced by si. Building on prior research on these constructions in spoken discourse (Lastres-López, 2018, 2019, 2020, 2021), the aim is to extend the analysis to written discourse in order to (i) examine how these constructions are used in newspaper discourse in particular, and (ii) unveil differences across speech and writing. Prior research has shown that, in addition to conveying prototypical conditional meaning, these constructions can also encode a wider range of functions in discourse (Ford & Thompson, 1986; Ford, 1997; Warchal, 2010; Lastres-López, 2020, 2021). However, with the notable exception of Ford and Thompson (1986), corpus-based studies exploring variation in these constructions in spoken and written discourse are scarce.

The analysis considers both patterns of subordination and insubordination. Within the former pattern, I explore conditional constructions, including both prototypical cause-consequence patterns, as in examples (1) and (2), and also other conditionals in which the conditional meaning is weaker, as illustrated in (3) and (4). On the other hand, the study also encompasses cases of insubordination (Evans, 2007; Evans & Watanabe, 2016; Beijering et al., 2019), in which an if/si-clause is used as a stand-alone construction whose meaning is fully complete in the discourse situation. The corpus analysis allows us to determine whether

if/si-insubordination, as shown in (5), is exclusive of spoken discourse or whether these constructions have also entered into written discourse.

(1) If they cannot agree, the government will soon run out of money, leaving it unable to pay items such as social security payments, military pay and interest payments. (CONTRAST-IT cnt_en_gua_bus_010)

(2) Si la inyección económica no viene, lo vamos a pasar mal (CONTRAST-IT cnt_es_mun_dep_006)

'If the economic injection does not arrive, we are going to have a hard time'

(3) We didn't move out or anything, that would have been too expensive, but we gave each other space if you get my meaning. (CONTRAST-IT cnt_en_gua_cul_004)

(4) Quería edificar su victoria a partir de un riesgo innecesario si se quiere (CONTRAST-IT cnt_es_pai_dep_024)

'He wanted to build his victory on an unnecessary risk if you want'

(5) If you could give me a couple of 39c stamps please (Evans, 2007: 380).

Rooted in the framework of Systemic Functional Linguistics (Halliday & Matthiessen, 2014), the data are analysed in terms of metafunction, distinguishing between if/si-constructions at the ideational, interpersonal and textual levels. Furthermore, the tokens are also annotated according to the following variables: degree of likelihood of the condition, position of the if/si-clause, markedness of the apodosis and presence of a modal verb in the apodosis.

Corpus data are retrieved from the English and Spanish components of CONTRAST-IT (De Cesare, 2018), which comprise data from online newspapers featuring different thematic sections (politics, economy, sports, etc.). Preliminary results show morphosyntactic and pragmatic variation in English and Spanish, as well as some differences in how these constructions are used in newspaper discourse as opposed to prior studies examining speech (Lastres-López, 2021). While spoken discourse shows higher proportions of interpersonal conditionals, written data in both languages indicates that these constructions are mostly used in newspaper discourse in their ideational metafunction.

References

Beijering, K., Kaltenböck, G., & Sansiñena, M. S. (2019). (Eds.). Insubordination: Theoretical and Empirical Issues. Mouton de Gruyter.

De Cesare, A. M. (2018). CONTRAST-IT. University of Basel. Available online at: https://contrast-it.philhist.unibas.ch/en/home/

Evans, N. (2007). Insubordination and its uses. In I. Nikolaeva (Ed.), Finiteness: Theoretical and empirical foundations (pp. 366-431). Oxford University Press.

Evans, N. & Watanabe, H. (2016). (Eds.). Insubordination. John Benjamins.

Ford, C. (1997). Speaking conditionally: Some contexts for if-clauses in conversation. In A. Athanasiadou & R. Dirven (Eds.), On conditionals again (pp. 387-413). John Benjamins.

Ford, C. & Thompson, S. (1986). Conditions in discourse: A text-based study from English. In E. C. Traugott, A. ter Meulen, J. S. Reilly & C. A. Ferguson (Eds.), On conditionals (pp. 353-372). Cambridge University Press.
Halliday, M. A. K. & Matthiessen, C. (2014). Halliday's introduction to Functional Grammar. Routledge.
Warchal, K. (2010). Moulding interpersonal relations through conditional clauses: Consensus-building strategies in written academic discourse. Journal of English for Academic Purposes, 9, 140-150. https://doi.org/10.1016/j.jeap.2010.02.002

Navigating Gendered Terrain: An Analysis of Politeness and Impoliteness Strategies in Female Instructional Texts

Walter Yared Armas Cáceres, University of La Laguna

Throughout history, the scientific realm has long been entrenched in a pervasive culture of male dominance, with gender biases deeply ingrained. This has created systemic barriers for women in science, impeding their full participation and recognition within the field. (Golinski, 2022). Women have encountered challenges in navigating and constructing their identities within this distant and traditionally male-dominated territory. This paper focuses on unweaving the relevance of discourse analysis research and contemporary technologies, probing their application in the examination and analysis of instructional texts. Reflecting on the insightful theories of politeness and impoliteness proposed by scholars such as Culpeper (2015) and Lakoff (2004), the goal is to identify the most prevalent maxims found in instructional texts and explore the reasons behind their use. I have, therefore, analysed the polite and impolite maxims in *The Housekeeper's Pocket Book* (Harrison, 1733), a guide and family cookbook that contains instructions for dishes, entertainment, and even monthly bills for seasonal provisions. Due to the quantitative nature of my findings and drawing from the aforementioned theories, I have provided a systematic classification of strategies for both politeness and impoliteness that can be applied to the text. Applying the sustained framework of Lakoff's *Language and Women's Place: Text and Commentaries* (2004), to determine which strategies are most associated with women, and Culpeper's Impoliteness Strategies (2015) to analyse which impolite strategies may also appear, I have reached some deductions. Thus, my analysis illustrates, despite the limitations of the sample size, that Harrison strategically employs impolite communication strategies as a means of crafting and asserting her identity. This deliberate choice reflects her conscious effort to navigate the challenges posed by the traditional male-dominated territory of the time. By opting for impolite strategies, Harrison may be seeking to establish a sense of authority and command attention in an environment where women were often marginalized or overlooked. This conclusion possibly underscores the necessity for deconstructing our biased notions about gender and considering that women do not possess a distinguished way of writing. In other words, her style is neither less authoritarian nor less complex. Therefore, discrediting the belief that men use more impoliteness than women in their writings. In the long term, my overarching objective is to analyse a diverse selection of instructive works authored by women to assess the applicability of this thesis to the whole corpus of female writing. With a broader horizon, the incorporation of artificial intelligence into the analysis of instructive texts could incorporate sentiment analysis tools to gauge the emotional tone of texts. Additionally, even natural language processing algorithms can be included to automate the analysis of

text data, such as syntactic parsing, or identification of politeness patterns. Eventually, this text will be part of the Corpus of Women's Instructive Writing in English (CoWITE).

References

Culpeper, J. (2015). Impoliteness strategies. In A. Capone, F. Lo Piparo & M. Carpezza (Eds.), *Perspectives in Pragmatics, Philosophy and Psychology* (pp. 421-445). Springer.

Goliński, J. (2002). The care of the self and the masculine birth of science. *History of Science, 40*(2), 125-145. DOI: https://DOI.org/10.1177/007327530204000201

Harrison, S. (1777). *The house-keeper's pocket-book, and compleat Family cook.* https://www.loc.gov/item/75313835/

Lakoff, R. T. (2004). *Language and woman's place: Text and commentaries.* Oxford University Press. https://ebookcentral-proquest-com.accedys2.bbtk.ull.es

Nueva terminología en el campo de los videojuegos: un estudio de corpus

Carmen Isabel Luján-García, University of Las Palmas de Gran Canaria

Como es bien sabido, el empleo de unidades léxicas anglicadas en las distintas lenguas romances, el francés, italiano y español, se ha convertido en algo muy habitual. En este trabajo, nos centraremos en el análisis de anglicismos en el entorno de los foros de videojuegos. Algunos trabajos anteriores dentro del campo del videojuego (Cabrera Álvarez, 2015; Morales Ariza, 2015; Mezones Galicia, 2015; Cotelo García, 2021; Ramírez Sánchez, 2022 ponen de manifiesto la tendencia al alza al empleo por parte de

hablantes de español a usar términos provenientes del inglés para referirse a todo lo relativo al videojuego. Para ello, los gamers hispanohablantes han hallado mecanismos para convertir palabras inglesas en españolas.

Con este trabajo se persigue dar respuesta a las siguientes preguntas de investigación:

1. ¿Se usan unidades léxicas anglicadas en el campo de los foros de videojuegos en

español?

2. ¿Qué tipo de adaptaciones se da entre los anglicismos extraídos del corpus?

3. ¿Con qué grado/frecuencia se usan de anglicismos adaptados en el corpus extraído?

El método consistió en la recogida de una muestra de 100.340 palabras del foro en línea llamado Foro3Djuegos (https://www.foro3djuegos.com/). La elección de este portal se debe a que es bastante activo y cuenta con múltiples foros en los que se puede interactuar con otros usuarios de distintos vídeojuegos tales como Playstation 3, 4 y 5, X box Series X, X box One, X box 360, Nintendo Switch, Nintendo Wii U, Nintendo 3DS, Nintendo Wii, Stadia, PSP, PS Vita, entre otros. https://www.foro3djuegos.com/foro-de-juegos/3080/0/0/todos/videojuegos/. La compilación de este corpus se realizó en julio de 2023 y una vez se finalizó la compilación, se procedió a insertarlo en la herramienta libre de análisis de corpus lingüísticos llamada *AntConc*.

Los resultados arrojaron luz sobre el creciente empleo de términos adaptados del inglés al español mediante las cuales muchos de los vocablos ingleses reciben la terminación –ar con el objeto de convertirlos en verbos en español. La muestra extraida está compuesta de los términos: banear, boostear, craftear, dropear, farmear, formatear, fut-ballers, gametar, levelear, linkear, moder, plis, postear, raidear, rankear, stunnear, trollear y testear. Con respecto a la frecuencia de uso de la muestra extraída del corpus analizado, la mayoría de los vocablos adaptados objeto de este estudio se emplean de una a dos veces en el corpus, por lo que podemos considerarlo una frecuencia baja. En cualquier caso, se trata de términos anglicados que están actualmente en uso en nuestra lengua y que probablemente tiendan al alza. Se podría concluir que la lengua inglesa se ha convertido en la principal lengua encargada de designar las nuevas realidades que van apareciendo en el terreno del videojuego y en cualquier ámbito vinculado la tecnología de forma más directa o indirecta.

Referencias

Cabrera Álvarez, C. (2015). *El léxico de las revistas de videojuegos españolas: Propuesta de normalización terminológica.* [Tesis Doctoral: Universidad de Las Palmas de Gran Canaria].

Cotelo García, R. (2021). El lenguaje de los videojuegos: Anglicismo y creatividad léxica en la plataforma Twitch. En B. Flores, A. Salud & R. Pérez (Eds.).

Mezones G. y Manuel M. (2015). *Anglicismos utilizados en los videojuegos online de tipo Shooters y Rpg.* [Tesis Doctoral: Universidad César Vallejo].

Morales Ariza, L. (2015). La terminología “gamer” en el contexto del videojuego multijugador en línea. *Revista Electrónica del Lenguaje, 2*, 1- 26.

Ramírez Sánchez, I. (2022). El léxico de los videojuegos y el ingenio de los jugadores. *La Provincia.* 30/05/2022. https://www.laprovincia.es/opinion/2022/05/30/lexico-videojuegos-ingenio-jugadores-66686883.html

Old English text generation. A viable strategy of data augmentation?

