Corpus Linguistics in the Digital Era: Genres, Registers and Domains

La lingüística de corpus en la era digital: géneros, registros y dominios

14th International Conference on Corpus Linguistics - May 10 - 12, 2023

XIV Congreso Internacional de Lingüística de Corpus

Book of abstracts – Libro de sumarios

aelinco

Universidad de Oviedo

Facultad de Filosofía y Letras

Corpus Linguistics in the Digital Era: Genres, Registers and Domains

La lingüística de corpus en la era digital: géneros, registros y dominios

14th International Conference on Corpus Linguistics - Oviedo, May 10 - 12, 2023

XIV Congreso Internacional de Lingüística de Corpus – Oviedo, 10 a 12 de mayo de 2023

The abstracts are in the form in which they were submitted by their authors with some minor editing for consistency and intelligibility, and also as a result of possible electronic cross-platform mismatches.

Corpus Linguistics in the Digital Era: Genres, Registers and Domains
La lingüística de corpus en la era digital: géneros, registros y dominios

TABLE OF CONTENTS – ÍNDICE DE CONTENIDOS

PANELS PAPERS, POSTERS AND SEMINAR - COMUNICACIONES, PÓSTERES Y SEMINARIO

POSTERS

CORPUS COMPILATION SEMINAR

PLENARY LECTURES

Creating urban place through discourse and other semiotic sign-making:
What do narratives of Brooklyn, NY, and those of the death of Queen Elizabeth II have in common?

Beatrix Busse – *University of Cologne*

Discursive place-making strategies and semiotic meaning-making are two interconnected aspects that shape people's experience and interaction with an urban environment. They also create it because the process of how a space becomes a place involves negotiating the multitude of voices, signs, interests, and identities that converge within the city (Busse 2019, 2021, 2022, Cresswell 2015, Busse & Warnke 2022). Place is socially constructed and it is dynamic, provisional, unpredictable, and discursively manipulated (Stokowski 2017).

Over the past decade my work on urban discourse in selected neighbourhoods of Brooklyn, New York, and London, has led me to analyze which semiotic forms and constructions take on foregrounded practices of so-called discursive urban-place-making. I investigate how these both stable and changing practices interact in the various neighborhoods of Brooklyn, in the virtual spheres related to them as well as in practices of other geolocations, such as London neighborhoods, or of capturing specific incidents, such as the historic passing away of Queen Elizabeth I. In this paper, I will show how these practices carry the potential for (re-)indexing specific social values of an urban neighborhood, of the borough itself or of iconic figures. These discourses and meaning-making semiotic practices position themselves in the social, cultural, political, and economic spheres of urbanity.

To answer these questions, I will analyse data from two subcorpora: (a) *heiUrban*- an ongoing project from 2012-2023 investigating urbanity and discursive place-making strategies in Brooklyn, New York, and (b) the more recent 2022-corpus on *Cultural Heritage in London: Queen Elizabeth II's legacy*.

The *heiUrban* dataset currently so far includes 1579 individual audio recordings of interviews with people in Brooklyn taken in 2012, 2015, 2017, and 2023 (16 hrs 37 mins 34 secs). The transcriptions of these interviews amount to ca. 536,923 words including annotations. Moreover, we collected 9754 photographs of the linguistic landscape and over approximately 160 million Tweets mentioning Brooklyn and those geo-tagged as coming from the area. In order to capture both the visible and 'hidden' discourse patterns, we also collected over local 47,000 Wi-Fi SSIDs in neighbourhoods of interest around Brooklyn in 2017, as well as more recent Wi-Fi data collected in 2023, which is currently being processed. For this presentation, the focus will be on the most recently collected 2023-data.

The *Cultural Heritage in London* project consists of 2062 photographs and 299 audio recordings (13 hrs 3 min 1 sec) of interviews with people who queued to see Queen Elizabeth lying-in-state in September and October 2022, as well as people in Shoreditch to see what language is used to describe the Queen, the different areas, and perceptions about the changes to the King's English and royal imagery.

As can be seen, the corpus is constructed through collecting and organizing language data from various modalities such as text, speech, image, and video within one digital interface – a rare and complex mix of data types, which also demands for a mixed-methods approach. Hence, this paper will also give a critical stance on multimodal corpus compilation and analysis (Knight and Adolphs 2020) and their value in urbanity studies. Furthermore, Pink (2008) and Busse and Warnke (2022) suggest that ethnographic researchers are also part of the place-making process itself, as the empirical and sensorial aspects of collecting data contributes to the overall perception and understanding of communicative practices in that place, not merely by engaging with people in the place of interest but also by walking, dining, and by truly embodying the experience of data collection as a researcher. Hence, I aim to explore the future of the field of both corpus linguistic and urban studies, given the recent technological advancement, our roles as researchers and the challenges and chances of doing justice to human practices of urban place-making.

References

Aist, G., Allen, J., Campana, E., Galescu, L., Gomez Gallo, C.A., Stoness, S., Swift, M., & Tanenhaus, M. 2006. Software architectures for incremental understanding of human speech, 1922–1925. In: *Proceedings of Interspeech* 2006, Pittsburgh.

Busse, B. 2019. Patterns of discursive urban place-making in Brooklyn, New York. In V. Wiegand & M. Mahlberg (eds.), *Corpus linguistics, context and culture,* 13–42. Berlin: Mouton de Gruyter.

Busse, B. 2021. Practices of discursive urban place-making in Brooklyn, New York: (hidden) digital and embodied discourse. *Text & Talk,* Vol. 41 (Issue 5-6), pp. 617-641.

Busse, B. 2022. 15. The HeiURBAN Database: A Brief and Unconventional Position Piece. In: Busse, B. and Warnke, I. ed. *Handbuch Sprache im urbanen Raum Handbook of Language in Urban Space.* Berlin, Boston: De Gruyter, pp. 394-414.

Busse, B. & Warnke, I. 2022. Urban Linguistics: Ideas and Anchor Points. In: Busse, B. and Warnke, I. ed. *Handbuch Sprache im urbanen Raum Handbook of Language in Urban Space.* Berlin, Boston: De Gruyter, pp. 1-32.

Cresswell, T. 2015. *Place: An introduction,* 2nd edn. Malden: Wiley Blackwell.

Knight, D., Evans, D., Carter, R. A. & Adolphs, S. 2009. Redrafting corpus development methodologies: Blueprints for 3rd generation multimodal, multimedia corpora. *Corpora,* 4(1), 1–32.

Knight, D. & Adolphs, S. 2020. Multimodal Corpora. In: Paquot, M., Gries, S.T. (eds) *A Practical Handbook of Corpus Linguistics.* Springer, Cham.

Mana, N., Lepri, B., Chippendale, P., Cappelletti, A., Pianesi, F., Svaizer, P., & Zancanaro, M. 2007. Multimodal Corpus of Multi-Party Meetings for Automatic Social Behavior Analysis and Personality Traits Detection. In: *Proceeding of Workshop on Tagging, Mining and Retrieval of Human-Related Activity Information at ICMI'07* (pp. 9–14). Nagoya, Japan.

Pink, S. 2008. An Urban Tour: The Sensory Sociality of Ethnographic Place-Making. *Ethnography* 9:2, 175–96.

Stokowski, P. A. 2002. Languages of Place and Discourses of Power: Constructing New Senses of Place. *Journal of Leisure Research* 34:4, 368-382.

***Tomorrow I'll go jogging*: A corpus-based analysis of the history of the English absentive construction**

Teresa Fanego – *University of Santiago de Compostela*

Keywords: *absentive; grammar network; grammatical obsolescence; narrative dissonance; prescriptivism; progressive.*

This presentation explores the history of the construction exemplified in the title, henceforth referred to as the Expeditionary *Go* construction, and its relation to the absentive. The absentive is an aspectual category closely related to the progressive; it can be defined as the grammatical expression of absence (Bertinetto et al. 2000; de Groot 2000; Abraham 2007). As first described by de Groot (2000), the absentive carries the information that the human Subject is away from the Speaker-deictic centre, engaged in a temporary activity; the notion of absence implies remoteness and distance: it is not possible to use an absentive construction when the Subject is visible and in the direct neighbourhood of the Speaker.

An example of the absentive from German, which employs for that purpose a periphrasis consisting of copula + infinitive, can be seen below:

(1) *Jan ist box-en.*
John is box-INF
'John is out boxing.'

Unlike the various Germanic and non Germanic languages discussed in de Groot (2000), Present-day English (PDE) is not considered to have an absentive construction distinguished by grammatical form. In this presentation, however, I will argue that, in earlier stages, English possessed a construction, identifiable by both meaning and morphosyntax, which was specialized in expressing absence from the deictic centre. In Old and Middle English the construction in question consisted of a generic motion verb with the meaning 'go' plus a phrase with the preposition *on/an* and a gerundial noun in *-ing*:

(2) *Þis child scholde wende* ***an hontingue.***
'This child should go [out] on hunting.'
(*c*1300 *St. Kenelm* (Laud) 148; *OED go* v. 30.f.(a))

Over the course of the history of English, several variants of the construction have co-existed, namely the *on/an* pattern illustrated in (2) and the patterns in (3)–(5) below:

(3) With phonetic reduction of *on/an* to *a*:

he looked upon coaches as horn-blowing contrivances, quite beneath the dignity of men, and only suited to giddy girls that did nothing but chatter and **go a-shopping**. (CLMET3.02 1839 Dickens, *Barnaby Rudge*)

(4) With the absentive marker *out*:

Thomas Foley gives an entertainment to the Admiralty; but I pleaded health, and remained at home. Neither will I **go out sightseeing**, which madness seems to have seized my womankind. (CLMET3.02 1825–1832 Scott, *The journal of Sir Walter Scott*)

(5) Without any marker intervening between the motion verb and the *-ing* form:

But he is stupid cos he **went hitchhiking** once and left a day early to avoid traffic. (BNC BYU 1992, *55 convs rec. by Josephine*)

The analysis seeks to assess when and why the variants of Expeditionary *Go* illustrated in (2)–(4) became obsolescent, so that the construction developed into the rather fixed pattern (5) it is today. A further concern of the presentation is to explore the reasons leading to the frequent derogatory overtones of expeditionary clauses, as seen in some of the above examples.

Data for the study have been gathered from a vast collection of corpora and databases including *Nerthus Lexical Database of Old English* (Martín-Arista et al. 2016), EEBO BYU (1470–1700, 755 million words), the *Corpus of Late Modern English Texts* (CLMET3.0, 1710–1920; 34 million words), and BNC BYU (1975–1994; 100 million words).

References

Abraham, Werner. 2007. Absent arguments on the absentive: An exercise in silent syntax. Grammatical category or just pragmatic inference? *Groninger Arbeiten zur Germanistischen Linguistik* 45.3–16.

Bertinetto, Pier Marco, Karen H. Ebert & Casper de Groot. 2000. The progressive in Europe. In Östen Dahl (ed.), *Tense and aspect in the languages of Europe*, 517–558. Berlin: Mouton de Gruyter.

BNC BYU = Davies, Mark. 2004–. *British National Corpus* (from Oxford University Press). URL: https: //www.english-corpora.org/bnc/

CLMET3.0 = De Smet, Hendrik, Hans-Jürgen Diller & Jukka Tyrkkö. 2013. *The Corpus of Late Modern English Texts, version 3.0*. Leuven: K. U. Leuven.

EEBO BYU = Davies, Mark. 2017. *Early English Books Online. Part of the SAMUELS project*. URL: https: //www.english-corpora.org/eebo/

Groot, Casper de. 2000. The absentive. In Östen Dahl (ed.), *Tense and aspect in the languages of Europe*, 641–667. Berlin: Mouton de Gruyter.

Martín-Arista, Javier, Laura García-Fernández, Miguel Lacalle-Palacios, Ana Elvira Ojanguren-López & Esaúl Ruiz-Narbona (eds.). 2016. *NerthusV3. Online Lexical Database of Old English*. Universidad de La Rioja: Nerthus Project.

Exploring the base of the iceberg: Writing processes in learner corpus research

Gaëtanelle Gilquin – *Université Catholique Louvain*

Learner corpus research has made an important contribution to the study of L2 writing through the analysis of large databases of authentic texts written by representative samples of learners. So far, however, the almost exclusive focus on the final texts, of which most learner corpora are composed, means that we have mainly considered the tip of the iceberg, that is, the visible part of writing. The different processes leading to the written product, such as editing, proofreading or use of external resources, are usually invisible in learner corpora, and hence in learner corpus research. Yet, the literature on writing has long highlighted the importance of processes in addition to products (e.g. Murray 1980, Krapels 1990, Sasaki 2000).

This presentation will explore the base of the iceberg, namely the processes that are involved in writing texts in L2, and will show how learner corpus research could contribute to a better understanding in this respect too. Special types of learner corpora will be introduced that give access to writing processes, including PROCEED, the Process Corpus of English in Education (Gilquin 2022). This corpus contains written texts, like traditional written learner corpora, but for each text it also provides a screencast video and a keylog file, which reproduce the writing process as it unfolds.

Illustrations of how a corpus like PROCEED can be used in learner corpus research will be given. We will see how the corpus makes it possible to investigate phenomena such as writing fluency or dictionary consultation, but also to speculate about more cognitive aspects of L2 acquisition. Examples of possible pedagogical applications will be given too. The challenges of dealing with writing process data will be outlined, as well as the potential for further development of this type of research.

References

Gilquin, G. 2022. The *Process Corpus of English in Education*: Going beyond the written text. *Research in Corpus Linguistics* 10(1): 31-44.

Krapels, A. R. 1990. An overview of second language writing process research. In B. Kroll (ed.), *Second Language Writing: Research Insights for the Classroom* (pp. 37-56). Cambridge: Cambridge University Press.

Murray, D M. 1980. Writing as process: How writing finds its own meaning. In T. R. Donovan & B. W. McClelland (eds.), *Eight Approaches to Teaching Composition* (pp. 3-20). Urbana, IL: National Council of Teachers of English.

Sasaki, M. 2000. Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing* 9(3): 259-291.

Corpus linguistics and political speaking:
The trends, tropes, and techniques of influencing people through words

Jukka Tyrkkö – *Linnaeus University*

Politics is a field of human activity that relies extensively on the careful and intentional use of language. Throughout history, there have been specific instances of political speaking that have become enshrined in culture ("I have a dream…", "ask not what your country can do for you…", "we shall fight on the beaches…"), but it is also evident that political speaking has gone through meaningful changes over time, and that there is much to be gained from the systematic analysis of not only individual speech events, but of political speaking as a genre that both responds to and shapes public attitudes and sensitivities.

In this talk I will discuss political speeches as a specific type of culturally and societally significant discourse, with particular reference to the value of establishing baseline evidence of linguistic practice through empirical analysis. Findings concerning both well-known rhetorical devices and entirely data-driven observations will be juxtaposed against contemporary historical, political and technological developments to show the sensitivity of political speaking to the context in which speeches are delivered.