Ana Elvira Ojanguren López, University of La Rioja

This paper intends to be a contribution to Natural Language Processing (Clark, Fox and Lappin 2010; Jurafsky & Martin 2020) and, more concretely, to the computational processing of historical languages (Piotrowski 2022). It deals with Old English in the context of the generalisation of artificial intelligence caused by the availability of Large Language Models. Large Language Models (LLMs; Rothman 2022), including GPT, LLaMA, or Mistral, are probabilistic computational models capable of predicting subsequent words within a given string and of identifying the missing word in a certain context. They can perform more specific tasks of Natural Language Processing, such as parsing, translation and text generation. Against this backdrop, this paper focuses on text generation because historical languages, including Old English, do not have enough written records to be processed computationally on a general basis (Schmid 2019; Anastasopoulos et al. 2020). The state of the art in computational processing with artificial intelligence is that data augmentation (Boulanger 2023) can be guided by the interaction with databases by means of Retrieved Augmented Generation (RAG; Lewis et al. 2021). This brings about a change in the paradigm of the interaction with LLMs. Instead of pre-training the LLM and fine-tuning it for a specific dataset or task, the new paradigm consists of pre-training, prompting and predicting (Liu et al. 2023). In other words, the linguistic interaction with the LLM through its language system becomes the central aspect because the textual prompt should be formulated in such a way that it is ordered to perform a task that maximally resembles those performed during the original training of the LLM (Brown et al. 2020). The aim of the paper is to carry out a preliminary test of LLMs as to the task of generating Old English text. The methodology includes the interaction with the LLM, or prompt building, and the assessment of the linguistic output. Prompts will be guided by the textual material in Tolkien's (2014) Sellic Spell, both in the Present Day English and in the Old English versions. The LLMs under review are GPT, LLaMA, Mistral and Bard. The language systems used for the analysis are ChatGPT, Copilot, HuggingChat and Bard. The assessment includes morphological and syntactic well-formedness as well as lexical selection. Syntactic well-formedness will be assessed on the basis of the categories and relations of the framework of Universal Dependencies (de Marneffe et al. 2021). Lexical selection will be checked against the relational databases of the Nerthus Project (www.nerthusproject.com), which will also provide lemmatised frequency lists and attested morphological generation. The results of this initial test could determine if generating Old English text is a viable strategy of data augmentation.

References

Anastasopoulos, A., Cox, C., Cruz, H., & Neubig, G. (2020). Endangered languages meet modern NLP. In Proceedings of the 28th International Conference on Computational Linguistics (Volume 1) (pp. 39-45).

Boulanger, H. (2023). Data augmentation and generation for natural language processing (Doctoral dissertation, Université Paris-Saclay). English. NNT: 2023UPASG019.

Brown, T. et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Clark, A., Fox, C., & Lappin, S. (Eds.). (2010). The handbook of computational linguistics and natural language processing. Oxford: Wiley-Blackwell.

de Marneffe, M.C. et al. (2021). Universal dependencies. Computational Linguistics, 47(2), 255-308.

Lewis, P. et. al. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv: 2005.11401v4 [cs.CL] 12 Apr 2021.
Liu, P. et al. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys, 55(9), 195. https://doi.org/10.1145/3560815
Piotrowski, M. (2022) [2012]. Natural Language Processing for Historical Texts. Springer.
Rothman, D. (2022). Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, Hugging Face, and OpenAI's GPT-3, ChatGPT, and GPT-4. Packt Publishing.
Tolkien, C. (Ed.) (2014). Beowulf. A Translation and Commentary by J.R.R. Tolkien. Together with Sellic Spell. London: Harper Collins Publishers.

Language models and systems

OpenAI. (2022). ChatGPT [Software]. https://openai.com/chatgpt
Microsoft. (2023). Copilot [Software]. https://copilot.microsoft.com/
Hugging Face. (2023). Huggingchat [Software]. https://huggingface.co/
Google. (2023). Bard [Software]. https://ai.googleblog.com/2022/04/introducing-bard.html

On generalized -s in nineteenth-century representations of south-western dialects

Javier Ruano, University of Salamanca

Generalized -s is one of the grammatical characteristics of the traditional dialects of the South-West of England (e.g. Ihalainen, 1994, p. 212) that has been reported to be in complementary distribution with periphrastic DO (Klemola, 2018 p. 282-89), and which refers to the use of verbal -s with subjects other than the third-person singular; thus:

(1) a. Zo I *buys* haaf a poun a gunpowder, an chuckled to mezelf

b. I *knaws* when I'm well-off

c. Thee *thinks* I be a zaft-head

d. they as *works* ael day lang and ael the year round

It has traditionally been associated with habitual aspect, as well as with punctual and continuous events (see Godfrey & Tagliamonte, 1999, p. 105-8; de Both, 2019, p. 29), while it has been used as a narrative marker, with the verb say predominating in formulaic expressions with subject I (e.g. I says, says I). Godfrey & Tagliamonte (1999, p. 107) argue that "[s]ince this patterning is not attested in the historical literature […] it may be a later development in which verbal -s has been reanalyzed as a marker for iconically ordered narrative events" or "as a marker of historic present regardless of person and number of the verb". Even though generalized *-s* has been well reported in the literature (see Rupp & Britain, 2019, p. 72–85 for a review), its distribution and contexts of use prior to the Survey of English Dialects (Orton et al., 1962–71) remain rather obscure. In fact, Wagner (2012, p.

926) underlines that early accounts of dialect tend to discuss non-standard features such as verbal *-s* "in terms of their presence (or absence)" with little information (if at all) about their "frequencies (relative and absolute) and distributional patterns".

This paper seeks to engage with ongoing discussion on generalized *-s* by looking at nineteenth-century representations of south-western dialects included in the Salamanca Corpus. I undertake a "frequentist approach" (de Both, 2019, p. 5) to determine whether literary recreations of dialect can inform our historical understanding of the frequency and geographical distribution of this grammatical feature. The analysis draws on thirteen representations of the dialects of Cornwall, Devonshire, Dorset, Somerset and Wiltshire, and examines all the positive declarative sentences with present verbs, which have been

annotated manually for dialect and aspect. The findings show, on the one hand, that generalized -s is a low-frequency feature and that, in line with later evidence (e.g. Klemola, 2018), few or no *-s* forms are documented in dialects where periphrastic DO is used (e.g. Somerset); however, *-s* marking is not more commonly used in representations of dialects where preverbal DO is rare, as in Devonshire (see Peitsara, 2002). On the other hand, the data indicates that instances of verbal *-s* are not confined to express habituality, being more commonly employed to encode non-habitual aspect and as a narrative marker, especially with the verbs *say, go* and *know*. The argument is made, therefore, that nineteenth-century representations of dialect not only report on and testify to the presence of a feature that was evaluated as characteristic of some south-western dialects. They also reveal patterns of historical distribution while they shed some light on patterns of use attributed to later developments.

References

de Both, F. (2019). Nonstandard periphrastic DO and verbal -s in the south west of England. *Journal of Historical Sociolinguistics, 5*(1), 1–35.

Godfrey, E. & Tagliamonte, S. (1999). Another piece for the verbal -s story: Evidence from Devon in southwest England. *Language Variation and Change, 11*, 87–121.

Ihalainen, O. (1994). The dialects of England since 1776. In R. Burchfield (Ed.), *Cambridge History of the English Language* (Vol. 5) (pp. 197–274). Cambridge University Press.

Klemola, J. (2018). The historical geographical distribution of periphrastic DO in southern dialects. In L. Wright (Ed.), *Southern English then and now* (pp. 262–92). De Gruyter.

Orton, H., Halliday, W., Barry, M., Tilling, Ph., & Wakelin, M. (Eds.). (1962–71). *Survey of English dialects (B) The basic material, 4 vols*. E. J. Arnold & Son.

Peitsara, K. (2002). Verbal -s in Devonshire: The Helsinki dialect corpus evidence. In H. Raumolin-Brunberg & T. Nevalainen (Eds.), *Variation past and present: VARIENG studies on English for Terttu Nevalainen* (pp. 211–230). Société Néophilologique.

Rupp, L. & Britain, D. (2019). *Linguistic perspectives on a variable English morpheme. Let's talk about -s.* Palgrave.

The Salamanca Corpus: Digital Archive of English Dialect Texts. (2011–). Compiled by M. F. García-Bermejo Giner, P. Sánchez-García & J. Ruano-García. Universidad de Salamanca. www.thesalamancacorpus.com.

Wagner, S. (2012). Late modern English: Dialects. In A. Bergs & L. J. Brinton (Eds.), *English historical linguistics: An international handbook* (Vol. 1) (pp. 915–38). De Gruyter.

Isabel Sofía Moskowich-Spiegel Fandiño, University of Coruña

María Begoña Crespo García, University of Coruña

This work aims at presenting a new subcorpus in the Coruña Corpus of English Scientific Writing (CC) which is currently under compilation. It is the case of CEGeT, Corpus of English Geography Texts.

One of the distinctive characteristics of the CC has always been to gather scientific texts written during what is traditionally known as “late Modern English period” in several twin corpora sharing identical principles (Monaco and Puente-Castelo, 2019; Moskowich 2019; Crespo and Moskowich, 2020). The first step of the compilation involved the delimitation of the time-span to be covered and the size of samples gatherEd. Although there are many different proposals for the periodisation of late Modern English, we agreed to accept 1700 and 1900 as time limits considering the situation of science at both ends. Concerning sample size, and after examining some of the texts from the period, we opted for the collection of 10,000-word extracts. At the moment of making this decision, specialised corpora were just a few and not much information about them was available, except for Biber’s (1993) claim that variation in specialised registers could be detected and studied in 1,000-word samples. Both our close reading of the material and our intention not to collect huge amounts of texts convinced us to select 10,000 words as the size for samples, instead. This decision, together with our determination to use XML-TEI both for text and metadata files, are now supported by other corpus compilers (VARIENG, 2016) that have partly adopted similar decisions. The Coruña Corpus has been organised into different sub-corpora, one per discipline, with the intention of representing late Modern specialised texts. Our delimitation of disciplines was based on the criteria of the historical moment in which texts were published, that is, that is, we adopted an inclusive perspective. All in all, the divisions proposed by UNESCO in 1988 were used as a starting point. This way, currently, several subcorpora have been published: The Corpus of English Texts on Astronomy (CETA, 2012) and the Corpus of English Philosophy Texts (CEPhiT, 2016), the Corpus of History English Texts (CHET, 2019) and the Corpus of English Life Sciences Texts (CELiST, 2020). Others, however, are found in different stages of the compilation process. This is the case of the Corpus of English Chemistry Texts (CECheT), the Corpus of English Texts on Languages (CETeL), the Corpus of English Texts on Physics (CETePh) and, finally, the Corpus of English Geography Texts (CEGeT), the object of study in this paper.

Each text sample is accompanied by a metadata file providing information both about the author and the text itself. Such metadata files, in combination with the Coruña Corpus Tool, can be used to narrow searches according to extralinguistic parameters (sex, age or geographical provenance of the author as well as date of publication, genre of the sample, etc.). It is precisely the contents of these files that will be presented as they embody the idiosyncrasy of CEGeT.

References

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*, 243-257.
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.
Lareo, I., Monaco, L. M., Esteve-Ramos, M. J., & Moskowich, I. (comps.) (2020). *Corpus of English Life Sciences Texts.* A Coruña: Universidade da Coruña. DOI: https://DOI.org/10.17979/spudc.9788497497848
LMEMT. (n.d.). Retrieved 24 June, 2016, from VARIENG research Unit for variation, contacts and change in English. http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/LMEMTind ex.html
Monaco, L. M. & Puente-Castelo, L. (2019). "A matter both of curiosity and usefulness": Compiling the Corpus of English Texts on Language. *Research in Corpus Linguistics, 7*, 47-68.
Moskowich, I., Lareo, I., Lojo Sandino, P., & Sánchez-Barreiro, E. (comps.) (2019). *Corpus of History English Texts.* Universidade da Coruña. DOI: https://DOI.org/10.17979/spudc.9788497497091
Moskowich, I. (2019). "An introduction to CHET, 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*, 42-56.
Moskowich, I., & Crespo, B. (2012). Corpus of English texts on astronomy (CETA). [CD-Rom] Included in *Astronomy 'playne and simple'. The writing of science between 1700 and 1900.* John Benjamins.
Moskowich, I., & Crespo, B. (2016). Classifying communicative formats in CHET, CECHeT and others. *EPiC Series in Language and Linguistics, 1,* 308-320.
Moskowich, I., Camiña-Rioboo, G., Lareo, I., & Crespo, B. (2016). Corpus of English philosophy texts (CEPhiT). [CD-Rom] Included in *'The Conditioned and the Unconditioned'. Late Modern English texts on philosophy.* John Benjamins.
Puente-Castelo, L., & Monaco, M. (2016). it is proper subserviently, to inquire into the nature of experimental chemistry': Difficulties to harmonize disciplinary particularities and compilation criteria during the selection of samples for CECheT. *EPiC Series in Language and Linguistics, 1*, 351-360.
UNESCO. (1988). *Proposed international standard nomenclature for fields of science and technology.* UNESCO.

Prescriptivism and pronominal case variation from 1710 to 1920

Miriam Criado-Peña, University of Granada

The English pronominal system has been the matter of heated debate among grammarians for the last 300 years. This is especially the case of the dichotomy nominative-accusative pronoun forms *I/me*; *he/him*; *she/her*; *we/us*; and *they/them* in a set of linguistic contexts. In Present-day English, the phenomenon does not appear to be a matter of correct vs. incorrect language, but one of formal vs. informal language, although a different state of affairs is found in earlier periods of English. Since the eighteenth century, a great concern for correctness grew among scholars, resulting in an upsurge in the publication of manuals as an attempt to regularize the language and "enforce a uniformity and conformity to some absolute standard" (Drake, 1977, p. 1). Throughout the century, the number of publications continued to increase, especially after the 1750s, leading to what is today known as the 'age

of prescriptivism', which culminated at the end of the following century (Tieken-Boon van Ostade, 2009, p. 3). Among the manifold grammatical aspects of English discussed in the books, a common point of concern among grammarians was the correct use of pronouns. Usage books, however, reveal a lack of consensus on the matter, and a look at the evidence demonstrates significant variation on their usage during the late Modern period. In this respect, some authors defended the use of the nominative case regardless of the linguistic context (e.g. Webster, 1804), others argued that case agreement was needed in such situations (e.g. Lowth, 1763) and there were also a few who supported the idea that the accusative case was always required (e.g. Priestley, 1771).

The use of personal pronouns in the late Modern English period has received some scholarly attention in the recent years, although the phenomenon has been mainly approached from a synchronic perspective, either focusing on the eighteenth century (Tieken-Boon van Ostade, 1994), on the nineteenth century (Nakayama, 2015), or on modern British English (Wales, 1996; Quinn, 2005) and American English (He & Wen, 2013). As far as I have been able to investigate, however, no studies have delved into case variation from a diachronic corpus-based approach. In light of this gap in the literature, the present paper aims to investigate the correlation between prescriptivist norms and actual usage as regards English pronominal forms over the course of the eighteenth and nineteenth centuries. Precepts are studied in view of prescriptive commentary in grammars whereas quantitative data is retrieved from CLMET3.1 (*The Corpus of Late Modern English Texts*). The purpose of this piece of work is therefore twofold: a) to analyse the historical distribution of pronoun case forms in postverbal position in the period 1710-1920; and b) to evaluate the influence that prescriptive rules may have had upon their actual usage over time. Case variation is thus explored in three different linguistic environments, including the use of pronouns after the linking verb *be* (e.g. 'it is me/I') and after the conjunctions *than* (e.g. 'she is taller than me/I') and *as* (e.g. 'she is as tall as me/I').

References

Drake, G. F. (1977). *The role of prescriptivism in American linguistics, 1820–1970.* John Benjamins.

He, Q., & Wen, B. (2015). Reflections on the grammatical category of the than element in English comparative constructions. *Ampersand, 2*, 101–108.

Lowth, R. (1763). *A short introduction to English grammar: With critical notes.* A. Millar and J. Dodsley.

Nakayama, M. (2015). *Grammatical variation of pronouns in nineteenth-century English novels* [Doctoral dissertation: The University of Tokyo].

Priestley, J. (1771). *The rudiments of English grammar, adapted to the use of schools; with notes and observations, for the use of those who have made some proficiency in the language.* J. and F. Rivington.

Quinn, H. (2005). *The distribution of pronoun case forms in English.* John Benjamins Publishing Company.

Tieken-Boon van Ostade, I. (1994). Standard and non-standard pronominal usage in English, with special reference to the eighteenth century. In D. Stein & I. Tieken-Boon van Ostade (Eds.), *Towards a Standard of English 1600-1800* (pp. 217–242). Mouton de Gruyter.

Tieken-Boon van Ostade, I. (2009). *An introduction to late modern English.* Edinburgh University Press.

Wales, K. (1996). *Personal pronouns in present-day English.* Cambridge University Press.
Webster, N. (1804). *A grammatical institute of the English language: Comprising an easy, concise, and systematic method of education, designed for the use of English schools in America. In three parts. Part II.* E. Duyckinck.

Presence and Absence of Laughter and Gestures. Examples from the BNC-Spoken 2014 and Dickens' Novels

Michael T.L. Pace-Sigge, University of Eastern Finland

This paper is concerned with non-verbal discourse markers in different types of conversation. Partington (2014) highlights that corpus-linguistics appears to be primarily concerned with presence; this is clearly manifested by John Sinclair's dictum of a "corpus-driven" approach to language studies which "trusts the text". When critically reviewing different forms of discourse this might, however, lead to the researchers turning a blind eye to the context and co-text within which these exchanges take place.

Two case studies will present the presence - or lack of it - of two extra-lingual features which provide crucial co-text markers that provide relevant pointers towards the context in which they occur in two separate corpora. Multi-modal research in linguistics, cognitive studies, and psychological studies have shown that, in spoken discourse, two key markers tend to be present in most situations: one is the use of laughter; the second is that of how gestures are employed to underscore and clarify what is being verbally transmittEd. In this, the chapter revisits work done by Alan Partington (2006), who looked at 'laughter in corpora'; for this case study, the corpus in question will be the most recent comprehensive corpus of casual spoken British English, namely the BNC2014 Spoken. The second case study looks at natural use of gesture in conversations as described by Gullberg (2006). For this case study, references to gesture during conversations will be observed in a corpus of the works of Charles Dickens, which presents an example of highly descriptive fiction.

In order to investigate usage patterns, a British casual spoken corpus, in particular, the most up-to-date, comprehensive and UK-wide spoken corpus, the BNC2014 is investigatEd. The BNC2014 has the occurrence of laughter tagged in XML format files and, as a consequence, the concordance lines where these tags occur are being calculated, using Wordsmith 8 (Scott, 2023). The second step is to review the concordance lines and expunge those where it is used as a lexical item (for example "he laughed his head off"). As a third step, near-collocates of the utterance marker are determined; in parallel, as a fourth step efforts are made to determine where laugh*, appears within an utterance: to se whether it is turn-initial, turn-final or mid-turn. For the second case-study, Mahlberg's Dickens corpus (DCorp: Mahlberg et. al, 2013) – which consists of 23 files, and a total number of 4,533,640 tokens - will be used in order to find word clusters and *co-text* for the target words movement, *gestures, point, hands* and *fingers*.

Overall, the research findings show that a well-constructed corpus can deliver a wealth of information of when and how laughter is employed by speakers even if the phonological qualities of these vocalisations are not available. This highlights the fact that, even with the absence of audio, a well-designed corpus can still provide a wealth of extra-lexical information which can form the basis of research.

References

British National Corpus (2014). *User Manual and Reference Guide, Version 1.1.* (BNC2014). http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf (last accessed 21/10/2023).

Gullberg, M. (2006). Handling discourse: Gestures, reference tracking, and communication strategies in early L2. *Language learning, 56*(1), 155-196.

Mahlberg, M., Smith, C., & Preston, S. (2013). Phrases in literary contexts: Patterns and distributions of suspensions in Dickens's novels. *International Journal of Corpus Linguistics, 18*(1), 35-56.

Partington, A. (2006). *The linguistics of laughter: A corpus-assisted study of laughter-talk.* Routledge.

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

Scott, M. (2023). *WordSmith Tools version 8.* Stroud: Lexical Analysis Software.

Reescrituras telecinemáticas: las transformaciones textuales de la novela Patria

Luisa Chierichetti, University of Bergamo Studies

La presente propuesta se adentra en una línea de análisis lingüístico de las adaptaciones fílmicas de textos literarios que es complementaria a los trabajos enfocados desde un punto de vista narratológico y, por otro lado, aspira a proporcionar elementos para una dimensión didáctica de la escritura de adaptaciones, siguiendo el camino marcado por Taranilla (2021). Se trata de una perspectiva poco explorada en el ámbito de los estudios de estas prácticas, para las que se ha argumentado elocuentemente la conveniencia de preferir el término de "reescritura" al de "adaptación", también para superar cierta idea parasitaria respecto del texto literario y, en general, el paso de una cultura "más alta" a una más popular (Pérez Bowie, 2010, p. 22).

Nuestra investigación se centra en la reescritura que el guionista Aitor Gabilondo realizó de la novela éxito de ventas *Patria* (2016) de Fernando Aramburu, con la homónima serie original producida por HBO España (2020) y dirigida por Félix Viscarret y Óscar Pedraza.

En un trabajo anterior (en prensa) examinamos los principales recursos llevados a cabo en el guion por lo que concierne a la apropiación y reelaboración de la estructura espaciotemporal de la novela, así como a la incorporación de eventos narrativos. Nuestro propósito es ahora el de llevar a cabo un estudio comparativo basado en corpus que considere también el producto final, la serie televisiva, examinando qué tipo de cambios experimenta el guion en la última etapa de la adaptación, en la que los actores finalmente convierten el diálogo escrito en discurso hablado.

Abordamos, pues, nuestro análisis partiendo de un corpus electrónico que contiene los 125 capítulos de la novela, los ocho capítulos de los guiones de la serie, y la transcripción de los ocho episodios de la serie con un total de 370 264 tokens y 301 648 palabras. A continuación, se indican los datos cuantitativos de los tres subcorpus que creamos y manejamos con el software de análisis textual Sketch Engine (http://www.sketchengine.eu):

Subcorpus	*Tokens*	*Palabras*	*%*
Novela	227.066	~186.017	61,3
Guion	107.357	~87.949	29
Serie	35.841	~29.199	9,7

Las herramientas de la lingüística de corpus nos permitirán centrarnos en los diálogos de los personajes en el producto seriado, enfocándonos en el léxico y en la organización sintáctica y ejemplificando los principales procedimientos que Rauma (2004) identifica con estas etiquetas: resumen, elaboración, expansión, reformulación, reubicación, reasignación, rectificación, eliminación, invento y transferencia directa. El examen de la adaptación del diálogo llevada a cabo primero en el guion y luego en el producto televisivo nos permitirá establecer con una sólida base cuantitativa ciertas propiedades de la narración oral definitiva que tiene lugar en la pantalla en comparación con la escritura del guion que, a su vez, se basa en la novela. Este acercamiento al discurso específico de la producción de cine y televisión creemos pueda contribuir a una definición más precisa de la escritura cinematográfica y, a la vez, ofrecer herramientas y ejemplos para la enseñanza de recursos lingüísticos en las instituciones de formación cinematográfica.

Referencias

Aramburu, F. (2016). *Patria*. Tusquets.

Bednarek, M. (2011). Expressivity and televisual characterization. *Language & Literature, 20*(1), 1-19.

Bednarek, M. (2015). An overview of the linguistics of screenwriting and its interdisciplinary connections, with special focus on dialogue in episodic television. *Journal of Screenwriting, 6*(2), 221-38.

Bednarek, M. (2018). *Language and television series. A linguistic approach to TV dialogue*. Cambridge University Press.