From a methodological standpoint, the presentation will discuss striking a balance between rich and "overly" rich metadata in the analysis of political language use, and the various ways in which corpus methods and related text analytical approaches can be used to explore conceptual and thematic developments in historical language data.

What corpora can and cannot do for journalistic translation research

Roberto A. Valdeón – *University of Oviedo*

This talk will discuss the usefulness of corpus-based approaches in the study of journalistic translation, a sub-area of research within Translation Studies that started in the last decade of the twentieth century and has burgeoned since 2000. I will start by looking at the introduction of corpus-based studies in the 1990s, following the work of Mona Baker, and its evolution thereafter. Corpus-based studies analyzed features of translated texts such as explicitation and simplification, giving way to the proposal of universals in translations. Corpus-based approaches were instrumental in the study of the interaction between languages and the impact it had on target texts in a variety of settings and genres, including audiovisual translation. For example, the Pavia Corpus of Film Dialogue, a parallel and comparable corpus of Anglophone films and their Italian versions, has been the basis of the work by Maria Pavesi, amongst others.

The interest in journalistic translation can be traced back to the late 1980s, when the first articles on the topic, by Akio Fujii and Karen Stetting, were published. Fujii applied the concept of gatekeeping to news translation while Stetting coined 'transediting', which has become a recurrent term in the literature. Both had worked as journalists. News translation, or more broadly journalistic research, is now an established research strand within translation, after the publication of special issues of journals such as *Language and Intercultural Communication*, *Meta*, *Across Languages and Cultures*, *Perspectives* and *Journalism* devoted to it as well as a number of monographs and edited collections. Researchers have focused on the product rather than on the process of translation with some notable exceptions (e.g. Lucile Davier, Esperança Bielsa, Marlie van Rooyen).

The study of news translation/production as a product has drawn on the work of journalism and translation scholars. Researchers have had recourse to concepts such as gatekeeping and agenda-setting from communication studies, the tenets of critical discourse analysis and the tools of functional linguistics. Methodologically speaking, authors have compared source and target texts using corpus-based approaches. However, given the peculiarities of news texts, corpus-based studies are limited to certain contexts. On the one hand, the role of translation in news production remains problematic: How is translation used by news writers? Are news writers translators? On the other, the very nature of news production/translation makes it difficult to ascertain whether we study translation, that is, is translated news translation? Is it possible to locate source and target text? This talk will discuss some of the challenges that we encounter in the study of journalistic translation, especially for corpus-based approaches.

ROUND TABLES

English historical corpora ten years on

Chair: Javier Calle Martín – *University of Málaga*

Participants: Carolina Amador-Moreno – *University of Bergen*), Isabel Moskowich – *University of A Coruña*, Paula Rodríguez Puente – *University of Oviedo,* Javier Ruano-García – *University of Salamanca*, Nuria Yáñez-Bouza – *University of Vigo & University of Manchester.*

The *4th International Conference of Corpus Linguistics*, held in the year 2012 at the University of Jaén, pioneered the inclusion of a number of pre-conference workshops on the design, compilation and applications of the corpora then compiled in different Spanish universities. One of these workshops, entitled *English Historical Corpora Compiled in Spain*, was a landmark in the field giving room and voice to different collections of texts, both synchronic and diachronic, which were then germinating in our country (see Vázquez 2012). It is now ten years since then and the picture has drastically changed, both in the number of corpora made available and in their architecture. This round table has been conceived of as a state of the art of English historical corpora ten years after the 2012 workshop, hosting some new proposals offered as primary sources for linguistic research. Carolina Amador will open the session with CORIECOR – *A Corpus of Irish English Correspondence* (c. 1700-1900), followed by Javier Calle and Isabel Moskowich, who will present their scientific collections of texts in *The Málaga Corpus of Late Modern English Scientific Prose (1700-1900)* and *The Coruña Corpus of English Scientific Writing*, respectively. Next, Javier Ruano's account of *The Salamanca Corpus: Digital Archive of English Dialect Texts (1500-1950)* will be followed by Nuria Yáñez's collection of ego-documents in "The Mary Hamilton Papers (c. 1740 – c.1850): A Treasure Trove for the Study of Literary and Linguistic Social Networks". Finally, Paula Rodríguez will close the session with her present-day English proposal entitled "CoCELD – Corpus of Contemporary English Legal Decisions, 1950-2021: Challenges and Benefits of Compiling a Corpus of Legal Discourse". After the individual presentations on each corpus, the participants in the round table will engage in an interactive discussion to share knowledge on the advantages of corpus compilation in tandem with the digital humanities, and they will also address the challenges faced during the process as well as for future work in the field.

References

Vázquez, Nila (ed.). 2012. *Creation and Use of Historical English Corpora in Spain*. Newcastle upon Tyne: Cambridge Scholars Publishing.

CORIECOR: A *Corpus of Irish English Correspondence*, C. 1700–1900

Carolina P. Amador-Moreno – *University of Bergen*

Keywords: *Irish English, historical sociolinguistics, data visualization, Irish emigration, private correspondence, corpus of Irish English Correspondence.*

High net emigration from Ireland over the last 400 years has resulted in large Irish and overseas archives that capture the evolution of IrE and its influence on other varieties.

As noted in recent research, systematic exploration of this variety's development is impossible without access to the right kinds of longitudinal data such as The Corpus of Irish English Correspondence (CORIECOR), a historical sociolinguistic corpus of letters from the late-17th to the early 20th century.

CORIECOR has provided a unique opportunity for researchers to trace the emergence and development of features of IrE and to explore stylistic, regional, and social variation along the lines of the historical sociolinguistic survey enabled by the creation of data such as the *Corpus of Early English Correspondence*, described in Nevalainen & Raumolin-Brunberg (2016).

CORIECOR was devised by Kevin McCafferty and the present author back in 2008, when it was first funded by the University of Bergen. The purpose of what became our overall project was to compile a corpus that would be available to researchers studying the evolution and spread of IrE in Ireland, issues of language contact between speakers of Irish and English during the language shift, and the influence of IrE on other Englishes around the world. Our aim was to gather and make available to the broader research community a large body of linguistic data that would enable us to study IrE through time, using larger amounts of more vernacular data from a long timespan.

For a new stage of CORIECOR initiative, a recent Spanish-funded project called CORIECOR Visualized (CORVIZ) has allowed us to focus on socio-pragmatic issues affecting language use in private communication, to combine Corpus Linguistics methodology and visualization techniques, and to make CORIECOR available to the broad academic (and non-academic) community through a website that will allow for linguistic searches and visualization of results: https://corviz.h.uib.no/index.php .

The material included in CORIECOR was compiled for use by families, historians and genealogists. Its original format was not suitable for linguistic purposes. For this reason, the letters contained in the corpus were initially reformatted and catalogued only in the simplest of ways (e.g. by year and country of origin of the letter). The CORVIZ project has allowed this material to be connected, reorganised and tested. The work conducted so far has employed visualizations tools in order to better explore the language use of individual writers who have been grouped into networks of family, friends, colleagues, business associates, etc., to permit research based on social network approaches that have proved fruitful in the study of the language both historically and in present-day contexts.

References

Nevalainen, T. & H. Raumolin-Brunberg 2003/2016. Historical sociolinguistics. Language change in Tudor and Stuart England. London: Longman.

The *Málaga Corpus of Late Modern English Scientific Prose*

Javier Calle-Martín – *University of Málaga*

Keywords: *corpus compilation, Late Modern English, Málaga Corpus, tagging.*

The Málaga Corpus of Early English Scientific Prose is a collection of English vernacular medical writing, consisting of three diachronically divided components, i.e. The Málaga Corpus of Late Middle English Scientific Prose (1350-1500); The Málaga Corpus of Early Modern English Scientific Prose (1500-1700); and The Málaga Corpus of Late Modern English Scientific Prose (1700-1900). The three components have been purposely designed so as to contain evidence from the three text types of medical writing in English, that is, theoretical treatises, surgical treatises and recipe collections. In itself, the corpus stems from actual linguistic evidence of the period, both handwritten and printed, standing out as the ideal input for diachronic linguistic research at the levels of spelling, morpho-syntax and lexis.

The present paper is particularly concerned with the third component of the corpus, *The Málaga Corpus of Late Modern English Scientific Prose* (1700-1900), which has been recently published and made available in the project's webpage (https://latemodernmss.uma.es). In its current form, the corpus amounts to 2.5 million words, of which 1.5 million belong to the 18th century and the other million to the 19th century. The corpus is offered in three different formats, that is, the plain text version, the modernised version and the tagged version. The CQP-web version is also available for online use (https://latemodernmss.uma.es/cqpweb/). The present paper first describes the rationale of the corpus considering the typology of texts, their chronology, the text types and authorship. Second, the paper delves into the process of compilation, which is a sequential process consisting of a) modernisation by means of VaRD (Variant Detector) and b) automatic tagging by means of CLAWS (Constituent Likelihood Automatic Word-tagging System). The paper closes with a brief demonstration of the corpus potential using the CQP-web version.

References

Calle-Martín, Javier, Miriam Criado-Peña, Verónica Hernández, Sinéad Linehan-Gómez and Juan Lorente-Sánchez. 2016. *The Málaga Corpus of Late Modern English Scientific Prose (MCLModESP)*. Málaga: University of Málaga. Available from https://latemodern.uma.es.

The *Coruña Corpus of English Scientific Writing* 20 years later

Isabel Moskowich – *Universidade da Coruña*

Keywords: *Coruña Corpus, diachrony, scientific discourse.*

The *Coruña Corpus of English Scientific Writing* is a purpose-built electronic corpus conceived as a resource for the study of eighteenth and nineteenth-century scientific writing in English. The project began in 2003 as a means of exploring the historical background of English as the language of science, and is now founded on solid grounds: socio-external considerations, theoretical principles of Corpus Linguistics, and technical robustness, all of which serve to confirm its value as a resource for research (Crespo and Moskowich, 2020). The Coruña Corpus is a specialised corpus in which texts dealing with different scientific disciplines are compiled. The texts belong to the late Modern English period as there are certain events clearly delimiting the history of science. Thus, our corpus stems from the belief in the inextricable relationship between language and society. It is this interest in the relationship between registers and their users that has led us to include information about the author (sex, age, place of education) and the text (genre, reception, date of publication) for each individual sample. This presentation aims at providing a complete description of the Coruña Corpus of English Scientific Writing and the changes it has been undergoing across the last 20 years (Moskowich et al, 2020), from the seminal idea in 2003 to the present. Therefore, I will present the compilation principles underneath, the different decisions on representativity and balance we had to make, as well as its structure and availability. Complying with the key principles promoted by the European Commission regarding open-access science, all sections of the Coruña Corpus are freely accessible at the University of A Coruña Repository, RUC (https://ruc.udc.es/dspace/handle/2183/21846), although a couple of them were originally published by John Benjamins. As an ongoing project, the Coruña Corpus grows slowly but uninterruptedly, and we will also introduce here the new suborpora which are in preparation as well as the five already available. The scientific disciplines for which we have already compiled text samples embrace Astronomy, Philosophy, History, Life Sciences and Chemistry. Thus, the first subcorpus compiled was *CETA*, *Corpus of English Texts on Astronomy* (2012), and the second was *CEPhiT, Corpus of English Philosophy Texts* (2016). Both were re- issued in open access later. The third, the *Corpus of History English Texts* (*CHET*) was released in 2019. Shortly afterwards, in 2020, the fourth corpus, the *Corpus of English Life Sciences Texts* (CELiST) was published (Moskowich, 2021), followed by the *Corpus of English Chemistry Text (CECheT*) in 2022. The *Corpus of English Texts on Languages* (*CETeL*) is currently being prepared for publication (Monaco and Puente-Castelo, 2019) and a preliminary version of the corpus on Physics is being put together.

References

Crespo, Begoña and Moskowich, Isabel. 2020. Astronomy, philosophy, life sciences and history texts: setting the scene for the study of modern scientific writing. *English Studies*, 101:6, 665- 684.

Monaco, Leida Maria and Luis Puente-Castelo. 2019. 'A matter both of curioſity and uſefulneſs': Compiling the Corpus of English Texts on Language. *Research in Corpus Linguistics*, 7, 47-68.

Moskowich, Isabel. 2021. The making of the Corpus of English Life Sciences Texts (CELiST), a bunch of disciplines. In Moskowich, Isabel; Lareo, Inés and Camiña Rioboó, Gonzalo (eds.), *"All families and genera": Exploring the Corpus of English Life Sciences Texts.* Amsterdam: John Benjamins. 2–19.

Moskowich, Isabel; Puente-Castelo, Luis; Crespo, Begoña and Camiña, Gonzalo. 2020. The Coruña Corpus of English Scientific Writing: Challenge and Reward. *Nexus*, 2020/2: 31-38.

***The Corpus of Contemporary English Legal Decisions, 1950-2021*: Challenges and benefits of compiling a corpus of legal discourse**

Paula Rodríguez Puente – *University of Oviedo*

Keywords: *corpus compilation; legal discourse; Plain Language Movement.*

Legal discourse is widely assumed to be resistant to change or "outside the ravages of time" (Görlach 1999: 145), and indeed legislative documents are extremely conservative with fixed and formulaic structures (see, e.g., Biber & Gray 2019). However, recent research has shown that changes can be observed in the lexico-grammatical features of some legal documents when examined diachronically (see, e.g., Rodríguez-Puente 2019, 2020), particularly since the emergence in the 1970s of the Plain Language Movement, which sought to draw attention to the "unnecessary complexity of the official language used by governments, businesses, and other organizations which are in linguistic contact with the public" (Crystal 2019: 401). Williams (2007), for example, notes a considerable reduction of the use of the passive voice in prescriptive legal documents between the 1980s and the 2010s, which may respond to a need for a more functional and accessible linguistic expression as a means of satisfying those demands for plainer language (see also Williams 2004, 2013; Bulatović 2013).

Despite the crucial changes in legal language in recent years, research in that direction is scarce to date, particularly in the British English variety, probably due, in part, to the shortage of specialized corpora that allow this kind of studies. Most of the available corpora gather different types of legal documents produced in the USA, such as those compiled at the Brigham Young University (see https://lawcorpus.byu.edu/), as well as the *Corpus of US Supreme Court Opinions* (Davies 2017). As far as British English is concerned, the *Corpus of Historical English Law Reports* (Rodríguez-Puente et al. 2018) contains judicial opinions produced from 1535 to 1999, but the twenty-first century is not represented in it. The *British Law Report Corpus* (Marín-Pérez & Rea-Rizzo 2012) also contains legal decisions produced between 2008 and 2010, but does not allow for any long-term diachronic investigation.