Bianchi, F. y Gesuato, S. (2020). Pride and prejudice on the page and on the screen: Literary narrative, literary dialogue and film dialogue. *Nordic Journal of English Studies, 19*(2), 166-98.
Hoffmann, C. y Kirner-Ludwig, M. (Eds.). *Telecinematic stylistics*. Bloomsbury Academic.
Mancilla, J. (2013). Acercamiento al problema de la adaptación cinematográfica de textos literarios: La transposición. *Logos. Revista de Lingüística, Filosofía y Literatura*, *23*(1), 32- 44.
McQueen, S. (2012). Adapting to language Anthony Burgess's and Stanley Kubrick's A Clockwork Orange. *Science Fiction Film and Television, 5*(2), 221-241.
Mahlberg, M. (2013). *Dickens and corpus stylistics*. Routledge.
Paget, D. (1999). Speaking out: The transformation of Trainspotting. En D. Cartmell & I. Whelehan (Eds.), *Adaptations: From Text to Screen, Screen to Text* (pp. 128-140). Routledge.
Pérez Bowie, J. A. (2010). Sobre reescritura y nociones conexas. Un estado de la cuestión. En J. A. Pérez Bowie (Ed.), *Reescrituras fílmicas: Nuevos territorios de la adaptación* (pp. 21-43). Universidad de Salamanca.
Piazza, R., Bednarek, M. y Rossi, F. (Eds.). (2011). *Telecinematic discourse: Approaches to the language of film and television series.* John Benjamins.
Rauma, S. (2004). *Cinematic dialogue, literary dialogue, and the art of adaptation: Dialogue metamorphosis in the film adaptation of The Green Mile.* [Tesis doctoral: Universidad de Jyväskylä]. http://urn.fi/URN:NBN:fi:jyu-2004949558
Sánchez-Escalonilla, A. (2016). *Del guion a la pantalla: Lenguaje visual para guionistas y directores de cine*. Ariel.
Taranilla, R. (2021). Transformaciones textuales en la reescritura cinematográfica de obras literarias: El caso de Zama. *Cuadernos Aispi, 18*, 179-200.
Torres, A. (2015). Literary film adaptation for screen production: The analysis of style adaptation in the film Naked Lunch from a quantitative and descriptive perspective. *Logos: Revista de Lingüística, Filosofía y Literatura,25*(2), 154-64.

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

Łukasz Grabowski, University of Opole

Spanish EFL learners' use of contrastive paratactic expressions across three CEFR levels and gender influence

Carmen Maíz-Arévalo, Complutense University of Madrid

The Common European Framework of Reference for Languages (2001) and its Companion Volume (2018) emphasize the importance of linking expressions for pragmatic competence. In the case of L1 Spanish EFL learners, the adverbial expression of contrast has proven to be particularly difficult, with learners both transferring from their L1 and overusing the same devices (Larsson et al., 2020; Navarro Gil and Roquet Pugès, 2020; Pérez-Paredes et al., 2012; among others). However, these studies have focused on the analysis of different cohesive devices at a specific level rather than contrasting the use of such devices across different levels and in connection with the sociological variable of gender. The aim of this

paper is to redress this imbalance by analysing the use of six specific contrastive connectors; namely, “however”, “on the other hand”, “nevertheless”, “nonetheless”, “yet” and “still” in a learner corpus across the levels B1, B2 and C1. More specifically, this study intends to answer these following research questions: (1) To what extent is the expression of contrast (i.e. by means of adverbials such as “however”) influenced by the learners’ level? More specifically, I will be analysing six contrastive expressions: “however”, “on the other hand”, “nonetheless”, “nevertheless”, “yet” and “still” and (2) to what extent (if any) does the learners’ gender influence their use of concessive expressions? To this purpose, a sub-corpus of the learner-based *FineDesc* corpus has been employed, which includes the written compositions of EFL students who take official accreditation exams to certify their level in different language centres all over Spain. For the present study, a total of 275 opinion essays have been scrutinized with the help of Sketch Engine. These essays belong to levels B1 (82 compositions, 41 by male and 41 by female students), B2 (140 compositions, 70 by male and 70 by female students) and C1 (53 compositions, 25 by female and 28 by male students). Surprisingly, lower-level (B1) users show a wide range of expressions similar to higher-level users, while those at B2 levels tend to avoid "risky" options. Interestingly, gender does not significantly influence learners' use of connectors in this corpus, contradicting earlier findings that suggested female learners use more connectors than males.

References

Larsson, T., Callies, M., Hasselgård, H., Laso, N. J., Van Vuuren, S., Verdaguer, I., & Paquot, M. (2020). Adverb placement in EFL academic writing: Going beyond syntactic transfer. *International Journal of Corpus Linguistics*, *25*(2), 156-185.

Navarro Gil, N., & Roquet Pugès H. (2020). Linking or delinking of ideas?: The use of adversative linking adverbials by advanced EFL learners. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, *33*(2), 505-535.

Pérez-Paredes, P., Sánchez-Tornel M., & Alcaraz Calero J. M. (2012). Learners’ search patterns during corpus-based focus-on-form activities: A study on hands-on concordancing. *International Journal of Corpus Linguistics*, *17*(4), 482-515.

The construction of VAW in the US Press: A corpus-informed discursive news values analysis

Miguel Fuster-Márquez, University of València

‘What is not named does not exist,’ a Whorfian statement often attributed to Steiner and often found in feminist literature addressing Gender-Based Violence or advocating for non-sexist language use. An examination of news stories from the corpus compiled by the NEWSGEN team at the University of València on Violence Against Women (VAW) reveals the power of lexical and phraseological choices in shaping the social and cultural news context of this kind of violence as portrayed by journalists. Notable recent contributions on this topic by NEWSGEN members include Santaemilia (2021), Maruenda-Bataller (2021), Author (2022), and Gregori-Signes (2022). The primary objective of this paper is to focus on

key terms (words and phrases) as pointers or linguistic indicators of news values (NVs), which determine what is newsworthy (Bednarek & Caple 2014, p. 136), in American news stories published by three highly influential quality newspapers, *The New York Times*, *The Washington Post*, and *The Boston Globe*. The three newspapers enjoy a high reputation and are considered to be ideologically center left. Therefore, ideological differences between these newspapers are not a relevant issue. Following Bednarek and Caple (2014, p. 135), we argue that NVs "exist in and are constructed through discourse", they do not exist outside language. The study digs into the most recent period in our corpus data, spanning from 2015 to 2020, employing a Corpus-assisted Discourse methodology (see Partington et al., 2013), to investigate journalistic aspects of interest through the lens of Discursive News Values Analysis (see Potts et al., 2015; Bednarek & Caple, 2015, 2017; AUTHOR & Gregori-Signes, 2019; AUTHOR, 2022) combined with Critical Discourse and Feminist approaches (Baker & Ellece, 2011, p. 26; Wodak & Meyer, 2008; Lazar, 2018; Motschenbacher, 2018). The journalistic corpus was obtained from FACTIVA, one of the largest news databases in the world. The corpus was uploaded to the Sketch Engine platform, which automatically performed POS tagging, lemmatization, and parsing. The annotated NEWSGEN-VAW corpus allowed us to select this subcorpus, which specifically contains hard and soft news stories. We then generated a list of key words and key multi-words to identify terms that deserved attention as pointers to NVs. After filtering out overlapping (multi)word terms, we closely examined the identified terms using the concordancing tool and, when necessary, referred to longer text fragments in these dailies. The paper then explores how newsworthiness is constructed by American journalists. It is of interest to us to examine the ethical commitment of journalists in reporting on sensitive issues of VAW because they play a key role in shaping Americans' perception, and their linguistic choices can affect how the readership understands and responds to these issues. It will be found that journalistic choices encompass social, cultural, and judicial systems when labeling and describing the representation of violent cases that construct the NV of negativity and impact, as witnessed in terms like *sexual violence, domestic violence, sexual assault, sexual misconduct,* or the presence of social actors that construct the NV of Eliteness or Personalization and Proximity, as shown in *The Patriots, Red Sox, Goodell,* etc. From an ethical and critical approach, it is important to see whether journalists who report on individual VAW cases also point to social implications.

References

Baker, P., & Ellece, S. (2011). *Key terms in discourse analysis*. Continuum.

Bednarek, M., & Caple, H. (2014). Why do news values matter? Towards a new methodological framework for analysing new discourse in critical discourse analysis and beyond. *Discourse & Society, 25*(2), 135-158.

Bednarek, M., & Caple, H. (2017). *The discourse of news values: How news organizations create newsworthiness*. Oxford University Press.

AUTHOR, M., & Gregori-Signes, C. (2019). La construcción discursiva del turismo en la prensa española (verano de 2017). *Discurso & Sociedad, 13*(2), 195-224.

Lazar, M. (2018). Feminist critical discourse analysis. In J. Flowerdew & J. E. Richardson (Eds.), *The Routledge handbook of discourse analysis* (pp. 372-387). Routledge.

Motschenbacher, H. (2018). Sexuality in critical discourse studies. In J. Flowerdew & J. E. Richardson (Eds.), *The Routledge handbook of discourse analysis* (pp. 388-402). Routledge.

Maruenda-Bataller, S. (2021). The role of news values in the discursive construction of the female victim in media outlets: A comparative study. In M. AUTHOR, J. Santaemilia, C.

Gregori-Signes, & P. Rodríguez Abruñeiras (Eds.), *Exploring discourse and ideology through corpora* (pp. 141-166). Peter Lang.
Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies*. John Benjamins.
Potts, A., Bednarek, M., & Caple, H. (2015). How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. *Discourse and Communication, 9*(2), 149-172.
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). *RiCL, 9*(2), 90-113.
Santaemilia, J., & Maruenda, S. (2013). Naming practices and negotiation of meaning: A corpus-based analysis of Spanish and English newspaper discourse. In I. Kecskes & J. Romero Trillo (Eds.), *Research trends in intercultural pragmatics* (pp. 439-457). De Gruyter Mouton.

The impact of MT as a writing tool on EFL Academic Writing: a qualitative linguistic analysis

Natalia Judith Laso Martín, University of Barcelona

Elisabet Comelles Pujadas, University of Barcelona

English has definitely become the lingua franca of scholarly communication and, most specifically, the international language of research dissemination. This has been boosted by Internet, where most of the research-related websites (e.g. online journals, institutional repositories, etc.) favour the use of this language.

The above underscores a significant hurdle that EFL/EAL English researchers have to navigate, as they experience increasing pressure to publish their research findings in internationally recognized peer-reviewed journals. As previously highlighted in existing scholarly works (Cargill & Burgess, 2008; Villagrán & Harris, 2009; Lillis & Curry, 2010; Cargill & O'Connor, 2013; Burgess et al., 2019), it is essential for the scientific community to share their outcomes with fellow professionals within their respective domains. Thus, they make use of different strategies and tools to improve their written English (Godwin-Jones, 2022), such as online dictionaries, grammar checkers, online writing aids and Machine Translation (MT). Although the latter was once regarded as a poor resource, producing low quality texts, the landscape has shifted with the emergence of MT engines based on Neural Networks (NN). Nowadays, the quality of Machine Translated (MTed) texts has markedly improved, and this technology has become a widely accepted resource among the general public, as well as within academic contexts (Lee, 2020). In this area, research has shown an improvement from phrase-based approaches (Groves & Mundt, 2015) to NMT systems, especially as regards vocabulary (Kol et al., 2018). Consequently, this development raises a compelling question: Can MT be a useful writing aid for the crafting of academic texts?

This paper explores if the use of MT can help EFL Spanish researchers write their papers in English. To this aim, we recruited 16 Spanish researchers and asked them to write an

abstract directly in English and another one in Spanish. The Spanish abstract was then MTed and post-edited by the researchers themselves. Next, a professional reviser was asked to revise both texts to make them suitable for publication. Finally, we compared the revised versions to calculate the number of changes performed by the professional reviser and analyse the typology of changes carried out.

The results obtained show that the combination of MT and self-postedition has had a relatively positive impact on the academic texts produced by the EFL Spanish researchers. The texts produced with the aid of MT show fewer accuracy issues, although there is still room for improvement to adjust texts to English academic writing conventions.

This study, although exploratory, provides valuable insights on those linguistic areas that researchers using MT must critically evaluate and refine to meet academic standards. Consequently, the findings of this study could serve as the basis of a series of targeted workshops addressed to academics and researchers willing to use MT as a writing aid. The outcomes of this research may be used to raise awareness of MT-related issues and the necessary changes required to adapt an MTed text to the conventions of academic writing.

References

Burgess, S., Martín, P., & Balasanyan, D. (2019). English or Spanish for research publication purposes? Reflections on a critical pragmatic pedagogy. In J. N. Corcoran, K. Englander, & L. Muresan (Eds.), *Pedagogies and policies for publishing research in English. Local initiatives supporting international scholars*. (pp. 128-140). Routledge.

Cargill, M., & Burgess, S. (2008). Introduction to the special issue: English for research publication purposes. *Journal of English for Academic Purposes*, *7*(2), 75-76. DOI: https://DOI.org/10.1016/j.jeap.2008.02.006

Cargill, M., & O'Connor, P. (2013). *Writing scientific research articles*. Oxford.

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology, 26*(2), 5–24. DOI: http://DOI.org/10125/73474

Groves, M. & Mundt, K. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes*, *37,* 112-121, DOI: https://DOI.org/10.1016/j.esp.2014.09.001.

Kol, S., Schcolnik, M., Spector-Cohen, E. (2018). Google translate in academic writing courses? *The Eurocall Review, 26*(2).

Lee, S-M. (2020). The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning,* *33*(3), 157-175. DOI: https://DOI.org/10.1080/09588221.2018.1553186

Lillis, T., & Curry, M. L. (2010). *Academic writing in a global context: The politics and practices of publishing in English*. Routledge.

Villagran, T., Harris, D., & Paul, R. (2009). Some key factors in medical writing. *Revista Chilena de Pediatría,* *80*(1), 70-78. http://www.revistachilenadepediatria.cl/index.php/rchped/article/view/2539

The relation between metaphor and evaluation revisited

Radoslava Trnavac, The National Research University "Higher School of Economics“

Kattie Patterson, University of Granada

Metaphor’s ability to perform a number of discourse functions accounts for their pervasiveness and prevalence in everyday communication. In public discourse, they are used to persuade, to influence, and to frame ideas, and, as a consequence, are able to shape the theory and practice of world affairs (Charteris-Black, 2018). In the political sphere, they provide a narrative structure through which facts are analyzed or challenged, assumptions are made and theories are formulated (Marks, 2018). As they make their way into the media, these cognitive frames of reference act as the glasses through which facts and theories are understood by the public.

In the linguistic research, it has been widely suggested that metaphor performs some sort of evaluative function, with which authors typically express their opinions or sentiments. Drawing on a previous work that described the relation between metaphors and evaluation in movie reviews (Fuoli et al., 2021), we investigate the extent to which evaluation is performed by metaphor in broadsheets and tabloids related to the news discourse on the AUKUS alliance, a trilateral security pact between Australia, the United Kingdom, and the United States. We analyze the following research questions: (1) What is the extent to which evaluation is performed by metaphor in broadsheets and tabloids in the news discourse on Aukus? (2) Do these two types of newspapers differ in terms what type of a metaphor is used to express evaluations? (3) Is there a relationship between the type of metaphors used and the polarity/explicitness of the evaluation in broadsheets and tabloids? (4) Can metaphors be associated with certain parameters of evaluation in the news discourse of broadsheets and tabloids? (5) How does the frequency of different source domains change with the genre of a newspaper? The results of this research have implications for sentiment analysis tasks and automated identification of metaphors.

Our newspaper corpus compiles a collection of news articles from two broadsheets and two tabloids from Britain and Australia for the period of 15 September 2021 until 31 October 2021 with a total number of 16.000 words for each sub-corpus. Drawing on a previous work of Marks (2018) on the relevance of metaphor to International Relations Theory, we apply the following two methodologies to analyze the data: a parameter-based theory of evaluation (Bednarek, 2006), and a recent protocol for the annotation of metaphors found in Fuoli et al. (2021).

The preliminary analysis shows that in both broadsheets and tabloids the majority of metaphors perform evaluative function, and mostly negative. Contrary to the findings of Fuoli et al. (2021), creative metaphors are not more likely than conventional to perform evaluation in the newspaper discourse. In addition, contrary to the results found in Tan (2023), conventional and novel metaphors do not find significant associations with the genre of newspapers. Metaphors are more frequently associated with opinions rather than emotions

in both broadsheets and tabloids, while emotivity expressed by them is significantly more frequent in tabloids, thus confirming a general tendency outlined in Bednarek (2006) that tabloid newspapers are characterized by their preference for emotivity of evaluations. Most of the source domains of metaphors in the broadsheets and tabloids achieve significant associations with the genre of a newspaper.

References

Bednarek, M. (2006). *Evaluation in media discourse: Analysis of a newspaper corpus.* Continuum.

Charteris-Black, J. (2018). *Analysing political speeches: Rhetoric, discourse and metaphor*. Bloomsbury Publishing.

Fuoli, M., Littlemore, J., & Turner, S. (2021). Sunken ships and screaming banshees: metaphor and evaluation in film reviews. *English Language & Linguistics, 26*(1), 75-103.

Marks, M. P. (2018). *Revisiting metaphors in international relations theory.* Palgrave Macmillan.

Tan, X. (2023). *Static and dynamic metaphoricity in U.S-China trade discourse: A transisciplinary perspective.* Vrije Universiteit Amsterdam.

The translation of “yet” as an adverb and conjunction: English language and translation teaching and learning through an English-Spanish parallel corpus

Sidoní López Pérez, International University of La Rioja

Both corpus-based and corpus-driven approaches and research have revolutionized linguistic disciplines since the 1990s (Doval Reixa & Sánchez Nieto, 2019). At first, corpora projects were mainly monolingual; however, parallel corpora started to emerge soon after “as a distinct field of research within corpus linguistics, itself a fairly young discipline” (Borin, 2002, p. 1). Parallel corpora consist of source texts and their translations into one or more languages (McEnery & Xiao, 2007), and they have been largely used in linguistics and translation studies (Yahya, Alotaibi & El- Dakhs, 2020). In addition, parallel corpora are also being used in foreign language and translation teaching and learning as they provide multiple translation suggestions through different examples of real use (Doval Reixa & Sánchez Nieto, 2019). Several studies have focused on the translation of different linguistic elements or expressions using parallel corpora, but their implications for teaching and learning a foreign language have not been fully explored yet (McEnery & Xiao, 2007).

This study focuses on the translations of “yet” in the English-Spanish parallel corpus PaEns. “Yet” is a word that is commonly used in English that has different functions. It can be used both as an adverb and a conjunction. As an adverb, it can be used with affirmative and negative statements, questions, and superlatives. As a conjunction, “yet” is used to show contrast, and it generally occurs after “and”. In addition, “yet” can also be used for emphasis

with a meaning similar to “even”. All these uses imply different translations of the English term into Spanish that our students should know about and be aware of. To this end, this study aims to provide different translation suggestions available in Spanish for the various functions of “yet” using the English-Spanish Parallel Corpus PaEnS. This corpus is a part of the Parallel Corpora Spanish project (PaCorES) that consists of multiple original texts in German, English, and French, and their Spanish translations. PaEnS is composed of two major components: the core corpus and the supplements. This research is based on the core corpus, which consists of 148 original texts in English and Spanish and their corresponding translations. It contains nearly 37 million tokens, and more than one million bisegments. In this case, two novels by Joanne K. Rowling have been selected: *Harry Potter and the Philosopher's Stone* (1997) and *Harry Potter and the Chamber of Secrets* (1998). Results indicate that “yet” appears in 62 examples, in which 21 are negative sentences, and “yet” is translated as “todavía” and “aún” in most of the cases; 18 are sentences in which “yet” is used as a conjunction, and translated as “sin embargo” and “pero” in most of the cases; 15 are examples in which “yet” is used for emphasis, and it is mostly translated as “otro/a” and “nuevamente”; 7 are questions, and “yet” is translated as “ya” in the majority of the examples; and, finally 1 sentence in which “yet” is used with a superlative and translated as “nunca”. These examples can help our students know more specifically about the different functions of “yet” in original texts, whilst also providing them with different translation options that can be useful to both English language and translation teaching and learning.

Acknowledgment

This study has been carried out in the framework of the research project PaCorES: *Online Spanish Parallel Corpora. A multifunctional tool for translation, language learning, and linguistic research* (PID2021-125313OB-I00, Head: Irene Doval) funded by the State Research Agency (AEI) of the Spanish Ministry of Science, Innovation and Universities. Since 2023, the Xunta de Galicia (2023-PG057-1, GI-1954) has also collaborated in the funding of the project. In addition, this study has also been partially carried out in the framework of the research projects VIEALI-Rioja (B036, Universidad Internacional de La Rioja, Spain, 2022-2024), and ANCORABI (Ref. PY20_01365, Retos–Junta de Andalucía, University of Málaga, Spain).

References

Borin, L. (2002). ...and never the twain shall meet? In Lars Borin (Ed.), *Parallel corpora, parallel worlds: Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999* (pp. 1–43). Rodopi.

Doval Reixa, I., & Sánchez Nieto, M. T. (2019). Parallel corpora in focus: An account of current achievements and challenges. In I. Doval Reixa & M. T. Sánchez Nieto (Eds.), *Parallel corpora for contrastive and translation studies: New resources and applications* (pp. 19-38). John Benjamins Publishing Company. DOI: https://DOI.org/10.1075/scl.90.01dov

McEnery, A., & Xiao, R. (2007). Parallel and comparable corpora: What is happening. In G. James & G. Anderman (Eds.), *Incorporating corpora. The linguist and the translator* (pp. 18–31). Multilingual Matters.

Yahya, N., Alotaibi, H., & El-Dakhs, D. A. S., (2020). Parallel corpora in EFL writing classrooms: Are they effective? *International Journal of Computer-Assisted Language Learning and Teaching, 10*(2), 23–39. DOI: https://DOI.org/10.4018/IJCALLT.2020040102

The translation of English derivational adjective compounds in Romance languages: a corpus-based case study

Raluca Nita, University of Poitiers

Ramón Martí Solano, University of Limoges

This study deals with the translation in Romance languages of English morphologically complex adjectives, formed simultaneously by compounding and derivation—Adjective/Noun1+$Noun_2$-*ed*, described as "derivational adjective compounds" (Adams, 1973, p. 99-101), "adjectival compounds that involve derived adjectives as heads" (Plag, 2003, p. 153), or *-ed* adjectival compounds (Labrador de la Cruz & Ramón García, 2010). Semantically, the $Noun_2$-*ed* refers mainly to body parts (Adams, 1973; Quirk et al., 1985; Tournier, 1991), *dark-haired, single-minded*, but also to any other elements constitutive of an object, *long-stemmed flower*, *oak-panelled wall*, and thus appears in a metonymical relation (part-whole) with the noun it qualifies (*stem – flower*). Generally, most of these denominal *-ed* adjectives (Schuwer, 1998)—to be distinguished from adjectives derived from past participles with inflectional *-ed*—cannot function on their own as they express an inherent property of the base they qualify (Hirtle, 1970), **an eyed/a haired man*. It is the modifier in the compound (adjective or noun) that transforms a stable, inherent property into an occasional, context-dependent one, as in a *blue-eyed, black-haired man.*

These adjectival compounds (AdjC) prove particularly interesting for a contrastive analysis with Romance languages, which cannot form such compounds and have, on the whole, different morphological features compared to English (Chuquet & Paillard, 2007, Paillard, 2011). For instance, in French the nominal derivational suffix in past participles (*-u*) is considerably less productive than the English derivational *-ed* (Lehmann & Martin-Berthet, 1998; Paillard, 2000).