In order to bridge this gap, we have embarked on the compilation of the *Corpus of Contemporary English Legal Decisions* (CoCELD), a corpus of British judicial decisions produced between 1950 and 2021. The corpus contains decisions held at the House of Lords, the Supreme Court and the Privy Council, which have been downloaded from the *British and Irish Legal Information Institute* (BAILII).[1] The House of Lords was UK's highest court of appeal until July 2009. From October 2009, the Supreme Court assumed jurisdiction on points of law for all civil cases in the UK and the criminal cases in England and Wales, as well as Northern Ireland. The Privy Council, in turn, is the highest court of appeal for certain British territories, some Commonwealth countries and a few UK bodies.

In this paper we present the structure and characteristics of the CoCELD, as well as the methodology used for its compilation. The new corpus, which will be released in February 2022, will contain two sample texts of roughly 2,500 words for each year from 1950 to 2021 (i.e. 288 files), which will add up to approximately 720,000 words. The corpus will contain files in raw text and POS-annotated files, and will be freely available for the research community under signed consent. With CoCELD we hope to contribute with a new, useful resource for linguists with an interest in legal language, from both a synchronic and a diachronic perspective.

Bibliography

Biber, Douglas & Bethany Gray. 2019. Are law reports an "agile" or an "uptight" register? Tracking patterns of historical change in the use of colloquial and complexity features. In Teresa Fanego & Paula Rodríguez-Puente (eds.), *Corpus-based Research on Variation in English Legal Discourse*, 149-169. Amsterdam: John Benjamins.

[1] http://www.bailii.org/

Bulatović, Vesna. 2013. Legal language: The passive voice myth. *ESP Today. Journal of English for Specific Purposes at Tertiary Level* 1(1): 93-112.

Crystal, David. 2019. *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press.

Görlach, Manfred. 1999. *Nineteenth-century England: An Introduction*. Cambridge: Cambridge University Press.

Davies, Mark. 2017. *Corpus of US Supreme Court Opinions*. Available online at https://www.english-corpora.org/scotus/

Marín Pérez, María José & Camino Rea Rizzo. 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.

Rodríguez-Puente, Paula. 2019. Interpersonality in legal written discourse: A diachronic analysis of personal pronouns in law reports, 1535-present. In In Teresa Fanego & Paula Rodríguez-Puente (eds.), *Corpus-based Research on Variation in English Legal Discourse*, 171-199. Amsterdam: John Benjamins.

Rodríguez-Puente, Paula. 2020.Historical legal discourse: British law reports. In Eric Friginal & Jack A. Hardy (eds.), *The Routledge Handbook of Corpus Approaches to Discourse Analysis*, 499-517. New York: Routledge.

Rodríguez-Puente, Paula, Teresa Fanego, María José López-Couso, Belén Méndez-Naya, Paloma Núñez-Pertejo, Cristina Blanco-García & Iván Tamaredo. 2018. *Corpus of Historical English Law Reports 1535-1999 (CHELAR), v.2*. University of Santiago de Compostela: Research Unit for Variation, Linguistic Change and Grammaticalization, Department of English and German.

Williams, Christopher. 2004. Legal English and plain language: An introduction. *ESP Across Cultures* 1: 111-124.

Williams, Christopher. 2007[2005]. Tradition and Change in Legal English. Verbal Constructions in Prescriptive Texts. Bern: Peter Lang.

Williams, Christopher. 2013. Changes in the verb phrase legislative language in English. In Bas Aarts, Joanne Close, Geoffrey Leech & Sean Wallis (eds.), *The Verb Phrase in English. Investigating Recent Language Change with Corpora*, 353-371. Cambridge: Cambridge University Press.

The Salamanca Corpus: **Challenges and avenues for future research in the history of English dialects**

Javier Ruano-García – *University of Salamanca*

Keywords: *Salamanca Corpus, dialects, British English, Early Modern English, Late Modern English, variation and change.*

The Salamanca Corpus (SC) is an ongoing project that has been designed as a digital archive of dialect texts written between 1500 and 1950 that can help us improve the history of English dialects, one that remains fragmented and poorly understood as a result of the scarcity of sources. The main aim of the SC is thus to recover and digitise older and hardly accessible texts including specimens of literary dialect and dialect literature (Shorrocks 1996), as well as provincial glossaries from all over England (García-Bermejo Giner 2012). They feature amongst the most relevant materials that have come down to us to investigate Early and Late Modern English dialects, as non-literary records of provincial speech, such as those documented in correspondence and journals are scant and often hard to locate (see, however, García-Bermejo Giner and Montgomery 1997; Auer et al. forthcoming).

This paper shows what the SC can tell us about the history of English dialects. It outlines some of the major challenges behind its compilation over the past decade, including the dearth of data for specific varieties and the difficult interpretation of the evidence offered by some texts. At the same time, I report on the findings of recent work so as to underline that, despite the pitfalls and widespread criticism about the use of such evidence for linguistic research, the data preserved in dialect writing and lexicography sheds useful light on the history of forms that remain little explored, while it can prove beneficial in reconstructing linguistic ideas about dialects over time. Attention will be paid to selected case studies, which include grammatical phenomena such as pronoun exchange (e.g. *her ate* 'she ate') and ongoing work on dialect enregisterment (see Ruano-García 2020, 2023; Schintu 2022). In so doing, this contribution to the Round Table seeks not only to provide an overview of the current scope and materials of the corpus, but also to highlight the areas that await further investigation, while I point to some avenues of research that can now be explored given the increasing availability of dialect sources.

References

Auer, Anita, Anne-Christine Gardner and Mark Iten. forthcoming. "Patterns of linguistic variation in Late Modern English pauper petitions from Berkshire and Dorset". In Markus Schiegg & Judith Huber (eds.). *Intra-Writer Variation in Historical Sociolinguistics*. Bern: Peter Lang.

García-Bermejo Giner, María F. 2012. "The Online Salamanca Corpus of English Dialect Texts". In Nila Vázquez-González (ed.). *Creation and Use of Historical English Corpora in Spain*. Newcastle upon Tyne: Cambridge Scholars Publishing, 67-74.

García-Bermejo, María F. and Michael Montgomery. 1997. "British regional English in the nineteenth century: The evidence from emigrant letters". In Alan R. Thomas (ed.). *Issues and Methods in Dialectology*. Bangor: University of Wales Bangor, Department of Linguistics, 167–183.

Ruano-García, Javier. 2020. "On the enregisterment of the Lancashire dialect in Late Modern English: Spelling in focus". *Journal of Historical Sociolinguistics* 6(1), 1–38.

Ruano-García, Javier. 2023. "'Well, taaking about he da bring inta me yead wat I promised var ta tell ee about': Representations of south-western speech in ninteenth-century dialect writing". *English Language and Linguistics* (Special issue "Speech representations in Late Modern English text types", eds. Anita Auer et al.).

SC = *The Salamanca Corpus: Digital Archive of English Dialect Texts*. Eds. María F. García-Bermejo Giner, Pilar Sánchez-García & Javier Ruano-García. 2011–. http://www.thesalamancacorpus.com.

Schintu, Paula. 2022. *The Enregisterment of Late Modern Derbyshire Dialect (1850–1950)*. PhD dissertation, Universidad de Salamanca.

Shorrocks, Graham. 1996. "Non-standard dialect literature and popular culture". In Juhani Klemola, Merja Kytö & Matti Rissanen (eds.). *Speech Past and Present: Studies in English Dialectology in Memory of Ossi Ihalainen*. Frankfurt am Main: Peter Lang, 385–411.

The Mary Hamilton Papers (c.1740 – c.1850): **A treasure trove for the study of literary and linguistic social networks**

Nuria Yáñez-Bouza – *Universidade de Vigo & University of Manchester*

Keywords: *digital edition, ego-documents, Late Modern English, letter writing, Mary Hamilton, reading practices, social networks.*

Over the past decades, interest in the study of ego-documents has been thriving from both a historiographic as well as a historical sociolinguistic perspective (Dekker 2000, van der Wal & Rutten 2013), and access to authentic historical materials has improved as a consequence of advances in methodological tools in corpus linguistics and digital humanities. In the context of Late Modern English, this has led to an increasing number of data sources in the form of linguistic corpora (e.g. *Corpus of Early English Correspondence Extension*, *Corpus of Irish English Correspondence*) or digital editions (e.g. *The Elizabeth Montagu Letters*, *The Collected Letters of Hannah More*). The project undertaking the compilation of *The Mary Hamilton Papers (c.1740–c.1850)* has added to this growing body of scholarship with a new historical source, which is offered both as an online digital edition via Manchester Digital Collections and as a linguistic corpus with searchable transcriptions via CQPweb. This paper offers an overview of its contents and the digitisation and annotation processes carried out in collaboration between the University of Vigo in Spain and the universities of Manchester and York in the UK.

Mary Hamilton (1756–1816) was a royal sub-governess in the court of George III and a member of the Bluestocking circle, thus a well-connected figure in royal, aristocratic and literary circles of the late eighteenth and early nineteenth ceturies (Crawley 2014). *The Mary Hamilton Papers* contains her private correspondence, diaries and other personal writings together with materials pertaining to her husband John Dickenson and his family. The time-span covered by the collection (c.1740–c.1850) and the wide range of topics addressed (court and royal life, literary interests, women's education, courtship and romance, the Bluestocking network, etc.) make this edition a unique data source to explore the intellectual and social world of Hamilton's day and to investigate important questions about literary practices, letter writing and everyday language in Georgian England, among others.

The digital edition incorporates c. 3,000 items displayed in high-resolution images (12,700 images) alongside text transcriptions (in diplomatic and normalised format) and basic metadata: 2,951 pieces of private correspondence, 38 diaries and travel journals, and 11 manuscript books of various kinds. Over half of these, c.1,600 items and 16 diaries, have been transcribed with a detailed level of XML mark-up following TEI guidelines (P5). It covers, for instance, structural elements of the text in the body; manuscript features like underlining, additions and deletions; editorial intervention like footnotes and non-modern spelling or sic-forms; content-based mark-up such as person and place names, foreign words; and customised tags for research analysis in reading practices and salutations in correspondence. The transcribed materials have furthermore been tagged for part of speech (CLAWS7) and semantic categories (USAS) (c. 1m tokens). We have also constructed a 'personography' database containing biographical data of all writers and addressees, as well as nearly everyone mentioned in the material that has been transcribed (c. 2,530 individual persons). Designed with multiple fields and with links to external authority files such as VIAF, this database is a core tool for the study of social networks.

All in all, the rich selection of data and the diversity of editorial practices in *The Mary Hamilton Papers* will facilitate an exciting new wave of linguistic, literary and historical studies in the late Georgian period (see Coulombeau *et al.* in prep.).

References

Corpus of Early English Correspondence Extension. See: Nevalainen, Terttu, Minna Palander-Collin & Tanja Säily (eds.). 2018. *Patterns of change in 18th-century English. A sociolinguistic approach*. Amsterdam/Philadelphia: John Benjamins.

Corpus of Irish English Correspondence. See: McCafferty, Kevin & Carolina P. Amador-Moreno 2012. *A Corpus of Irish English Correspondence* (CORIECOR): A tool for studying the history and evolution of Irish English. In Bettina Migge & Máire Ní Chiosáin (eds.), *New perspectives in Irish English*, 265–288. Amsterdam/Philadelphia: John Benjamins.

Coulombeau, Sophie, David Denison & Nuria Yáñez-Bouza. In preparation. *Mary Hamilton and her circles: An edited collection*. Manchester: Manchester University Press.

CQPweb. Lancaster server. https://cqpweb.lancs.ac.uk/

Crawley, Lisa. 2014. A life recovered: Mary Hamilton, 1756–1816. *Bulletin of the John Rylands Library* 90(2), 27–46.

Dekker, Rudolf (ed.) 2000. Egodocuments and history. Autobiographical writing in its social context since the Middle Ages. Hilversum: Verloren.

The Collected Letters of Hannah More. http://hannahmoreletters.co.uk/Letters

The Elizabeth Montagu Letters. http://www.elizabethmontagunetwork.co.uk

The Mary Hamilton Papers (c.1740–c.1850). Compiled by David Denison, Nuria Yáñez-Bouza, Tino Oudesluijs, Cassandra Ulph, Christine Wallis, Hannah Barker and Sophie Coulombeau, University of Manchester. https://doi.org/10.48420/21687809

van der Wal, Marijke & Gijsbert Rutten (eds.). 2013. *Touching the past. Studies in the historical sociolinguistics of ego-documents*. Amsterdam/Philadelphia: John Benjamins.

Bilingual corpora and hybrid text production: Assisted writing, translation and post-editing (I)

Chair: Noelia Ramón García – *University of León*

Participants: Marlén Izquierdo – *University of the Basque Country*, Belén Labrador – *University of León*, Leticia Moreno – *University of Valladolid*, Noelia Ramón García – *University of León*.

The ACTRES Research Group (Contrastive Analysis and Translation English-Spanish: https://actres.unileon.es/wp/) was set up at the University of León over 20 years ago now by Dr. Rosa Rabadán Álvarez, thus founding a leading group in the field of corpus linguistics in Spain. Today the ACTRES team includes researchers from several different Spanish universities: León, Valladolid, País Vasco, Deusto and Universidad Europea Miguel de Cervantes. In addition, researchers from different international institutions have also been part of the ACTRES team at some point in time or are still collaborators, including colleagues from the Universities of Ottawa (Canada), Bergen, (Norway), Brighton and Surrey (UK).

The original line of research, which was focused on English-Spanish contrastive analysis and translation, has over time divided into several different lines of research with the advent of new technologies, the compilation of new types of corpora and the challenges raised by new projects. Our long-term experience with the compilation and mining of different types of corpora has yielded not only numerous PhD theses and research papers, but also a number of different computerized applications designed to meet the needs of various potential users. This attempt to bridge the gap between corpus- based contrastive research and real-life needs in multilingual globalized contexts is the main aim of the current research line of the ACTRES Research Group.

First, we will outline a general description of the evolution of the ACTRES Research Group over the past 20 years: the background, aims, methods and results obtained until now. We will also introduce the list of the main corpora compiled, both parallel and comparable, different tools built ad hoc for specific research needs (a rhetorical tagger and a browser) and applications developed within the framework of the ACTRES project, including a translation post-editing tool (PETRA 2.0) and 15 text generators to assist native speakers of Spanish in the technical writing of promotional texts in the food and drink industry.

We will describe the compilation process of the bidirectional parallel corpus PACTRES 2.0. This corpus contains over 5 million running words of contemporary English and Spanish texts and their corresponding translations into the other language. It includes texts from five different registers: fiction, non-fiction, newspapers, magazines and miscellanea. PACTRES provided the empirical data to feed the first application derived from the ACTRES Research Group: the post-editing programme PETRA.