The contrastive approach has not yet been applied to *-ed* AdjC. Translations into French, Spanish and Romanian will point out the semantic complexity and discourse features of the

synthetic English compounds when compared with their analytical equivalents. We combine a multilingual literary corpus—in English, French, Spanish and Romanian, of about 550,000 words—with translations from various registers, manually collectEd. Indeed, despite the productive potential of the *-ed* AdjC shown in the literature (Adams, 1973; Hirtle, 1970; Quirk *et al.*, 1985) and their expected adequacy to fiction as a genre involving character descriptions, our multilingual literary corpus provides only 35 hits of this subtype. This sample is nevertheless relevant to conduct an exploratory survey of Romance equivalents although further examples are deemed necessary as they add new morphological and semantic features to our initial corpus.

The 112 occurrences of our current corpus are analysed according to the morphological, syntactic and semantic features of their translational equivalents. The three Romance languages share two dominant equivalents for Adj/N_1+N_2-*ed* qualifying a noun (Nq): N_2 Adj+Nq (more than 60%), Adj+Nq (around 20%). In the first case, the semantic relation of possession between the N2 and the qualified noun Nq in English is made explicit either by prepositions locating the N2 relative to the Nq, a *keen-eyed concierge* is translated as *un concierge au regard acéré, un conserje de mirada aguda, un administrator cu ochi ageri*, or sometimes, as in French, by the verb *avoir* (have), as in We are *big-hearted > Nous avons un grand cœur*. In the second case, the N_2 is omitted and an adjective, either the equivalent of the English adjective or an adjective derived by affixation from the N_2, modifies the Nq, as in *big-boned face > son visage osseux*.

The remaining translations involve other types of grammatical changes (*-ed* AdjC transformed into a verb phrase or an adverb, *literal-minded: il ne faut pas prendre les choses au pied de la lettre*) and/or changes of perspective ('modulation', according to Vinay & Darbelnet 1977) triggered mainly by the metaphorical nature of the compound (*loud-mouthed > au verbe haut*). The corpus shows that the metaphorical features (Tournier, 2004, p. 93) are reinforced by the collocation with abstract nouns, by the metaphorical use of the Adj or N_1 and by the emotional, rather than physical features attached to the N2: *mean-minded resentfulness > rancœur mesquine, resentimiento mezquino, resentiment răutăcios; hare-brained conclusions > conclusions farfelues, conclusiones descabelladas, concluzii nechibzuite.*

Translations point out the variety of meanings associated with *-ed* AdjC according to the type of nouns they qualify (*hard-nosed approach > une approche intransigeante, énergétique*), their syntactic position, and the register features of the text (*open-mouthed > bouche bée, à embouchures ouvertes*). Metaphorical meanings are in line with the polysemy of -ed AdjC and may be difficult to grasp as shown by discrepancies of translations between languages, as with *light-headed*, meaning either 'dizzy' or 'frivolous'.

Romance languages may occasionally compensate the lack of morphological creativity by means of other stylistic effects, thus displaying specific pragmatic features—French, for instance, tends to use more often its equivalents as foregrounded appositions, located at the

left of the nucleus clause (Huddleston & Pullum, 2002, p. 1408), contrary to the attributive, predicative or standard appositive position found in English.

References

Chuquet, H., & Paillard, M. (2007). Les adjectifs composés en X + V-ing : prédication, collocations, traduction. *Palimpsestes, 19*, 13-34. DOI: https://DOI.org/10.4000/palimpsestes.113

Hirtle, W. H. (1970). -Ed adjectives like 'verandahed' and 'blue-eyed'. *Journal of Linguistics, 6*(1), 19-36. DOI: https://DOI.org/10.1017/S0022226700002334

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language.* Cambridge University Press. DOI: https://DOI.org/10.1017/9781316423530

Labrador de la Cruz, M. B., & Ramón García, N. (2010). Bollywood-loving countries and Bollywood-inspired films: English compound adjectives in Spanish translations. In R. Rabadán, T. Guzmán & M. Fernández (Eds.), *Lengua, traduccion, recepcion: En honor de Julio Cesar Santoyo/ Language, translation, reception: To honour Julio Cesar Santoyo* (pp. 247-270). Universidad de León.

Lehmann, A., & Martin-Berthet F. (1998). *Introduction à la lexicologie. Sémantique et morphologie.* Dunod.

Paillard, M. (2000). *Lexicologie contrastive anglais-français - Formation des mots et construction du sens.* Ophrys.

Paillard, M. (2011). English–French contrasts in word-formation. Morphological patterns and stylistic effects. *Poznań Studies in Contemporary Linguistics, 47*(4), 913-927. DOI: https://DOI.org/10.2478/psicl-2011-0043

Plag, I. (2003). *Word-formation in English.* Cambridge University Press.

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

Schuwer, M. (1998). Étude sur les contraintes syntaxiques des adjectifs en -ed en anglais. *Cahier du CIEL 1999,* 110-151.

Tournier, J. (1991). *Structures lexicales de l'anglais.* Nathan.

Tournier, J. (2004). *Précis de lexicologie anglaise.* Ellipses.

Vinay, J-P., & Darbelnet, J. (1977). *Stylistique comparée du français et de l'anglais.* Didier.

Unveiling connotative reasons for term variations within a vertically stratified specialized corpus

Rossella Resi, University of Innsbruck

This work investigates term variations in a German corpus of specialized texts produced within the same institution (ift Rosenheim) in the door and window industry. A clear horizontal classification of the domain (von Hahn, 1983; Roelcke, 2014) made it possible to focus on connotative explanations for term variations occurring throughout a vertical stratification of specialized language (Ischreyt, 1965; Hoffmann, 1985; Roelcke, 2014).

The starting point for the analysis was the standardized terminological database (Theoriesprache1) of the institute with its corpus of normative texts (Resi, 2022). A secondary corpus was then defined, consisting of non-normative *Verteilersprache* (V-texts) and *Werkstattssprache* (W-texts) produced within the same institution. These included magazine articles from the ift-Akademie (V 1-14), reports on LinkedIn (V 15-27), Newsletters (V 28-35), reports of on-site ift-audits (W 1-9) and verbal interactions during ift-audits (W 10-

19)$_2$. Each text-related communicative context, was annotated with seven selected external objective criteria$_3$ to establish whether the causes for connotative variation might lie in the presence or absence of distinct factors.

The secondary corpus was searched for alternative designations (variaciön denominative, Freixa, 2006) of the original terms (in the sense of Daille, 2017) as well as polysemous variations (Pelletier, 2012) with analysis of any resulting implications.

The variant Nullschwelle4 of the original term barrierenfreie Schwelle$_5$, for example, was found in non-spontaneous and non-anonymous texts (V5,18,19,32,35). There were heated discussions in 2018 on the need for a low (not zero) wheelchair-friendly door threshold profile for insulation purposes, as opposed to the idea that door thresholds always constitute barriers. Subsequently, the usage of Nullschwelle in V-texts became an implicit declaration of intent and an expression of personal opinion. This dynamic also emerged from a proliferation of term clusters (amongst others niveaugleiche Schwelle$_6$) in an effort to avoid ideological implications. However Nullschwelle, preserved the mere referential meaning of “zero-height of the threshold profile”in W-texts. This is not surprising, considering that the same polysemy occurs for the economic variant of Schwelle$_7$, which substitutes the standardized term Schwellenprofil8 in V and W texts.

The standardized term Schwelle refers only to the measurement of height from the ground. It was also observed that the inclination to adopt more economic expressions for transparent but overly long original terms, sometimes also concealed an implicit persuasive component. An example of this is Compriband (a variant of the original term Fugendichtungsband), a product name chosen by a well-known company. In immediate non-public operative texts (W 10-18), the use of the proper noun is only a matter of brevity (testified by the fact that installers used this name even when they were fitting a different product brand), while in public written texts the choice is based on personal preference. Most W-texts were characterized by operativity, which encouraged directness over and above individuality and personal choice, despite the oral and non-anonymous nature of these texts.

The contribution will demonstrate on the basis of empirical data that certain features of communicative contexts can encourage reformulations of original terms with a more or less intentional addition of implied connotations to the conceptual reference. This challenges once again the idea that highly specialized texts, and therefore their terminology, are intrinsically anonymous.

References

Daille B. (2017). *Term variation in specialised corpora: Characterisation, automatic discovery and applications* (Vol. 19). John Benjamins Publishing Company.

Freixa J. (2006). Causes of denominative variation in terminology. A typology proposal. *Terminology, 12*(1), 51-77.

Gülich E., & W. Raible (1975) *Textsorten. Differenzierungskriterien aus linguistischer Sicht.* Frankfurt, Athenäum-Skripten Linguistik; 5. Athenäum-Verlag.

Hoffmann L. (1982). Probleme und Methode der Fachsprachenforschung. In *Wissenschaftliche Zeitschrift der Universität Leipzig.* Gesellschaftlich-Sprachwissenschaftliche Reihe*, 31*(1), 25-34.

Hoffmann L. (1985). *Kommunikationsmittel Fachsprache. Eine Einführung*. Narr.

Ischreyt H. (1965). *Studien zum Verhältnis von Sprache und Technik: Institutionelle Sprachlenkung in der Terminologie der Technik.* Schwann.

Pelletier J. (2012). *La variation terminologique: un modèle à trois composantes.* [Thèse de doctorat en linguistique, Département de langues, linguistique et traduction, Faculté des lettres, Université Laval, Québec].

Resi R. (2022). Concettualizzazione terminologica italiano-tedesco al servizio della traduzione di norme tecniche nel settore serramentistico. In E. Chiocchetti et al. (Eds.), *Risorse e strumenti per l'elaborazione e la diffusione della terminologia in Italia*. Eurac Research, Bolzano.

Roelcke T. (2014). Zur Gliederung von Fachsprache und Fachkommunikation. In *Fachsprache International Journal of Specialized Communication 36*(3–4), pp. 154–178.

Von Hahn W. (1983). Fachsprache. In *Lexikon der Germanistischen Linguistik.,* 2 (pp. 390-395). Aufl. Tübingen.

Using corpora to inform sociological studies of translation

Paolo Canavese, Dublin City University

Corpus linguistics (CL) offers flexible tools and methods that can be applied across a range of social sciences (McEnery & Brezina, 2022; Crosthwaite, 2023; Zinn, 2019; Rubtcova et al., 2017). This paper illustrates how CL has been used to inform SWIFT, an ongoing research project that aims to shed light on the profiles and needs of in-house translators working for the Swiss Confederation. The project is rooted in the field of the "sociology of translation" (Wolf & Fukari, 2007), which means that institutional translation is seen as a social activity, with a focus on sociological concepts such as agents, networks and fields (Buzelin, 2005; Gouanvic, 2005). To this end, it primarily uses a questionnaire and interviews as a method to collect federal translators' ideas and perceptions. At the same time, the analysis of an ad hoc corpus containing 250 job announcements for the recruitment of federal translators published between 2016 and 2023 was extremely useful as an additional datapoint and for triangulation purposes. Building on existing competence models (European Commission, 2022; Hurtado Albir et al., 2023) and previous studies of other institutional contexts (Prieto Ramos & Guzmán, 2022; Lafeber 2022), this study provides insights into the tasks performed by translators, which go far beyond merely translating, and the plethora of competences required.

The corpus underwent thematic analysis in NVivo, using over 80 descriptive codes. The macrocodes were predefined, i.e. drawn from the literature, while finer-grained codes were added, refined, and clustered during the annotation rounds to carefully capture the content of the job announcements. The quantitative data, complemented by a qualitative analysis of the coded segments, made it possible to characterise the profiles of federal translators, and revealed some unexpected trends. For example, it turned out that "personal and interpersonal competences" is the second most common requirement (i.e. it is mentioned in 96.4% of job announcements), after having a tertiary education (98.8%), which raises the question of how, for example, soft skills can be included in the training curricula for translators. With regard to "thematic and cultural competences" (67.2%), the multi-word expression "general knowledge" frequently co-occurs with quantifying adjectives such as "very good" or "excellent"; instead, when referring to the thematic fields covered by the language service, the noun "interest" is more frequently used, signalling that specialised knowledge can be acquired on the job. In terms of "instrumental competence" (75.2%), the most frequently required skills are familiarity with computer-assisted translation tools and general IT skills. Machine translation and post-editing skills are scarce (N=9), which is unexpected in the age of artificial intelligence.