Another stage in the long-term ACTRES project has been the compilation of specialised comparable corpora of online promotional texts. This change of focus from a parallel to a comparable corpus was prompted by the need to assist professionals in writing specialized texts necessary to sell their products worldwide. The final result was a set of generators of promotional texts as writing aids for non-native speakers of English. Current ACTRES research is focusing on new challenges in building a controlled natural language, particularly in promotional texts in the food and drink industry. Capitalizing on previous multi-layer analysis including POS and semantic information, rhetorical and pragmatic tagging, we will comment upon the treatment of multi-word expressions (MWE) for the construction of a bilingual text production environment.

In this round table we will provide an outline of the know-how of the ACTRES research team for successful transfer of descriptive linguistic data to real-life applications.

P-ACTRES 2.0.: A bidirectional parallel corpus for joint-contrastive-translation research

Marlén Izquierdo – *Universidad del País Vasco*

Keywords: *ACTRES Parallel Corpus, (English-Spanish) joint-contrastive-translation studies, empirical description, functionalism.*

Research conducted within the ACTRES Project may be safely credited for contributing to the development of empirical translation studies as a scientific discipline (Rabadán, 2008). First, it has always been conducted from a corpus approach, hence warranting the authenticity of the data analysed. Second, it has unfolded in accordance with an underlying integral point of view, connecting the three branches of Translation Studies, namely, product-, process-, and function-oriented research (Holmes 1972). As such, Rabadán's seminal theoretical revision of the concept of Translation Equivalence (1991) brought to light the need for descriptive research that would reveal translation equivalents, on the basis of which several applications have been developed.

This part of our round table focuses on the development and use of the ACTRES Parallel Corpus, today P-ACTRES 2.0. (Sanjurjo-González & Izquierdo, 2019), a tool aimed to make descriptive studies robust and functionally meaningful. The first version of P-ACTRES 2.0. (Izquierdo, Hofland & Reigem, 2008) furthered comparable-corpus-based contrastive analyses and enabled joint-contrastive-translation studies, thus instantiating not only what ACTRES was truly conceived of as, but also what for (Rabadán, 2002).

We will first comment upon the compilation (2008) and extension (2019) processes, discussing the paramount importance of a linguist-engineer alliance. Without the needs and wants of the former and the practical know-how of the latter, it would have been daunting for corpus linguistics to take off in empirical research the way it did. In this regard, P-ACTRES is a clear example of cross-disciplinary research efforts. We will also identify key criteria for corpus building such as intended purposes of use, availability of texts, digitisation and programming. The first aspect plays a role as both cause and consequence in the case of the ACTRES Parallel Corpus. In fact, P-ACTRES came to being to bridge a gap identified in previous comparable-corpus-based contrastive analyses (Labrador de la Cruz, 2000; Ramón García, 2003). The issue of availability has always triggered hot debates around the ethics involved in corpus compilation; we will first recount our experience with the possibility of using copyrighted materials, and then move on to today's institutional requirements to pass an ethics board. This building criterion is bound up with a key feature of corpus linguistics, namely, representativeness. We will explain this aspect attending to both quantitative and qualitative aspects. Any collection of authentic language use to be considered a corpus must be digital and apt for electronical support (Sinclair, 1991). We will revise the early steps of digitisation of P-ACTRES, which were shaped precisely by challenges posed by availability, with critical eyes. Finally, programming the corpus is paramount for sound corpus-based research. At this stage, it is the linguists' role to specify what corpus techniques a powerful corpus should be equipped with: primarily POS-tagging that enables advanced searches, then basic -yet essential- browsing functionalities such as KWIC concordances, keyword lists, cluster searches, etc.

Bibliography

Holmes, James. 1972. The name and nature of translation studies. In L. Venuti (Ed.). *The Translation Studies Reader*, 1st ed. London: Routledge, 172–185.

Izquierdo, Marlén., Hofland, Knut. & Reigem, Øystein. 2008. The ACTRES parallel corpus: an English-Spanish translation corpus. *Corpora* 3(1): 31–41.

Labrador de la Cruz, Belén. [2000] 2005. *Estudio contrastivo de la cuantificación inglés-español*. León: University of León.

Rabadán, Rosa. 2008. Refining the idea of "applied extensions". In A. Pym, M. Shlesinger and D. Simeoni, (eds.) *Beyond Descriptive Translation Studies. Investigations in homage to Gideon Toury*. Amsterdam/Philadelphia: John Benjamins, 103-117.

Rabadán, Rosa. 2002. Análisis Contrastivo y traducción inglés-español: El programa ACTRES. In *NuevasPerspectivas de los Estudios de Traducción,* J.M. Bravo & P. Fernández (eds.), 35-55. Valladolid: University of Valladolid.

Rabadán, Rosa. 1991. Equivalencia y Traducción. Problemática de la equivalencia translémica inglés-español. León: University of León.

Ramón, Noelia. 2003. Estudio contrastivo inglés-español de la caracterización de sustantivos. León: University of León.

Sanjurjo-González, Hugo and Izquierdo, Marlén. 2019. P.ACTRES 2.0. A parallel corpus for cross-linguistic research. In: I. Doval & M. Sánchez Nieto (eds.) *Parallel Corpora for Contrastive and Translation Studies*. Amsterdam/Philadelphia: John Benjamins. 215-232. https://doi.org/10.1075/scl.90.13san.

Sinclair, John. 1991. *Corpus, concordance, collocation.* Oxford: OUP

ACTRES comparable corpora and text generators

Belén Labrador – *Universidad de León*

Keywords: *comparable corpora, promotional texts, text-generators, second-language writing, English-Spanish contrast.*

This part of the general presentation on the history of ACTRES aims at describing the process of compilation of specialized comparable corpora of different text types and the subsequent design of generators of texts as writing aids for non-native speakers of English. The corpora were tagged syntactically (part-of-speech) and rhetorically - following Swales' (2004) move-step model.

In the first stage, we worked on a variety of texts from several specialized fields which were problematic and therefore interesting from a contrastive perspective: biomedical abstracts, culinary recipes (Rabadán et al 2016), electronic product descriptions (Labrador et al 2014), meeting minutes (Pizarro 2017), wine-tasting notes (López-Arroyo 2016), technical reports (Ramón & Labrador 2015), directors' reports (Rabadán et al 2021), audit reports, instruction manuals for household appliances (Cristobalena 2014), football match reports (Díez Fernández 2009), opinion articles (Pérez Blanco 2020), clinical trials, and promotional texts for rural houses. In the second stage, we focused on online promotional texts related to the food industry: cheeses, herbal teas, wines, bakery and pastry, and dried meats.

The corpora enabled us to identify cross-linguistic differences between English and Spanish at different levels: lexis, grammar, and rhetorical structure. These detailed analyses gave rise to building applications addressed to Spanish-speaking professionals who need to write texts in English for specific purposes. The resulting text-generators (Labrador & Ramón 2020; Moreno Pérez & López Arroyo 2021; Pérez Blanco & Izquierdo 2021) guide the users through the process of second-language writing with tips, pop-up menus, built-in glossaries, in-complete sentences for them to fill in and examples extracted from the corpora. The underlying assumption is that the user has a working knowledge of English and can write their texts in English, especially if they are pro-vided with prompts and assistance at every step of the writing process. This computer-assisted writing process will increase the quality of the final text, which will have a better impact on the readers, as it will follow linguistic conventions characteristic of that specific text type, thus ensuring readability.

Our late interest in food-related online promotional texts sprang from the pragmatic richness of these types of texts, which reflect descriptive and persuasive functions. Also, the food industry is one of the most important business sectors in our region, with small and medium-sized manufacturing companies looking for expanding their markets and increasing their sales worldwide; hence the need for promoting their products in English on their websites. The pilot projects of the generators were tried by professionals in the field, who informed us of their needs and offered feedback, which was implemented in the final version of the generators. Finally, the generators were registered as intellectual property by the University of León.

References

Cristobalena, A. 2014. Text Generator: An Aid for Writing in the Tertiary EST Class. *CORELL: Computer Resources for Language Learning* 4: 27-41.

Díez Fernández, M. A. 2009. Análisis Contrastivo Inglés-Español De Las Crónicas Futbolísticas En La Prensa Escrita. Tesis Doctoral. León: Universidad de León.

Labrador, B.; N. Ramón; H. Alaiz and H. Sanjurjo-González. 2014. Rhetorical structure and persuasive language in the subgenre of online advertisements. *English for Specific Purposes* 34: 38-47. DOI: 10.1016/j.esp.2013.10.002.

Labrador, B. and N. Ramón. 2020. Building a second-language writing aid for specific purposes: Promotional cheese descriptions. *English for Specific Purposes* 60: 40-52. DOI: 10.1016/j.esp.2020.03.003

López Arroyo, B. and R. P. Roberts. 2016. Differences in Wine Tasting Notes in English and Spanish. *Babel* 62(3): 370-401. DOI: 10.1075/babel.62.3.02lop.

Moreno Pérez, L. and B. López Arroyo. 2021. Atypical Corpus-Based Tools to the Rescue: How a Writing Generator Can Help Translators Adapt to the Demands of the Market. *Monti 13(1)*: 251-279. DOI: 10.6035/MonTI.2021.13.08

Pérez Blanco, M. and M. Izquierdo. 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*. Amsterdam/Philadelphia: John Benjamins. 148-173. DOI: 10.1075/btl.158.06per.

Pérez Blanco, M. 2020. Epistemic adjectives in English and Spanish journalistic opinion discourse. *Journal of Pragmatics*, 170, 112-124.

Pizarro, I. 2017. A corpus-based analysis of genre-specific multi-word combinations: Minutes in English and Spanish. In: Egan, T. and D. Hildegunn (eds.) *Cross-Linguistic Correspondences. From lexis to genre*. 221-252. ISBN 978-9027259561.

Rabadán, R. V. Colwell and H. Sanjurjo-González. 2016. BiTeXting your food: Helping the gastro industry reach the global market. In: Moreno Ortiz, A. and C. Pérez-Hernández (eds.) *CILC2016 (EPiC Series in Language and Linguistics*, *vol.1*): 361-371.

Rabadán, R.; I. Pizarro and H. Sanjurjo-González. 2021. Authoring support for Spanish language writers: A genre-restricted case study. *RESLA, Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics* 34 (2) pp. 677–717. DOI: 10.1075/resla.19048.rab.

Ramón N. and B. Labrador. 2015. The rhetorical structure of technical brochures: A proposal for technical writing. *Procedia* 173: 241-245. DOI: 10.1016/j.sbspro.2015.02.059.

Swales, J. 2004. *Research genres. Explorations and applications*. Cambridge: Cambridge University Press DOI:10.1017/CBO9781139524827.

Building a CNL for the food and drink industry: The challenge of multiword expressions

Leticia Moreno-Pérez – *University of Valladolid*

Keywords: *MWE, multiword expression, CNL, food and drink industry.*

After devoting 20 years to the mining of parallel and comparable corpora, moving from translation to the creation of writing aid tools, the ACTRES Research Group is now working to build a Controlled Natural Language (CNL). Using the thorough corpus-based mining emerged from the creation of the writing generators, the Group is now in the process of developing a CNL addressed to the food and drink industry, with a view to help companies in the sector promoting their products with a more sophisticated tool.

Although this phase could be considered a paradigm shift in the career of the Research Group, the truth is that ACTRES had already been working on a CNL, as the language of the writing generators fulfills all the criteria to be considered a CNL according to Kuhn's definition (2014: 3): "[a] controlled natural language is a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax, and/or semantics while preserving most of its natural properties".

Even though the suitability of the characteristics of a CNL varies broadly "[d]epending on application area, environment, and goal" (Kuhn, 2014: 12), and the CNL of the writing generators unarguably fulfilled its purpose as a functional tool that greatly helps in the communication process, it is restricted in terms of what users can produce. A more sophisticated CNL would allow users to write free texts in the field, rather than choosing from a catalogue of sentence patterns.

In part, the freedom and sophistication required to take the writing generators one step further entails a better management of multiword expressions (MWEs). The reason for that is that MWEs comprise a relevant part of language (Jackendoff, 1997; Mel'cuk, 1998; Sinclair, 2000; Gray and Biber, 2015, among others) and, since they are used to "express precisely ideas and concepts that cannot be compressed into a single word" (Villavicencio et al, 2005: 2), if MWEs are too scarce or poorly managed, there is some information that might not be expressed neither properly nor naturally. This becomes especially relevant in the case of CNLs for specialized purposes, since terminological units are mostly MWEs (Justeson and Katz, 1995).

At this stage, the ACTRES projects have (partially) overcome some relevant challenges related to the management of MWEs in NLP, as are the identification of MWEs – i. e. "the task of determining individual occurrences of MWEs in running text" (Baldwin and Kim, 2010: 280); labeling of MWEs – terminology in the ACTRES corpora is already tagged into four dimensions: part-of-speech, semantic, rhetorical and pragmatic tagging; and the equivalence of MWEs in the two languages of the project, English and Spanish – terms in both languages are already matched.

Nevertheless, there remain some challenges ahead, as narrowing down what a MWE is in the language of promotional texts in the food and drink industry; retrieving common patterns for MWEs in the language of promotional texts in the food and drink industry; refining tagging with a new perspective, as the Group now has a new aim and resulting IT environment; planning and building the multi-layer architecture of the CNL; expanding automatic MWE tagging; and automatic entity recognition.

References

Baldwin, T. and Kim, S. N. 2010. Multiword Expressions. In N, Indurkhya and F. J, Damerau (Eds.). *Handbook of Natural Language Processing* (2nd ed., pp. 267-292). Boca Raton, USA: CRC Press.

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

Jackendoff, R. 1997. *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.

Justeson, J., and Katz, S. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, 1(1),* 9-27.

Kuhn, T. 2014. A survey and classification of controlled natural languages. *Computational Linguistics, 40(1)*, 121-170.

Mel'cuk, I. 1998. Collocations and Lexical Functions. In A. P. Cowie (Ed.). *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press.

Sinclair, J. 2000. Lexical Grammar. *Darbai Ir Dienos 24*, 191-205.

Villavicencio, A., Bond, F., Korhonen, A., and McCarthy, D. 2005. Editorial: Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language, 19(4),* 365–377.

Bilingual corpora and hybrid text production: Assisted writing, translation and post-editing (II)

Noelia Ramón García – *Universidad de León*

As the chair of the session, I will provide an overview of the main objectives of the ACTRES Research Group (Contrastive Analysis and Translation English-Spanish; URL: actres.unileon.es), describe briefly the most important bilingual corpora compiled by our team, and list the main applications developed using the findings obtained.