Combining different sources of data – and the strengths and weaknesses thereof – allowed for a comprehensive picture of the phenomena under analysis. On the one hand, the corpus analysis provided a solid basis to better frame the subjective views of the participants. On the other hand, the picture that emerged from the corpus analysis was necessarily partial, as it could not account for factors such as the evolution of profiles over the course of a career,

the frequency of tasks or actual profiles, which might differ from ideal ones. Questionnaires and interviews prove useful to refine this. Ultimately, this paper aims to stimulate the corpus community to engage with other disciplinary approaches in order to provide richer answers to their research questions, or even to formulate broader ones.

References

Buzelin, H. (2005). Unexpected allies. *The Translator, 11*(2), 193-218. DOI: https://DOI.org/10.1080/13556509.2005.10799198

Crosthwaite, P. (2023). Corpus linguistics: Mixed-methods research. In *The Encyclopedia of Applied Linguistics.* Wiley. DOI: https://DOI.org/10.1002/9781405198431.wbeal20019

European Commission. (2022). *European Master's in Translation Competence Framework 2022. European Commission.* https://commission.europa.eu/system/files/2022-11/emt_competence_fwk_2022_en.pdf

Gouanvic, J. M. (2005). A Bourdieusian theory of translation, or the coincidence of practical instances. *The Translator, 11*(2), 147-166. DOI: https://DOI.org/10.1080/13556509.2005.10799196

Hurtado Albir, A., Rodríguez-Inés, P., Prieto Ramos, F., Dam, H. V., Dimitriu, R., Haro Soler, M. M., & Zethsen, K. K. (2023). *Common European Framework of Reference for Translation – Competence Level C (specialist translator): A proposal by the EFFORT Project.* Effort project. https://www.effortproject.eu/

Lafeber, A. (2022). Skills and knowledge required of translators in institutional settings. In T. Svoboda, Ł. Biel, & V. Sosoni (Eds.), *Institutional Translator Training* (Vol. 30-48). Routledge.

McEnery, T., & Brezina, V. (2022). *Fundamental principles of corpus linguistics.* Cambridge University Press.

Prieto Ramos, F., & Guzmán, D. (2022). Institutional translation profiles. A comparative analysis of descriptors and requirements. In T. Svoboda, Ł. Biel & V. Sosoni (Eds.), *Institutional Translator Training* (pp. 49-72). Routledge.

Rubtcova, M., Vasilieva, E., Pavenkov, V., & Pavenkov, O. (2017). Corpus-based conceptualization in sociology: Possibilities and limits. *Espacio Abierto, 26*(2), 187-199. https://www.redalyc.org/journal/122/12252818011/html/

Wolf, M., & Fukari, A. (Eds.). (2007). *Constructing a sociology of translation.* John Benjamins.

Zinn, J. O. (2019). Utilising corpus linguistic tools for analysing social change in risk. In A. Olofsson & J. O. Zinn (Eds.), *Researching risk and uncertainty: Methodologies, methods and research strategies* (pp. 337-366). Palgrave Macmillan.

Vocabulary additions in teacher talk: What kinds of words do language instructors add to the textbook?

Nausica Marcos Miguel, Denison University

Silvia Aguinaga Echeverría, University of Navarra

Oihane Muxika Loitzate, University of the Basque Country

Studies focusing on the ability of instructors to rate target words have shown that instructors are well-equipped to assess words' importance for their learners (He & Godfroid, 2019; Robles-García et al., 2023; Sánchez-Gutiérrez, Robles-García, Hoyos Álvarez, 2023). Moreover, their textbooks are one of their usual resources to select those target words

(Allen, 2008; Sánchez-Gutiérrez, Robles-García, & Pérez Serrano, 2022). Given that instructors adapt their textbooks (McGrath, 2013; Masuhara, 2022), they probably also adapt their textbooks' target words to their contexts. However, internal and external factors influencing instructors' adaptations of target vocabulary within the classroom have barely been addressed.

This case study focuses on three instructors teaching the same second-semester course at a university in the United States. This study examines what kind of words each instructor added to the chapter's glossaries, which comprised 94 words related to body parts, health providers, health, and medical treatments. Specifically, this study explores four factors: frequency of the added words, whether the words were also in the textbook despite not being in the glossary, whether the words belonged to current events, and whether the words were part of the language varieties of the instructors.

The three instructors had different language backgrounds: Instructor A was from Argentina, Instructor B was from Spain, and Instructor C was born in a non-Spanish-speaking country but had spent a year learning in Spain. Each instructor had over ten years of experience teaching Spanish in the United States. Four synchronous online classes of each instructor working on the same chapter were recorded. In those sessions, there were around 20 students with a proficiency level between A1 and A2.

Once the 12 sessions were transcribed, the online platform SketchEngine generated a list of all the words uttered by the instructors. After examining the list, 90 additional words related to the glossary topics were found. 39 of the words belonged to the 3,000 most frequently used words in Spanish, with a total of 52 belonging to the 5,000 most frequently used words. The remaining 38 words were above these frequency ranges. Despite not being in the glossary, 50 appeared in the chapter. Of the remaining 40, nine words belonged to the 3,000 most frequent words, with a total of 25 belonging to the 5,000 most frequent words. The remaining 15 were above those bands. Within the low frequency words, four words dealt with the health situation during the study (COVID, COVID-19, *pandemia*, and coronavirus). No regional variation was found among the instructors. Still, when collocations were examined, clearer variational patterns emerged. For example, Instructor B used collocations from her variety, such as *estar fatal, estar desganado*, and *estar congestionado*.

These findings confirm that instructors use the textbook as a primary source of vocabulary, both for glossary and non-glossary words. When following the textbook, instructors' use of vocabulary remained within the recommended range of frequencies for teaching Spanish at this level. By providing additional words from current world events and/or their own experiences, they contributed to increasing knowledge of lower frequency bands. Sociolinguistic variation was perceived in collocations rather than at the word level. Therefore, this study offers new points for reflection as further factors beyond the four examined here need to be considered to explain the differences between instructors' additions.

References

Allen, H. W. (2008). Textbook materials and foreign language teaching: Perspectives from the classroom. *NECTFL Review*, 62, 5-28.

He, X., & Godfroid, A. (2019). Choosing words to teach: A novel method for vocabulary selection and its practical application. *TESOL Quarterly*, 53(2), 348-371. https://doi.org/10.1002/tesq.483
Masuhara, H. (2022). Approaches to materials adaptation. In J. Norton & H. Buchanan, *The Routledge handbook of materials development for language teaching* (pp. 277-290). Routledge.
McGrath, I. (2013). *Teaching materials and the roles of EFL/ESL teachers: Practice and theory*. A&C Black.
Sánchez-Gutiérrez, C. H., Robles-García, P., & Hoyos Álvarez, C. (2023). Vocabulary in the L2 Spanish classroom: What students know and what their instructors believe they know. In L. Marqués-Pascual & I. Checa (Eds.), *Current Perspectives in Spanish Lexical Development*, 68, 245.
Sánchez-Gutiérrez, C. H., Robles-García, P., & Pérez Serrano, M. (2022). L2 Spanish vocabulary teaching in US universities: Instructors' beliefs and reported practices. *Language Teaching Research*, 0(0). https://doi.org/10.1177/13621688221074443
Robles-García, P., Stewart, J., Nicklin, C., Vitta, J. P., McLean, S., & Kramer, B. (2023). 'The wisdom of crowds': When teacher judgments outperform word-frequency as a predictor of students' vocabulary knowledge. *Language Teaching Research*, 0(0). https://doi.org/10.1177/13621688231176067

What's done can't be undone': Verbal Contractions in Modern English

Marta Pacheco-Franco, University of Málaga

Javier Calle-Martín, University of Málaga

The habit of contracting words originates in the intrinsic tendency of languages to assimilate the pronunciation of two neighbouring sounds. In Present-day English, this phenomenon is most plainly observed in the so-called 'telescoped phrases', a term that refers to all those cases of elision affecting auxiliary verbs, as exemplified by *it's* and *doesn't* (Peters, 2004, p. 126). Among these, there are, on the one hand, the contractions featured in operators, i.e., "the larger class containing the NICE [i.e. negation, inversion, code and emphasis] verbs in *all* their uses" (Huddleston & Pullum, 2002, p. 104, our emphasis). In these cases, the element contracted is a verb, either in its functional or lexical uses, and the elision takes place at the beginning of the word, as in *it's* for *it is* or *they've* for *they have*. On the other hand, the negative adverb not may also be contracted to *n't* when it modifies an operator, thereby losing its mid-word vowel and its free-word status in the process. The examples *doesn't* for *does not* and *isn't* for *is not* illustrate this phenomenon, with the exception of the negative modal *can't*, which stands for the univocal full form *cannot*. Today, contractions are typically employed in colloquial registers, whether they be spoken or written. However, it remains unclear when these structures first became widespread in the language and how they became an index of informal English.

The present paper thus puts forward a corpus-based study on contractions that has been conceived as a register analysis of the phenomenon. The study is then concerned with the standardisation of the contractions in the period 1600-1999 and pursues the following

objectives: a) to study the development of these structures over time, assessing their origin and spread in English; b) to analyse their distribution across registers and gender; and c) to consider the factors which may have participated in the informality of the constructions in Present-day English. For those purposes, the analysis explores the instances of contraction of the auxiliary and lexical verbs *be* and *have* and of the modal auxiliaries will and would, as well as the shortened forms of the negative particle *not* in *A Representative Corpus of Historical English Registers* or *ARCHER* 3.2. The preliminary analysis of the data points to a wider distribution of the phenomenon in colloquial registers since their inceptio, which, more importantly, seems to be a result of the pull exercised by female authors in the overall dissemination of the structure. The considerable rise of the contractions from the 1900s onwards, which in itself was a critical point for the acquisition of civil rights on the part of women, possibly answers to the overall democratisation of the language taking place in the 20th century.

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.llc.manchester.ac.uk/research/projects/archer/

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

Peters, P, (2004). *The Cambridge guide to English usage.* Cambridge University Press.

Words are syntactically distributed for efficient use: Evidence from syntactic neighborhood density

Phillip G. Rogers, University of Pittsburgh

A growing body of research has identified patterns of systematicity within and among features of the lexicon that reflect cognitive and communicative pressures on language acquisition and use (Dingemanse et al., 2005). For example, the relationship between form and meaning has traditionally been considered arbitrary (de Saussure, 1916; Hockett, 1960), yet various cross-linguistic studies have revealed a widespread correlation across the lexicon, beyond what can be accounted for by sound symbolism (Dautriche et al., 2016; Monaghan et al., 2014). These regular correspondences between form and meaning offer advantages for learning and memory (Imai & Kita, 2014; Kirby et al., 2015).

At the same time, recent psycholinguistic research has demonstrated that our knowledge of words includes fine-grained information about the syntactic contexts in which they are likely to participate (Lester, 2018). For example, words with similar syntactic distributions prime each other (Lester et al., 2017). These syntactic distributions are defined as the probability distribution of a word's occurrences in various dependency roles (e.g., nouns may be more or less likely to serve as the subject of a verb or as the head of an adjective modifier—see Figures 1 & 2).

Building on research in phonology and semantics, the current study investigates the relationship between syntactic neighborhood density and frequency. Previous studies have shown that more frequent words have denser phonological neighborhoods (Mahowald et al., 2018). Since shared associative pathways among close phonological neighbors are known to facilitate learning (Storkel, 2004) and production (Stemberger, 2004), these results suggest that the most frequent words in a language are also the most optimized for efficient communication. Semantic neighborhood density is known to exhibit the same correlation with frequency, but to our knowledge the potential cognitive advantages of this pattern have not been explored to the same extent as for phonology. We hypothesize that frequent words will also have denser syntactic neighborhoods, presumably offering similar advantages to learning and production.

We extract syntactic distributional information for lemmas in 48 languages from the Universal Dependencies Treebanks (de Marneffe et al., 2021). Defined as the average distance to a lemma's *n* nearest neighbors, we compute syntactic neighborhood densities for *n* = 3, 10, and 50. We performed two statistical analyses to test our hypothesis. In a correlational analysis, we computed the Pearson correlation between syntactic neighborhood density and frequency for each language, comparing this correlation to a permuted baseline. We found that, for most languages and neighborhood sizes, there is a significant correlation between syntactic neighborhood density and frequency, with more frequent words having denser syntactic neighborhoods. We also performed a mixed-effects regression analysis predicting frequency from syntactic, semantic, and orthographic neighborhood densities for a neighborhood size of 10. The model included a full random effects structure with random intercepts and slopes for languages and subfamilies, and a model selection process was used to determine whether interactions and curvature should be included. The final model is complex (see Figure 3), but it supports the hypothesis that denser syntactic neighborhoods predict higher frequencies.

We interpret this finding as a design feature of language, and it supports the idea that the lexicons are structured for efficient use (Gibson et al., 2019). More broadly, these studies on the syntactic distributions of words present a challenge to traditional assumptions concerning the division between grammar and lexicon (Chomsky, 1995).

References

Chomsky, N. (1995). *The minimalist program.* MIT Press.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. (2016). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science, 41*(8), 2149–2169.
Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2005). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences, 19*(10), 603–615.
Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389–407.
Hockett, C. (1960). The origin of speech. *Scientific American, 203*, 88–96.
Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1651), 20130298.
Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition, 141*, 87–102.
Lester, N. (2018). *The syntactic bits of nouns: How prior syntactic distributions affect comprehension, production, and acquisition* [Doctoral thesis: University of California Santa Barbara].
Lester, N., Feldman, L. B., & Moscoso del Prado Martín, F. (2017). You can take a noun out of syntax...: Syntactic similarity effects in lexical priming. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 2537–2542.
Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Wordforms are structured for efficient use. *Cognitive Science, 42*(8), 3116–3134.
de Marneffe, M.-C., Manning, C., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics, 47*(2), 255–308.
Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language. *Philosophical Transactions of the Royal Society B, 369*(1651), 20130299.
de Saussure, F. (1916). *Course in general linguistics.* McGraw-Hill.
Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language, 90*(1), 413–422.
Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics, 25*(2), 201–221.

Figures

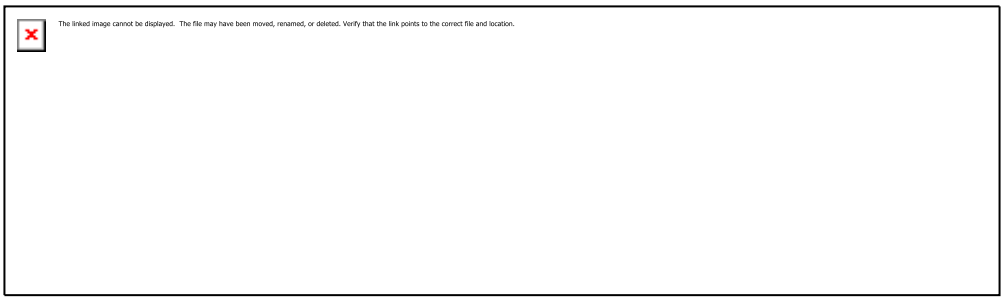


Figure 1: An example of the Universal Dependencies Treebanks dependency framework from Spanish. Syntactic dependencies are represented by arrows pointing from heads to their dependents, and each dependency is labeled for the type of relation. The translation of the sentence is ‘We want to get at least four or five gold medals.’

Figure 2: Partial probability vectors for the participation of three Spanish lemmas in different syntactic roles and relations. The height of each bar indicates how often that lemma participates in that dependency type relative to other syntactic dependency types. The probabilities shown are corrected for sample bias with the James-Stein shrinkage estimator. These three distributions illustrate how *oro* 'gold' is much more similar syntactically to *paz* 'peace' than to *medalla* 'medal', despite being more similar semantically to the latter.

Figure 3: Mixed-effects model predictions for frequency showing interactions between semantic and syntactic neighborhood densities (left) and semantic and orthographic neighborhood densities (right). Densities are measured by average distance to ten nearest neighbors, so low numbers represent denser neighborhoods. Color represents predicted word frequencies. In both interactions, higher frequencies are predicted for denser neighborhoods on both axes.

Written Learner Corpus of Brazilian Portuguese

Shintaro Torigoe, Osaka University

This project, as a follow-up to the *Portuguese Vocabulary Profile* (2015–2017; Torigoe, 2016, 2017), constructs written learner corpora of Brazilian Portuguese and compares them with those of L2 European Portuguese. In a previous study (Torigoe, 2016), the author proposed a wordlist for learners of Portuguese as a foreign language based on two written learner corpora of European Portuguese: the *Corpora do PLE*, constructed by the University of Lisbon, and the *Corpus de PEAPL2*, constructed by the University of Coimbra. In this wordlist, the author proposed a target vocabulary of approximately 500 words for three levels of the *Common European Framework of Reference* (CEFR) –namely, A1-A2, B1-B2, and C1-C2 – based mainly on word frequency. However, the total vocabulary size was half that of Capel's (2010, 2012) English Vocabulary Profile and Tono's (2013) *CEFR-J Wordlist*. Thus, in a further study (Torigoe, 2017), the author refined the wordlist by integrating frequency information from a Portuguese native speaker corpus and range information from a corpus of Portuguese textbooks published in Japan. The author proposed a new vocabulary list of approximately 1500 words for levels A1-A2 and B1-B2, almost equivalent to the lists of Capel and Tono. Nevertheless, issues remained in terms of unexpected exclusion of basic vocabulary due to the small size of the learner corpora. In addition, as *PLE* and *PEAPL2* are based on data from learners of European Portuguese, the wordlist would not necessarily be suitable for an environment in which Brazilian Portuguese is mainly taught and learnt, such as in Japan.

As part of this project, the author collects an L2 Brazilian Portuguese learner corpus and will use it as a starting point to build a new *Portuguese Vocabulary Profile* applicable to environments where Brazilian Portuguese is mainly taught and learnt. The corpus is collected in collaboration with the Federal University of Fluminense and the Federal University of Paraná. The methodology for collecting the corpus followed that used for *PEAPL2*, through L2 Portuguese learners' free writing on one of nine themes. This makes the corpus compatible with *PEAPL2* and *PLE*. As of January 2024, the author constructed a corpus of about 15,000 words from a total of 84 informants with the two universities, comprising native speakers of 16 languages in total. More specifically, 33.3% of the informants were native Spanish speakers, and 29.8% were native French speakers. The native Spanish speakers were mainly from Latin America, and many of the native French speakers were from West Africa and Haiti, which were areas not considerably covered by *PEAPL2* and *PLE*, of which the majority of informants were ERASMUS students.

In a statistical comparison, the Brazilian and European Portuguese learner corpora yielded a very high correlation (r = 95.8) with respect to lexical frequency. Although a chi-square test identified some vocabulary that was produced significantly more by learners of Brazilian Portuguese than by learners of European Portuguese, it seems to be because of the task influence rather than the difference in variants, except for function words and the most frequent vocabulary. Therefore, the author believes that there is currently no difference in terms of vocabulary between the learner corpora of the two variants. Thus, the Brazilian learner corpus will be integrated with *PEAPL2* and *PLE*, as well as the native speaker and textbook corpora, to develop a new *Portuguese Vocabulary Profile.*

References

Capel, A. (2010). A1-B2 Vocabulary: Insights and issues arising from the English profile wordlists project. *English Vocabulary Journal, 1.*

Capel, A. (2012). Completing the English vocabulary profile: C1 and C2 vocabulary. *English Vocabulary Journal*, *3,* 1-14.

Tono, Y. (2013). *CEFR-J handbook, A resource book for using CAN-DO descriptors for English language teaching* (written in Japanese). Taishukan.

Torigoe, S. (2016). Seeking the Portuguese vocabulary profile: Pilot study*. EPiC Series in Language and Linguistics, CILC2016.*

Torigoe, S. (2017). Portuguese vocabulary profile: Uma lista de vocabulário a aprendentes do PL2/PLE, baseada nos corpora de aprendentes e de livros de Ensino. *Revista da Associação Portuguesa de Linguística, 3*(2017), 387-400.

"You'll really appeal to your customers and push out your products." The phrasal verb use in European student business case competitions

Siyang Zhou, Hong Kong University of Science and Technology

Xinyue Zhang, University of Manchester

Formulaic language is an area of vocabulary research that has been gaining traction in recent decades (Wray, 2013). Phrasal verbs (PVs), a useful but difficult type of formulaic language, are two-part verbs consisting of a lexical verb and an adverbial particle (Gardner & Davies, 2007). They widely appear in informal spoken language of English-as-first-language users, yet they pose significant challenges to English-as-second-language learners (Liu, 2011). Therefore, researchers have compiled several corpus-based high-frequency PV lists to inform teaching and learning (Liu, 2011). The language styles of business English vary by genres and contexts, but it was observed that colloquialism seems to be a trend which enables business messages to be delivered in a natural, friendly, and accessible way (Jiang, 2015).

This study investigated PV usage in university students' business presentations, focusing on both the frequency and accuracy of such linguistic structure. The study drew on spoken data from 11 business case presentations delivered by 44 university students participating in the Copenhagen Business School (CBS) Case Competition. A 50,000-word learner corpus was compiled, and AntConc (Anthony, 2023) was used to extract and analyze the PVs. Case competitions were chosen because of the wide applicability of business presentations in real-life scenarios, which can provide important implications for ESP (English for specific purposes) instructors.

The mixed-methods analysis of the data yielded several noteworthy findings. Firstly, quantitative analysis revealed a strikingly high accuracy in the production of PVs by university students, as well as a diverse range of PV types, showcasing students' high language proficiency. Secondly, the qualitative analysis found that students exhibited a propensity for employing PVs inventively or creating new PVs and extending meanings that were not found in dictionaries. While half of the PVs identified in this study aligned with those in Liu's (2011) most frequent PV list, the other half of the PVs not on the list further signified that students incorporated less common but more distinctive expressions. This result highlighted students' mastery of English and their capability for expressive communication in the context of business presentations. It also indicated the possibility of using less common PVs in specific domains. Semantically, PVs were found to play a crucial role in both describing business strategies (see the example in the title) and signposting and transitioning within presentations. The findings illuminated the strategic use of PV for guiding the audience through various sections of the presentation, contributing to tighter overall coherence and clarity in communication. Despite the overall high accuracy observed, rare instances of incorrect or unnatural PV usage were noted. These instances were speculated to stem primarily from inadvertent grammatical slips, likely reflective of typical oral communication errors.

This study not only contributes to our understanding of advanced English speakers' PV usage in spoken business English but also offers practical implications for language educators and researchers seeking to enhance language proficiency and practical communication skills in university students.

References

Anthony, L. (2023). *AntConc* (Version 4.2.2) [Computer software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly, 41*(2), 339–359. DOI: https://DOI.org/10.1002/j.1545-7249.2007.tb00062.x

Jiang, Y. (2015). Study of language features of business English. *Higher Education of Social Science, 8*(5). DOI: https://DOI.org/10.3968/6939

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly, 45*(4), 661–688. DOI: https://DOI.org/10.5054/tq.2011.247707

Wray, A. (2013). Formulaic language. *Language Teaching, 46*(03), 316–334. DOI: https://DOI.org/10.1017/S0261444813000013