The ACTRES Research Group was set up at the University of León over 20 years ago now by Dr. Rosa Rabadán Álvarez. The initial aim was to carry out empirical studies in contrastive linguistics with the working languages English and Spanish. The results of these contrastive studies were then used to explore differences between original and translated language. More recently, the project began to focus on specialized texts and genres, and the ACTRES team developed *ad hoc* tools for cross-linguistic research (taggers and browsers), and also online applications addressed to Spanish-speaking professionals working in different domains. A parallel research line has exploited contrastive results to develop a tool to improve translation (post)editing. All the tools and applications developed by the ACTRES research team have been registered for intellectual property protection.

All the studies carried out within the ACTRES research group have made use of corpus data from its inception. The need to access large amounts of translated Spanish prompted the compilation of our English-Spanish parallel corpus P-ACTRES 2.0 in the year 2004. This corpus contains now nearly 6 million words of original texts in English with their corresponding translations into Spanish, as well as original texts in Spanish with their translations into English (Sanjurjo-González & Izquierdo 2019). Dr. Izquierdo is going to describe this parallel corpus in detail.

The availability of this large parallel corpus enabled the ACTRES research team to develop the application known as PETRA 1.0 for translation post-editing (Ramón & Gutiérrez-Lanza 2018). This tool has been designed to detect differences between original and translated Spanish, which may be due to the fact that the texts have been translated from English. The application also suggests more idiomatic options available in original Spanish.

A few years later, a new research line was taken up by the ACTRES research team: prompted by the needs arising in a globalized world, the focus was placed on specialized texts and genres in different professional fields. Academia meets the industry. In Spain, in particular small and medium-sized companies, often require linguistic services to promote their products internationally, and English is the *lingua franca* for trade today. Comparable corpora in English and Spanish have been compiled in different domains for this purpose. Dr. Belén Labrador is going to describe this line of research as well as the writing aids derived from it (Labrador et al. 2014, Labrador & Ramón 2020).

More recently, the ACTRES research team has focused on the construction of a controlled natural language for promotional texts in the food and drink industry. This current project is still in progress and aims at improving the pre-edition phase of documents for the international promotion of products and services, in English and Spanish. Dr. Leticia Moreno will comment on this line of research, in particular on the area related to how multiword expressions (MWEs) need to be tackled in the construction of a controlled natural language.

References

Labrador, B., N. Ramón, H. Alaiz and H. Sanjurjo-González. 2014. Rhetorical structure and persuasive language in the subgenre of online advertisements. *English for Specific Purposes* 34: 38-47. DOI: 10.1016/j.esp.2013.10.002.

Labrador, B. and N. Ramón. 2020. Building a second-language writing aid for specific purposes: Promotional cheese descriptions. *English for Specific Purposes* 60: 42-52. DOI: 10.1016/j.esp.2020.03.003.

Ramón, N. and C. Gutiérrez-Lanza. 2018. Translation description for assessment and post-editing: The case of personal pronouns in translated Spanish. *Target* 30(1): 112- 136. DOI: 10.1075/target.15098.ram.

Sanjurjo-González, H. and M. Izquierdo. 2019. P.ACTRES 2.0. A parallel corpus for cross-linguistic research. In: I. Doval & M. Sánchez Nieto (eds.) *Parallel Corpora for Contrastive and Translation Studies*. Amsterdam/Philadelphia: John Benjamins. 215-232. DOI: 10.1075/scl.90.13san.

De la teoría a los datos y de los datos a la teoría: aplicaciones estadísticas en lingüística de corpus

Chair: Javier Pérez-Guerra – *Universidade de Vigo*

Participants: Pascual Cantos Gómez – *University of Murcia*, Daniela Pettersson-Traba – *Complutense University of Madrid*, Yolanda Fernández-Pena – *Universidade de Vigo*, Iván Tamaredo Meira – *Complutense University of Madrid*, David Tizón-Couto – *University of Vigo*.

En respuesta al creciente interés en los métodos de investigación empírica y cuasi- experimental en el ámbito de la investigación lingüística y de la lengua, esta sesión temática reúne presentaciones sobre el protagonismo y la aplicación de técnicas estadísticas en datos textuales proporcionados por corpus. Dirigidas a una audiencia no especializada en estas técnicas, las presentaciones propondrán retos reales de análisis y soluciones proporcionadas por la estadística. La sesión se organiza de este modo:

Pascual Cantos Gómez "Análisis estadístico multivariable". El hecho de que muchas preguntas de investigación lingüística son demasiado complejas para ser abordadas mediante técnicas estadísticas univariadas justifica la creciente demanda de métodos cuantitativos más sofisticados. La estadística multivariable permite el análisis y la interpretación del comportamiento de múltiples variables de interés, asociadas a un mismo fenómeno y del que se dispone de un gran número (conglomerado) de observaciones. Lejos de ofrecer una presentación exhaustiva de todas las técnicas multivariantes, se demostrará la contribución que estos recursos pueden ofrecer a los estudios lingüísticos, especialmente a aquellos basados en corpus, a la vez que se sugerirán otros métodos multivariantes emergentes.

David Tizón Couto "Regresiones". Los modelos estadísticos nos ofrecen una imagen representativa del fenómeno lingüístico que estudiamos. A través de métodos multifactoriales podemos perseguir una línea confirmatoria o exploratoria en nuestra investigación. Un método válido para perseguir ambas líneas es la regresión, que consiste en una extensión de la estadística correlacional simple a situaciones en las que se explora el comportamiento de una variable de respuesta (a menudo el efecto de un escenario hipotético de causa-efecto) con respecto a cómo varía en función de múltiples variables predictoras (a menudo las causas en ese escenario hipotético de causa-efecto). Se presentarán ejemplos en los que el uso de la regresión ha resultado útil: (a) regresión lineal para la estimación de efectos sobre la longitud del constituyente inicial en dislocaciones a la izquierda en inglés moderno, (b) regresión logística para la estimación de efectos sobre variantes de pronunciación reducidas del inglés norteamericano contemporáneo y

(c) regresión logística Bayesiana para la estimación de efectos sobre una alternancia morfosintáctica en diferentes variedades del inglés contemporáneo.

Iván Tamaredo Meira "Árboles de inferencia condicional y bosques aleatorios". En la actualidad, las técnicas de análisis predictivo se han convertido en herramientas fundamentales en la lingüística de corpus. Mientras que muchas de estas técnicas estadísticas tienen un largo recorrido en lingüística (por ejemplo, la regresión logística binaria, ya empleada en sociolingüística en los años 70 bajo el nombre de "variable rules" o análisis Varbrul), otras son aún poco conocidas por la comunidad investigadora debido a su todavía reciente introducción a la disciplina. Nos centraremos en dos técnicas no paramétricas de análisis predictivo: los árboles de inferencia condicional y los bosques aleatorios. Estas técnicas son especialmente útiles en situaciones en las que el tamaño de la muestra es pequeño pero el número de variables independientes alto, así como en aquellos casos en los que existen numerosas interacciones entre los factores estudiados.

Javier Pérez Guerra "Agrupando datos con sentido (estadístico)". Con frecuencia, un estudio de corpus

requiere la sistematización de observaciones mediante la simplificación de tendencias de coaparición o de exclusión, esto es, de semejanza o de contraste. Daremos cuenta de esto mediante dos estudios de casos en los que se aplica, respectivamente, el análisis de factores del conocido análisis multidimensional de Douglas Biber con el fin de agrupar registros en un periodo de la historia del inglés, y el Behavioral Profiles, que proporciona clústers de niveles tras la sistematización de los valores de las variables como vectores más o menos cercanos/alejados en el "espacio estadístico", aplicado al agrupamiento de tipos de texto según sus características sistémico-funcionales.

Yolanda Fernández Pena "Fenogramas". Se explorará el uso de NeighborNets para la representación visual de fenogramas, un tipo de red filogenética basada en una matriz de distancia en la que las relaciones entre los diferentes niveles de una variable se representan mediante una mayor o menor distancia entre dichos niveles, sin imponer ninguna jerarquía entre los mismos. A modo de ejemplo, se empleará NeighborNets para analizar la distribución de fragmentos en inglés contemporáneo en los diferentes registros representados en el componente británico del *International Corpus of English* (ICE-GB).

Daniela Pettersson Traba "Métodos estadísticos para el análisis de colocaciones". Desde la mitad del siglo XX, el análisis de colocaciones ha desempeñado un papel fundamental en el campo de la lingüística en áreas tan diversas como la semántica léxica, la enseñanza y el aprendizaje de lenguas y la lingüística computacional, entre otros. No obstante, en las últimas décadas han surgido nuevas técnicas que nos permiten realizar análisis de colocaciones que van más allá de las simples medidas de asociación. Entre ellas se encuentran los llamados modelos de espacios vectoriales semánticos (SVS por sus siglas en inglés) y las redes colocacionales, dos técnicas que se pueden utilizar de forma complementaria.

Análisis estadístico multivariable

Pascual Cantos-Gómez – *Universidad de Murcia*

Palabras clave: *estadística multivariante, análisis multidimensional, análisis factorial, análisis de conglomerados y análisis discriminante lineal.*

El hecho de que muchas preguntas de investigación lingüística son demasiado complejas para ser abordadas mediante técnicas estadísticas univariadas justifica la creciente demanda de métodos cuantitativos más sofisticados. La estadística multivariable permite el análisis y la interpretación del comportamiento de múltiples variables de interés, asociadas a un mismo fenómeno y del que se dispone de un gran número (conglomerado) de observaciones. La estadística multivariada analiza dos o más variables de interés que se correlacionan entre sí en diferentes grados. Su objetivo principal es estudiar cómo se relacionan las variables entre sí y cómo interactúan juntas para distinguir entre los casos en los que se realizan las observaciones. Estas estadísticas se aplican en cualquier disciplina de investigación y se usan para clasificar datos, determinar posibles taxonomías de agrupación de datos o probar hipótesis sobre un conjunto de variables. El uso de estos métodos estadísticos nos permite analizar conjuntos complejos de datos, ya que reduce la dimensionalidad de estos, con las ventajas que ello conlleva: mejor comprensión del modelo de clasificación final y un aumento en la eficiencia y eficacia del modelo en sí. Además, hay disponible un amplio abanico de aplicaciones estadístico-informáticas que permiten aplicar estos métodos, pero es necesario asegurarse cuándo y qué estadística multivariada usar, ya que no siempre son adecuadas ni aplicables a determinados conjuntos de datos.

Lejos de ofrecer una presentación exhaustiva de todas las técnicas multivariantes, se demostrará la contribución que estos recursos pueden ofrecer a los estudios lingüísticos, especialmente a aquellos basados en corpus. Compararemos dos técnicas de estadística multivariada, Análisis Factorial (AF) y Análisis de Componentes Principales (ACP), además de describir el Análisis de Conglomerados (AC) como una técnica de clasificación exploratoria útil en la investigación lingüística.

Referencias bibliográficas

Berber Sardinha, T. y Veirano Pinto, M., ed. 2019. *Multi-Dimensional Analysis: Research Methods and Current Issues*. Londres: Bloomsbury.

Biber, D. 1988. *Variation across Speech and Writing*, Cambridge: Cambridge University Press.

Biber, D. 1995. Dimensions of Register Variation: A Cross-linguistic Comparison. Cambridge: Cambridge University Press.

Biber, D., S. Conrad y R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Cantos Gómez, P. 2013. Statistical Methods in Language and Linguistic Research, Sheffield: Equinox.

Cantos Gómez, P. 2014. "Sketching a 'Low-Cost' Text-Classification Technique for Text Topics in English". *Ibérica,* 27: 165-84.

Cantos Gómez, P. 2019. "Multivariate Statistics Commonly Used in Multi-dimensional Analysis". En *Multi-Dimensional Analysis: Research Methods and Current Issues*, eds. T. Berber Sardinha y M. Veirano Pinto, 97-124. Londres: Bloomsbury.

Gries, S. T. y A. Stefanowitsch. 2010. "Cluster Analysis and the Identification of Collexeme Classes". En *Empirical and Experimental Methods in Cognitive/Functional Research*, eds. S. Rice and J. Newman, 73-90, Stanford: CSLI Publications.

Oakes, M. 1998. *Statistics for Corpus Linguistics*, Edimburgo: Edinburgh University Press.

Parodi, G. 2005. "Lingüística de corpus y análisis multidimensional: exploración de la valoración en el corpus PUCV-2003". *Revista española de lingüística*, *35*(1), 45-76.

Randolph, K. A. y L. L. Myers. 2013. *Basic Statistics in Multivariate Analysis*, Oxford: Oxford University Press.

Venegas, R. 2015. "Caracterización multidimensional del Corpus del Español PUCV-2006". En *Géneros académicos y géneros profesionales. Accesos discursivos para saber y hacer,* ed. G. Parodi, 121-146. Valparaíso: Ediciones Universitarias de Valparaíso.

Agrupando datos significativamente: análisis de correspondencias y fenogramas

Yolanda Fernández-Pena – *Universidade de Vigo*

Palabras clave: *análisis de correspondencias*, *fenogramas*, *redes filogenéticas*, *fragmentos*, *colectivos.*

En esta intervención nos centraremos en la agrupación de datos mediante dos técnicas estadísticas: el análisis de correspondencias y los fenogramas. El análisis de correspondencias se emplea para la identificación sistemática de relaciones entre variables entre las que no se asume ningún tipo de jerarquía. Utilizando la distancia chi-cuadrado como medida de proximidad/alejamiento de las categorías, permite visualizar las principales tendencias en varias dimensiones (Levshina 2015: 369; Brezina 2018: 200; Schützler 2022: 260). Se explorará también el uso de NeighborNet para la representación visual de fenogramas, un tipo de red filogenética basada en una matriz de distancia en la que las relaciones entre los diferentes niveles de una variable se representan mediante una mayor o menor distancia entre dichos niveles, sin imponer ninguna jerarquía entre los mismos.

En primer lugar, se hará una breve introducción a cada una de las técnicas estadísticas en cuestión. Para ejemplificar su uso, se emplearán dos bases de datos diferentes. El análisis de correspondencias se llevará a cabo como parte de un estudio de corpus que explora la distribución de las estrategias de reconstrucción de las estructuras fragmentarias en inglés contemporáneo en los diferentes registros representados en el componente británico del *International Corpus of English* (ICE-GB; Nelson et al. 2002) (Pérez-Guerra & Fernández-Pena forthc.). Por estrategia de reconstrucción entendemos el proceso mediante el cual las/os hablantes reconstruyen el significado proposicional de un fragmento (p.ej. *Well done to Giles*), aquel que es equivalente al de la oración completa correspondiente, apoyándose en su conocimiento de la construcción (p.ej. 'say sth to sb', Say *well done to Giles*), el contexto lingüístico inmediato o el contexto extralingüístico. El análisis de correspondencias tratará de determinar si hay estrategias de reconstrucción que son más características de algún/os registro/s (p.ej. monólogo, diálogo, ficción, informativo).

Por otra parte, se empleará el análisis de redes filogenéticas para analizar la variación diatópica de la concordancia verbal de número con frases colectivas complejas (p.ej. *a bunch/group of* N). Para ello, se empleará un estudio de corpus, con datos de seis variedades nativas de inglés extraídos del *Corpus of Global Web-based English* (GloWbE; Davies 2013), en el que se analizó la concordancia verbal de número de 23 nombres colectivos que toman complementos preposicionales introducidos por la preposición *of* (Fernández-Pena 2020). El análisis de redes filogenéticas permitirá determinar y representar gráficamente las relaciones de similitud y la distancia entre las variedades del inglés investigadas en función del grado de variación en la concordancia de número que admiten. Dado que la longitud de las ramas de la red filogenética es proporcional a la distancia lingüística (Bryant & Moulton 2004; Szmrecsanyi 2012), la proximidad entre varias variedades indicará similitudes en los patrones de concordancia.

Referencias

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A practical guide.* Cambridge University Press.

Bryant, David & Vincent Moulton. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2). 255-265.

Davies, Mark. 2013. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE). https:// www.english- corpora.org/ glowbe

Fernández-Pena, Yolanda. 2020. Reconciling synchrony, diachrony and usage in verb number agreement with complex collective subjects. New York: Routledge.

Levshina, Natalia. 2015. *How to do Linguistics with R.* Amsterdam & Philadelphia: John Benjamins.

Nelson, Gerald, Sean Wallis & Bas Aarts. 2002. Exploring natural language: Working with the British Component of the International Corpus of English. Amsterdam & Philadelphia: John Benjamins.

Pérez-Guerra, Javier & Yolanda Fernández-Pena. Forthc. An ecology of fragmentary constructions in English: A corpus-driven cognitive categorisation.

Schützler, Ole & Julia Schlüter. 2022. *Data and methods in Corpus Linguistics: Comparative approaches*. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt. 2012. Typological profile: L1 varieties. In Bernd Kortmann & Kerstin Lunkenheimer (eds.), *The Mouton atlas of variation in English*, 826-843. Berlin: Mouton de Gruyter.

Agrupando datos con sentido (estadístico)

Javier Pérez-Guerra – *Universidade de Vigo*

Palabras clave: *análisis multimensional, clúster, registro, sistémico-funcional, factor.*

Con frecuencia, un estudio de corpus requiere la sistematización de observaciones mediante la simplificación de tendencias de coaparición o de exclusión, esto es, de semejanza o de contraste. Daremos cuenta de esto mediante dos estudios de caso en los que se aplica, respectivamente, el análisis de factores del conocido análisis multidimensional de Douglas Biber con el fin de agrupar registros en un periodo de la historia del inglés, y el Hierarchical Agglomerative Cluster Analysis, que proporciona clústeres de niveles tras la sistematización de los valores de las variables como vectores más o menos cercanos/alejados en el 'espacio estadístico', aplicado al agrupamiento de tipos de texto según sus características sistémico-funcionales.

En primer lugar, la metodología del análisis multidimensional (MDA) ha sido empleada en Yáñez-Bouza y Pérez-Guerra (en prensa) en el estudio de cinco subregistros dentro del dominio del discurso científico: filosofía (humanidades), historia (ciencias sociales), ciencias de la vida y astronomía (ciencias naturales) y textos médicos. Con datos del Coruña Corpus of Scientific Writing y del corpus Late Modern English Medical Texts, se realiza un análisis multidimensional de un millón de palabras del inglés científico del siglo XVIII que permite escalar los cinco subregistros a lo largo de dos dimensiones principales de variación: discurso 'Implicado/Interpersonal versus Narrativo/Abstracto' (véase, a modo ilustrativo, gráfico 1) y 'Complejo/Elaborado versus No elaborado'. El análisis confirma, en primer lugar, que existen diferencias sustanciales entre los subregistros en términos de distribución y presencia de los rasgos lingüísticos distintivos y, en segundo lugar, que la fluctuación en el discurso en prosa es una característica general de la escritura científica del inglés moderno tardío.

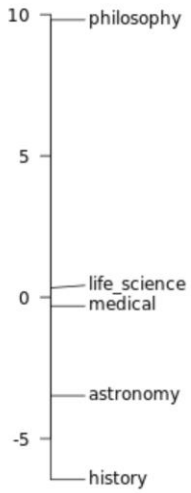

Gráfico 1. Análisis multidimensional

En segundo lugar, en Lingüística Sistémico-Funcional (SFL), la elección del elemento inicial de la cláusula o 'Tema' se ha reivindicado como indicador de registro, género o tipo de texto. En Pérez-Guerra (2021) ponemos a prueba esta premisa utilizando un análisis a gran escala basado en un corpus de Temas en el inglés americano escrito de nuestros días. El análisis incluye muestras de quince registros, dirigidos a diferentes audiciencias, con diferentes propósitos comunicativos y rasgos estilométricos. Se ponen en tela de juicio dos enfoques principales de Tema: la definición de 'primer elemento (ideacional)' de la cláusula (Halliday y Matthiessen 2014) y la hipótesis

'preverbal' (Berry 1995), según la cual el Tema se extiende, respectivamente, bien hasta el primer elemento ideacional o bien hasta el verbo. Cada uno de los Temas identificados en el corpus según estas definiciones se tipifica según su función sintáctica y estatus sistémico-funcional (textual, interpersonal, experiencial). La agrupación de registros basada en Hierarchical Agglomerative Cluster Analysis, la cual utiliza información lingüística sobre la categoría Tema, revela que el enfoque de 'primer elemento' (véase gráfico 2) es una métrica de disimilitud plausible para los registros, lo que demuestra que el concepto de Tema en SFL puede tomarse como un predictor de la categorización del registro.

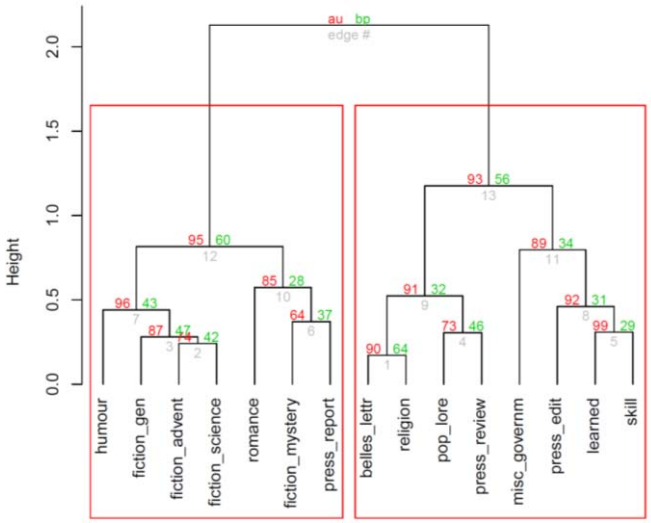

Gráfico 2. Hierarchical Agglomerative Cluster Analysis

Referencias bibliográficas

Berry, Margaret. 1995. Thematic options and success in writing. En Mohsen Ghadessy ed, *Thematic development in English texts*. Londres: Pinter, 55–84.

Halliday, Michael A. K. & Matthiessen, Christian M.I.M. 2014. *Halliday's introduction to Functional Grammar*, cuarta edición. Londres: Routledge.

Pérez-Guerra, Javier. 2021. Theme as a proxy for register categorization. En Elena Seoane y Douglas Biber eds, *Corpus-based approaches to register variation*. Ámsterdam: John Benjamins, 85–110.

Yáñez-Bouza, Nuria y Javier Pérez-Guerra. En prensa. The history of English registers. En Joan C. Beal ed, *New Cambridge history of the English language, Volume III: Change, transmission and ideology*. Cambridge: Cambridge University Press.

Métodos estadísticos para el análisis de colocaciones

Daniela Pettersson-Traba – *Universidad Complutense de Madrid*

Palabras clave: *colocaciones*, *espacios vectoriales semánticos*, *redes colocacionales*, *sinonimia*, *cambio semántico.*

Desde mediados del siglo XX, el análisis de colocaciones (Firth 1957, Sinclair 1966) ha desempeñado un papel fundamental en el campo de la lingüística en áreas tan diversas como la semántica léxica, la enseñanza y el aprendizaje de lenguas y la lingüística computacional, entre otros. No obstante, en las últimas décadas han surgido nuevas técnicas que nos permiten realizar análisis de colocaciones que van más allá de las simples medidas de asociación. Entre estas técnicas se encuentran los llamados modelos de espacios vectoriales semánticos (*SVS* por sus siglas en inglés) y las redes colocacionales; dos técnicas que se pueden utilizar de forma complementaria.

Los SVS, que se implementaron originalmente en el ámbito de la lingüística computacional, están orientados específicamente a comparar el contexto de dos o más palabras o significados de palabras. Por esta razón, es un método particularmente útil para investigar la (dis)similitud semántica entre palabras relacionadas, como pueden ser los sinónimos, en base a sus comportamientos colocacionales. El resultado de un análisis SVS consiste en una serie de valores de distancia entre las palabras o significados analizados en función de sus colocaciones. Es decir, cuantas más colocaciones compartan las palabras objeto de estudio, más similares serán en términos semánticos. La principal desventaja del SVS es que los valores de distancia se calculan en base a un gran número de colocaciones de palabras, razón por lo que no siempre es fácil identificar exactamente dónde se encuentran las diferencias o similitudes entre ellas. En palabras de Heylen et al. (2015: 154-155) el SVS es "too much of a black box". Por tanto, es conveniente recurrir a análisis más detallados para complementar los resultados obtenidos del SVS y así poder arrojar luz sobre las similitudes o diferencias identificadas. Un análisis que se podría utilizar para este fin son las redes colocacionales (Brezina, McEnery y Wattam 2015). En resumen, este método consiste en visualizar las colocaciones más significativas en redes de colocaciones en las que estas están conectadas a las palabras examinadas mediante flechas (véase Figura 1). El tamaño de estas flechas indica el grado de asociación entre el elemento examinado y la colocación: cuanto más corta es la flecha, es decir, cuanto más cerca se encuentra la colocación de la palabra examinada dentro de la red, mayor es el grado de asociación.

Figura 1: Ejemplo de una red de colocación

Las aplicaciones de este método son muy diversas. En un estudio reciente, Baker (2017: 95-101) adaptó este método al análisis diacrónico de colocaciones de las pares de sinónimos *on* y *upon* y *round* y *around*, investigando sus colocaciones en distintos períodos de la historia de la lengua inglesa. De esta manera, Baker mostró como se podría visualizar de una manera sencilla cómo estas palabras atraían y perdían colocaciones específicas con el tiempo.

En esta intervención se ejemplificará el uso de las dos técnicas mencionadas en párrafos anteriores mediante un estudio diacrónico de corpus de tres sinónimos adjetivos pertenecientes al campo semántico del olfato en inglés americano (1810-2009): *fragrant*, *perfumed* y *scented*. Mientras que el análisis SVS proporcionará una vista panorámica del comportamiento distribucional de estos adjetivos, las redes de colocaciones servirán para realizar una investigación más cualitativa, centrándonos en sus colocaciones más prominentes en distintas épocas.

Referencias

Baker, Paul. 2017. *American and British English: Divided by a common language*. Cambridge: Cambridge University Press.

Brezina, Vaclav, Tony McEnery y Stephen Wattam. 2015. "Collocations in context. A new perspective on collocation networks". *International Journal of Corpus Linguistics* 20(2): 139–173.

Firth, John. R. 1957. "A synopsis of linguistic theory 1930-1955". En John R. Firth (ed.), *Studies in linguistic analysis*, 1–32. Oxford: Philological Society.

Heylen, Kris, Thomas Wielfaert, Dirk Speelman y Dirk Geeraerts. 2015. "Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis". *Lingua* 157: 153–172.

Sinclair, John. 1966. "Beginning the study of lexis". En Charles E. Bazell (ed.), *In memory of J.R. Firth*, 410–429. Harlow: Longman.

Árboles de inferencia condicional y bosques aleatorios

Iván Tamaredo Meira – *Universidad Complutense de Madrid*

Palabras clave: *análisis predictivo, estadística no paramétrica, árboles de inferencia condicional, bosques aleatorios, omisión de pronombres sujeto.*

En la actualidad, las técnicas de análisis predictivo se han convertido en herramientas fundamentales en la lingüística de corpus. Mientras que muchas de estas técnicas estadísticas tienen un largo recorrido en lingüística (por ejemplo, la regresión logística binaria, ya empleada en sociolingüística en los años 70 bajo el nombre de "variable rules" o análisis Varbrul; Cedergren y Sankoff 1974), otras son aún poco conocidas por la comunidad investigadora debido a su todavía reciente introducción a la disciplina. Nos centraremos en dos técnicas no paramétricas de análisis predictivo: los árboles de inferencia condicional y los bosques aleatorios (Tagliamonte y Baayen 2012). Estas técnicas son especialmente útiles en situaciones en las que el tamaño de la muestra es pequeño pero el número de variables independientes o predictoras alto, así como en aquellos casos en los que existen numerosas interacciones entre los factores estudiados. Asimismo, otra ventaja de los bosques aleatorios es que estos modelos son relativamente inmunes a problemas de colinealidad entre variables, es decir, cuando existen correlaciones entre variables predictoras. De este modo, los árboles de inferencia condicional y los bosques aleatorios, que pueden y suelen ser empleados de forma complementaria, se convierten en herramientas de gran utilidad en la lingüística de corpus, motivo por el cual se han empleado recientemente en numerosos estudios y con diferentes objetivos (Arslan, Gür y Felser 2017; Rezaee y Golparvar 2017; Fonteyn y Nini 2020; Kruger y Van Rooy 2020).

Tras una breve introducción a las técnicas mencionadas, se ejemplificará su uso a través de un estudio de corpus sobre la omisión de pronombres sujeto en tres variedades del inglés: el inglés de Gran Bretaña, el inglés de Singapur y el inglés de la India. Se analizarán los efectos de una serie de variables predictoras (así como de sus posibles interacciones) en la probabilidad de que los hablantes de estas tres variedades omitan pronombres en función de sujeto. Las variables predictoras examinadas serán (i) la accesibilidad del antecedente, (ii) si el pronombre en cuestión ocurre en una oración coordinada o no, (iii) su posición en la cláusula, (iv) el pronombre omitido específico, (v) si la cláusula en la que se omite el pronombre es una cláusula principal o subordinada, (vi) el tipo de texto (se clasificarán los textos del corpus analizado según su formalidad y medio), (vii) la clase del verbo que ocurre con el pronombre omitido (léxico, modal o auxiliar no modal) y, finalmente, (viii) la variedad del inglés (Gran Bretaña, Singapur o la India). Debido al alto número de variables predictoras y al hecho de que la omisión de pronombres sujeto no es particularmente frecuente en ninguna de las variedades examinadas, los árboles de inferencia condicional y los bosques aleatorios resultan herramientas particularmente útiles para el estudio de este fenómeno lingüístico.

Referencias

Arslan, Seçkin, Eren Gür y Claudia Felser. 2017. "Predicting the Sources of Impaired *Wh*-Question Comprehension in Non-Fluent Aphasia: A Cross-Linguistic Machine Learning Study on Turkish and German". *Cognitive Neuropsychology* 34 (5): 312–331, DOI: 10.1080/02643294.2017.1394284

Cedergren, Henrietta J. y David Sankoff. 1974. "Variable Rules: Performance as a Statistical Reflection of Competence". *Language* 50 (2): 333–355.

Fonteyn, Lauren y Andrea Nini. 2020. "Individuality in Syntactic Variation: An Investigation of the Seventeenth-Century Gerund Alternation". *Cognitive Linguistics* 31 (2): 279–308.

Kruger, Haidee y Bertus Van Rooy. 2020. "A Multifactorial Analysis of Contact-Induced Change in Speech Reporting in Written White South African English (WSAfE)". *English Language and Linguistics* 24 (1): 179–209.

Rezaee, Abbas Ali y Seyyed Ehsan Golparvar. 2017. "Conditional Inference Tree Modelling of Competing Motivators of the Positioning of Concessive Clauses: The Case of a Non-native Corpus". *Journal of Quantitative Linguistics* 24 (2-3), 89–106.

Tagliamonte, Sali A. y R. Harald Baayen. 2012. "Models, Forests, and Trees of York English: *Was/Were* Variation as a Case Study for Statistical Practice". *Language Variation and Change* 24: 135–178.

Aplicaciones de la regresión múltiple a la lingüística de corpus

David Tizón-Couto – *Universidade de Vigo*

Palabras clave: *regresión múltiple, regresión mixta, estadística y corpus.*

Los modelos estadísticos nos ofrecen una imagen representativa del fenómeno lingüístico que estudiamos. A través de métodos multifactoriales podemos perseguir una línea confirmatoria o exploratoria en nuestra investigación (cf. Agresti 2002: 212). Un método válido para perseguir ambas líneas es la regresión múltiple, que consiste en una extensión de la estadística correlacional simple (p.ej. t-test, Pearson's correlation) a situaciones en las que se explora el comportamiento de una variable de respuesta (a menudo el efecto de un escenario hipotético de causa-efecto) con respecto a cómo varía en función de múltiples variables predictoras (a menudo las causas en ese escenario hipotético de causa-efecto) (cf. Gries, to appear: 10-11).

Se presentan tres ejemplos de la aplicación de la regresión múltiple a estudios de corpus previos: (a) regresión lineal para la estimación de efectos sobre la longitud del constituyente inicial en dislocaciones a la izquierda en inglés moderno (1500-1914; p.ej. *Tom Carrew, he was an honest man*; Tizón-Couto 2017), regresión logística (o binaria) para la estimación de efectos sobre variantes de pronunciación reducidas del inglés norteamericano contemporáneo (p.ej. 'needa' para *need to*; Lorenz & Tizón-Couto 2017) y (c) regresión logística Bayesiana para la estimación de efectos sobre una alternancia morfosintáctica en diferentes variedades nativas del inglés contemporáneo (*try and*/to V_{inf}; Tizón-Couto 2022).

La regresión lineal múltiple analiza variables dependientes que se pueden medir a través de una escala continua (p.ej. la longitud de un constituyente en un corpus escrito, o la duración temporal de un elemento en un corpus oral); la regresión logística binaria modela la razón de probabilidades de que una variable binaria reciba un valor en lugar del otro (p.ej. *try and* V_{inf} vs. *try to* V_{inf}; forma completa [*trying to*] vs. forma reducida [*tryna*]).

Se explican las (a) asunciones centrales de la regresión lineal múltiple (p.ej. la asunción de 'linealidad', o la necesidad de una distribución 'normal' de los residuales; Hilpert & Blasi 2020) y logística binaria (p.ej. evitar la 'colinealidad'), (b) los conceptos de variables fijas, interacciones entre variables y variables aleatorias (*random effects*; Schäfer 2020) y (c) buenas prácticas en el uso de este método de análisis estadístico multivariable, que le pueden conferir ventajas en comparación con métodos estadísticos que presuponen la selección automática de variables (p.ej. la selección deductiva de variables; Tizón-Couto & Lorenz 2021).

Bibliografía

Agresti, Alan. 2002. *Categorical data analysis*. Hoboken, NJ: Wiley.

Gries, Stefan Th. To appear. Corpus linguistics and the cognitive/constructional endeavor. In Mirjam Fried & Kiki Nikiforidou (eds.), *Cambridge Handbook of Construction Grammar*. Cambridge: Cambridge University Press.

Hilpert, Martin & Damián E. Blasi. 2020. Fixed-effects regression modeling. In Magali Paquot & Stefan Th. Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 505-533. Springer: Cham.

Lorenz, David & David Tizón-Couto. 2017. Coalescence and contraction of V-to-Vinf sequences in American English – Evidence from spoken language. *Corpus Linguistics and Linguistic Theory*. Advance online publication. https://doi.org/10.1515/cllt-2015-0067.

Schäfer, Roland. Mixed-effects regression modeling. In Magali Paquot & Stefan Th. Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 535-561. Springer: Cham.

Tizón-Couto, David. 2017. Exploring the Left Dislocation construction by means of multiple linear regression. *Belgian Journal of Linguistics* 31: 299-325.

Tizón-Couto, David. 2022. A multivariate account of particle alternation after bare-form try in native varieties of English. *English Language and Linguistics* 26.4: 645-676

Tizón-Couto, David & David Lorenz. 2021. Variables are valuable: making a case for deductive modeling. *Linguistics* 59.5: 1279-1309.

PANELS PAPERS

***Sleep well in Småland, whether you prefer a castle or a hut*: Persuasion through patterns of *you* in tourism discourse**

Annelie Ädel, Åsa Öhqvist & Sadjad Shokoohi – *Dalarna University*

Keywords: *tourism websites; persuasion; discourse patterns; audience orientation.*

Persuasion is always at work in language, even if it is especially prominent in genres where the directive function of language is key. This is highlighted in the definition of persuasion as "those linguistic choices that aim at changing or affecting the behavior of others or strengthening the existing beliefs and behaviors of those who already agree" (Virtanen & Halmari, 2005). Contemporary studies of persuasion tend to centre on domains in the public sphere such as advertising, politics and media discourse. In our talk, we focus on tourism discourse.

The complete English-version material from Sweden's official tourism website (www.visitsweden.com), whose mission is to market Sweden as a tourist destination, is analysed qualitatively and quantitatively. All verbal text (excluding images) from the 2019 version was compiled into a specialised corpus, comprising 53,296 words and 2,673 sentences. A frequency word lists generated through Sketch Engine showed that second-person *you* was highly frequent, ranked #9 after function words such as *the*, *and* and *of*. Based on frequent 2-4- word n-grams, it was also observed that *you* was commonly included in larger patterns. This justified the decision to use all examples involving *you* as a basis for the study, where tourism discourse is considered from the perspective of audience orientation, with a special focus on extended patterns that serve a persuasive function. While *you* has been studied in tourism research, it has not been in focus and observations regarding patterns *you* is involved in have been limited to collocations. As previous work has not considered the wider rhetorical patterns of *you*, our research will contribute to a broader perspective.

In the qualitative analysis, three coders participated in an inductive process of identifying persuasive rhetorical functions in a concordance list of 456 unique examples of *you* (with duplicates removed). The findings show that the *you* examples cover a scale from relatively informational to highly persuasive. This is in line with previous research having observed that tourism discourse generally serves both an informational and a persuasive function (e.g. Bosnar-Valkovic & Jurin, 2019; Calvi, 2010; Malekina & Ivanov, 2018). The majority of the *you* examples are clearly or even highly persuasive and can be divided into categories based on their rhetorical function: Building the writer-reader relationship; Anticipating reader reactions; Imagining scenarios; Presenting options; Offering tourist identities; Presenting tourist values; and Presenting a welcoming destination. Many of the examples serve more than one of these functions, as illustrated in the title. Quantitative findings on how the rhetorical functions are distributed will be presented in the talk.

There are two distinct speaker roles involved in the material: the "guide" (who is also the writer) and the potential "visitor" (the reader). Their relationship is asymmetrical from the perspective of knowledge about the destination, but the guide does not have the power to direct the audience (unlike e.g. a teacher in an educational context), so the potential visitor needs to be persuaded to follow the advice of the guide. As persuasion is clearly linked both to the discourse type— tourism discourse—and the discourse feature—*you* and associated patterns—in this study, we may expect to find the essence of persuasion at this intersection.

References

Bosnar-Valkovic, B. & Jurin, S. 2019. Communication and manipulation strategies of travel guides presenting Croatia: a linguistic perspective. *Tourism in Southern and Eastern Europe* 5, 121-138.

Calvi, M. V. 2010. Los géneros discursivos en la lengua del turismo: una propuesta de clasificación. *Ibérica* 19, 9-31.

Malenkina, N. & Ivanov, S. H. 2018. A linguistic analysis of the official tourism websites of the seventeen Spanish autonomous communities. *Journal of Destination Marketing & Management* 9, 204-233.

Virtanen, T. & Halmari, H. 2005. *Persuasion across genres: Emerging Perspectives.* In T. Virtanen & H. Halmari (Eds.), *Persuasion across genres: A Linguistic Approach* (pp. 3- 24). John Benjamins.

Lexical variety, lexical sophistication and lexical density in EFL Spanish undergraduates: a corpus-driven study in English for Fashion.

María Adsuara Martínez (ESNE Universidad de Diseño y Tecnología)

Keywords: *lexical richness, lexical variety, lexical sophistication, lexical density, ESP.*

The current study, which follows a corpus-driven approach, aims to both quantitatively and qualitatively compare the differences observed throughout an academic course in the written productions of a group of 2nd-year students enrolled in an ESP course at a Spanish university, namely English for Fashion. Thus, the study delves into the linguistic aspects of a group of participants (N= 150) in order to identify patterns of improvement or change in their written performance, as observed at the beginning and at the end of the academic year. This will be achieved by measuring quantitatively the corpus' lexical richness, observed across different constructs, which are hierarchically sub-elements: lexical variety, lexical sophistication, and lexical density. Lexical variation will be operationalized as Guiraud, Maas' index, MATTR, MTLD, and HD-D. Lexical sophistication will be measured using LFP (Lexical Frequency Profile) and the BNC/COCA frequency list, operationalized at B2K, that is, by looking at the ratios of tokens and types which fell beyond the most frequent two thousand words. Lexical density analysis will be based on the ratio of content words to all tokens in the texts. Analyses will be performed with the following corpus analysis tools: Lexical variety measures will be calculated with Taaled (Kyle et al. 2021); lexical sophistication will be calculated with AntWordProfiler (Anthony, 2022); and lexical density will be calculated with UAM Corpus Tool (O'Donnell, 2008). Accordingly, the dataset will be coded and analyzed on account of the participants' (1) amount of different words used; (2) proportion of relatively unusual or advanced words employed and (3) ratio of content words to all tokens in the dataset. In conclusion, the current study offers a comparison in the levels of lexical richness depicted in the written performance of a group of 150 undergraduates taking into account the employment of their own linguistic resources to convey the desired message and the relationship between learners' L2 production and their domain-specific competence.

Bibliography

Anthony, Laurence. 2022. *AntWordProfiler (Version 2.0.1)* [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software.

Crossley, Scott A., Cobb, Tom & McNamara, Danielle S. 2013. Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, *41*(4), 965-981.

Kyle, Kristopher, Crossley, Scott A., & Jarvis, Scott. 2020. Assessing the validity of lexical diversity using direct judgements. *Language Assessment Quarterly.*

O'Donnell, Michael. 2008. UAM Corpus Tool. http://www.corpustool.com/index.html

Individual collocational style across genres: Corpus-based and multi-dimensional authorship analysis

Maram Al Rabie & Alison May — *Reading University*

Keywords: *style, style markers, collocation, authorship attribution, multi-dimensional analysis.*

The concept of individual style is at the heart of authorship studies, which is defined as the unconscious (Conklin and Schmitt, 2012), unique and habitual choices in the use of linguistic forms (McMenamin, 2020). Although previous research claims that stylistic features are often unconsciously used, it is also widely accepted that linguistic choices are not entirely independent of genre and register restrictions (Leech and Short, 2005), making genre and register confounding variables in stylistic approaches to authorship analysis. This means that having texts of the same type is an important consideration when investigating style for authorship attribution and identification purposes. Therefore, the primary aim of this study is to investigate the effects of genre and register variation on individual style. Drawing on the psycholinguistic concepts of lexical priming and collocation (Hoey, 2005; Conklin and Schmitt, 2012) and using a corpus-based approach (Sinclair, 1991), this study examines the stability of collocation as a 'style marker' (McMenamin, 2020) across texts of different genres: fiction and non-fictional prose. A specialised corpus has been compiled of 155 texts, amounting to 23 million words, written by six prolific Victorian authors who wrote across the two types of genre, including Anthony Trollope, Charles Dickens, George Eliot, Harriet Martineau, Henry James and Wilkie Collins. The oeuvres of these authors are fully accessible via Project Gutenberg (Project Gutenberg, 2021) and the Internet Archive (Internet Archive, 2022).

Because genre and register variation are critical factors in this study, it is essential to look at the extent to which texts are similar or different. For this purpose, the multi-dimensional analysis framework introduced by (Biber, 1988) has been adopted, and the two sub-corpora of each author (i.e. fiction against non-fiction) were compared in terms of the six dimensions of variation in English. The underlying assumption of the multi-dimensional approach to register variation is that registers differ in continuous rather than dichotomous dimensions of variation. Each dimension reflects a communicative function associated with the statistical co-occurrence of linguistic features (Biber, 1988). Results show that all authors' sub-corpora have significant differences in the first and second dimensions, while the other four dimensions' scores do not show significant differences. The authors' fictional writings are confirmed as more involved and narrative-like. In contrast, their non-fictional prose is confirmed as more informational and non-narrative. Using corpus linguistic methods, such as skip-grams (Scott, 2020), the collocational frames of both positive and negative linguistic features associated with these two dimensions are analysed to investigate their collocational patterns across texts with different dimensional scores. Ultimately, this analysis helps us understand whether individual collocational styles are influenced by the genre and register variation and which communicative functions/dimensions are more or less stylistically stable across the two genres. The implications for authorship analysis are discussed.

Bibliography

Biber, D. 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press.

Conklin, K. and Schmitt, N. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics*. 32, pp. 45-61.

Hoey, M. 2005. *Lexical priming: a new theory of words and language*. London: Routledge. Internet Archive. [online]. [Accessed on 2 August 2022]. Available from: https://archive.org/

Leech, G.N. and Short, M. 2007. *Style in fiction: a linguistic introduction to English fictional prose*. 2nd ed. Great Britain: Pearson Education.

McMenamin, G.R., 2020. Forensic stylistics: The theory and practice of forensic stylistics. In: Coulthard, M, May, A and Sousa-Silva, R (eds.) *The Routledge handbook of forensic linguistics*. London: Routledge, pp. 539-557.

Project Gutenberg. [Online]. [Accessed on 26 September 2021]. Available from: https://www.gutenberg.org

Scott, M., 2020*, WordSmith Tools* version 8, Stroud: Lexical Analysis Software. Sinclair, J.M. 1991. *Corpus concordance collocation.* Oxford: Oxford University Press.

Speech representation in the *Hansard Corpus* (1803–2005)

Marc Alexander – *University of Glasgow*

Keywords: *parliamentary discourse, Hansard, representation of speech and thought, diachronic discourse analysis.*

The 1.6 billion words of the *Hansard Corpus* 1803-2005 (Alexander & Davies 2015, the corpus itself consisting of the official reports of speeches of the UK Parliament) encompass over two centuries of changes in practice and policy, meaning the corpus is made up of complex and shifting discourse practices which complicate any analysis. In particular, the nationalisation of the enterprise in 1909 (see Vice and Farrell 2017) is said to mark the shift in the record from third-person reportage to detailed first-person representation of speech, and we therefore expect corpus results to follow this pattern. However, a corpus analysis of the full range of the text reveals that the representation of speech in *Hansard* is more complex than this. This presentation investigates this phenomenon and uses corpus results of selected markers of speech representation to describe the distribution of discourse styles across the two centuries of the corpus, complementing the analysis in Alexander 2023.

An analysis of speech verbs in Hansard shows that from the beginning of the *Hansard Corpus*, selected first person speeches are reproduced in full, and there are erratic swings from first- to third- person speech presentation (in particular during the period from the 1880s to the 1910s). Figure 1, for example, shows the changes in use of 'said' per million words, as a proxy for reported speech versus direct speech. The changes in practice during the first period (1800s-1880s) are minor but clearly shown, and then during the Fourth Series of the publication there are rapid and substantial changes in practice. The post-1910 period is rather more consistent. Using the theoretical model of Leech & Short 2007 [1981] and Semino & Short 2004, the sections heavy in the use of markers such as 'said' (the part of the corpus on the left of Figure 1) can be shown through concordance analysis to be made up of Narrative Reports of Speech Acts (NRSA) with occasional Free Direct Speech (FDS), and later sections to vary between the two modes erratically, before the post-1910 shift to FDS.

Importantly, these two types of speech representation differ in the length of text they require to represent the same stretch of speech; NRSA is far briefer than FDS, and so sections of FDS will be overrepresented in the corpus compared to those NRSA sections. Corpus analyses of parliamentary discourse (eg Alexander & Struan 2022) therefore need to take into account the shifting speech characteristics of the corpus to accurately analyse the prominence of lexical results.

The paper therefore examines the representation of speech in the *Hansard Corpus* from a corpus perspective show how speech is warranted and presented in the run of British Parliamentary reporting since 1803, and to give an illustration of how and where researchers must take changing discourse types into account in analysing long diachronic data.

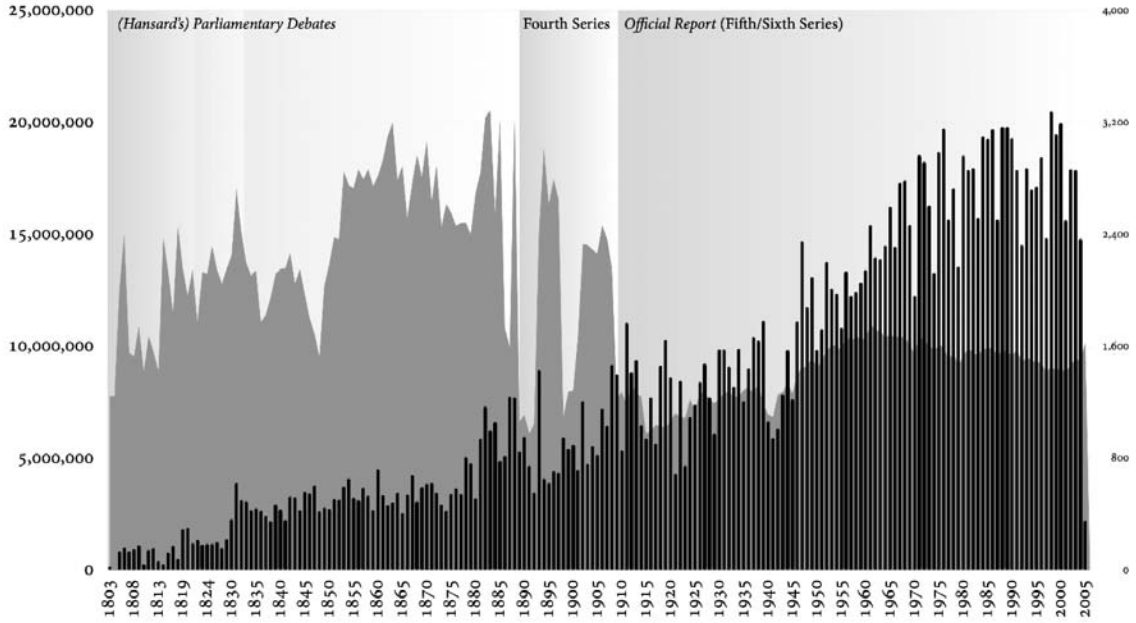

Figure 1: 'said' per million words in the *Hansard Corpus* (right axis) alongside the number of words contained in the corpus per year (left axis), and with major Series markers provided.

References

Alexander, Marc. 2023. Speech in the British *Hansard*. In: Korhonen, M., Kotze, H., and Tyrkkö, J. (eds.) *Exploring Language and Society with Big Data: Parliamentary discourse across time and space*. John Benjamins: Amsterdam.

Alexander, Marc & Andrew Struan. 2022. "In barbarous times and in uncivilized countries": Two centuries of the evolving uncivil in the *Hansard Corpus*. *International Journal of Corpus Linguistics* 27(4), pp. 480-505.

Alexander, Marc & Mark Davies. 2015. *The Hansard Corpus, 1803-2005*. Available online at http://www.english-corpora.org/hansard.

Leech, Geoffrey & Mick Short. 2007. *Style in Fiction*, 2nd edn. London: Routledge. [Reprints certain unaltered chapters from the 1981 first edition.]

Semino, Elena & Mick Short. 2004. Corpus Stylistics: Speech, Writing and Thought. Presentation in a Corpus of English Writing. London: Routledge.

Vice, John & Stephen Farrell. 2017. *The History of Hansard*. London: House of Lords Library and House of Lords Hansard.

Criteria for the selection of collocations in a plurilingual approach to phraseodidactics: A report of the procedures employed in the *PhraseoLAB* project

Moisés Almela Sánchez – *University of Murcia*

Keywords: *plurilinguistic approach, phraseodidactics, collocations, corpus linguistics, Open Education Resources.*

This presentation focuses on the procedures for selecting lexical collocations in a plurilingual approach to phraseodidactics. In line with current language education policies at European level (Council of Europe, 2001), plurilinguistic approaches promote the acquisition of a transversal and composite communicative competence in which knowledge of different languages is interconnected (Candelier et al., 2012; Cenoz & Gorter, 2013; a similar conception is prefigured in Cook, 1999). Following this framework, the European project PhraseoLAB aims at creating an Open Education Resource in which the learner's knowledge of phraseological units in English (L2) is used as a springboard for promoting phraseological competence in German (L3).

Collocations constitute one of the three categories of phraseological patterns targeted in PhraseoLAB (the other two are idioms and routine formulae). One of the main challenges for the inclusion of collocations in a phraseological database is the diversity of criteria in the literature for distinguishing collocations and syntactic/open-choice combinations. In the specialized literature, it is customary to distinguish two main approaches to the concept of collocation. With diverse names, these correspond roughly to a qualitative/semantically-oriented approach and a quantitative/frequency-based approach (Siepmann, 2005; Orlandi and Giacomini, 2016; among others).

In this paper, I will argue that, for the specific needs addressed in the PhraseoLAB database, the selection of collocations needs to follow a combination of both qualitative and quantitative criteria. Qualitative criteria are required for two reasons. The first one is that, following principles formulated in pedagogical lexicography (Benson 1989; Hausmann, 1997; Alonso Ramos, 2010; inter alia), the combinations that bear special relevance to foreign language learning are those which show an idiosyncratic restriction on one of the components (the *collocator*). The second reason for using qualitative criteria is the importance of assessing the degree of equivalence between phraseological units of a specific language pair. Chrissou (2020) proposes a scale of difficulty of phraseological units based on linguistic contrastive criteria. His proposal is informed by experiments with Greek learners of German, but a similar principle can be applied to other language pairs, such as the English (L2) - German (L3) language pair selected in PhraseoLAB. Thus, in activities designed for learners with an A2 level of German, collocational pairs with a high degree of lexical equivalence (e.g. Germ. *eine Reise buchen*/Eng. *(to) book a trip*) can be given priority over pairs with a lower degree of equivalence (e.g. Germ. *eine Auswahl treffen*/Eng. *(to) make a choice/selection*).

As Hallsteinsdóttir et al. (2006) explained, frequency of use in corpora is one of the factors that help determine the optimal phraseological material for German as a foreign language (see also Hallsteinsdóttir, 2020). This implies that qualitative criteria need to be complemented with quantitative filters. Möhring (2011) has already illustrated how the selection of German collocations for pedagogical purposes can be informed by a combination of qualitative and quantitative criteria, but his contribution is framed within a different, more lexicographically oriented context.

This presentation will show examples illustrating how the three aforementioned criteria (idiosyncratic restriction, equivalence degree, frequency in corpus) have been applied to the selection of collocations for the PhraseoLAB database. I will also outline the characteristics and the advantages of the two corpora employed in this process: the *DWDS Referenz- und Zeitungskorpora* (a freely accessible German corpus) and LEXMCI (an English corpus available in Sketch Engine).

Bibliography

Alonso Ramos, M. 2010. No importa si la llamas o no *colocación*, descríbela. In C. Mellado Blanco, P. Buján, C. Herrero-Kaczmarek, N. M. Iglesias Iglesias, A. Mansilla Pérez (Eds.), *La fraseografía del s. XXI. Nuevas propuestas para el español y el alemán* (pp. 55-80). Berlin: Frank & Timme.

Benson, N. 1989. The structure of the collocational dictionary. *International Journal of Lexicography 2*(1), 1–14.

Candelier, M. (Coord.), A. Camilleri-Grima, V. Castellotti, J-F. de Pietro, I. Lörincz, F-J. Meissner, A. Noguerol, A. Schröder-Sura 2012. *FREPA. A Framework of Reference for Pluralistic Approaches to Languages and Cultures. Competences and Resources.*

Cenoz, G. and D. Gorter 2013. Towards a plurilingual approach in English language teaching: Softening the boundaries between languages. *TESOL Quarterly 47*(3), 591–599. doi: 10.1002/tesq.121.

Chrissou, M. 2020. Sprachkontrastive Aspekte der Niveauzuordnung für den DAF- Unterricht: Hinweise aus der Unterrichtspraxis. In F. M. Mena Martínez and C. Strohschen (Eds.), *Teaching and Learning Phraseology in the XXI Century. Challenges for phraseodidactics and phraseotranslation / Phraseologie Lehren und Lernen im 21 Jahrhundert. Herausforderungen für Phraseodidaktik und Phraseoübersetzung* (pp. 117– 135). Berlin: Peter Lang.

Cook, V. 1999. Going beyond the native speaker in language teaching. *TESOL Quarterly 33*(2), 185–209.

Council of Europe 2001. Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge University Press. https://rm.coe.int/1680459f97

Hallsteinsdóttir, E. (020. Korpuslinguistische Ansätze in der Phraseodidaktik für Deutsch als Fremdsprache. In F. M. Mena Martínez and C. Strohschen (Eds.), *Teaching and Learning Phraseology in the XXI Century. Challenges for phraseodidactics and phraseotranslation / Phraseologie Lehren und Lernen im 21 Jahrhundert. Herausforderungen für Phraseodidaktik und Phraseoübersetzung* (pp. 137–157). Berlin: Peter Lang.

Hallsteinsdóttir, E., M. Šajánková and U. Quasthoff 2006. Phraseologisches Optimum für Deutsch and Fremdsprache. Ein Vorschlag auf der Basis von Frequenz- und Geläufigkeitsuntersuchungen. *Linguistik Online 27*(2), 117–136.

Hausmann, F. J. 1997. Semiotaxis und Wörterbuch. In K-P. Konerding and A. Lehr (Eds.), *Linguistische Theorie und lexikographische Praxis. Symposiumsvorträge, Heidelberg 1996* (pp. 171–179). Berlin: De Gruyter.

Orlandi, A. and L. Giacomini 2016. Introduction. In Orlandi, A. and L. Giacomini (Eds.), *Defining Collocation for Lexicographic Purposes. From Linguistic Theory to Lexicographic Praxis* (pp. 9–18). Bern: Peter Lang.

Siepmann, D. 2005. Collocation, colligation, and encoding dictionaries. Part I: Lexicological aspects. *International Journal of Lexicography 18*(4): 409–`443.

Performance vs. reader-oriented translations of theatre plays (English-Spanish): A corpus-based study on conversational markers

Olaia Andaluz-Pinedo – *University of the Basque Country & University of León*

Keywords: *theatre*, *performance-oriented translations*, *reader-oriented translations*, *prefabricated orality*, *conversational markers.*

Conversational markers (CMs) are typical linguistic features of spoken language used in the particular mode of discourse of dramatic texts, which has been termed *prefabricated orality* (Baños and Chaume 2009; Baños 2009). The cohesive and interactive functions of CMs in spoken language are mirrored in dramatic texts, recreating spontaneous orality for audiences, and thus promoting credibility and quality (Chaume 2007: 74). The presence and important role of CMs in dramatic dialogue, as well as the difficulty they entail for translation point to the relevance of further research into this area (Chaume 2004: 843; Mattsson 2009: 3; Ramón 2011: 486; Gutiérrez-Lanza in press). Nevertheless, the study of CMs in English-Spanish theatre translations seems to have received little attention.

In this contribution, we seek to describe and compare the use of CMs in performance-oriented and reader-oriented English-Spanish translations of theatre plays. Following the classification by Romero-Fresco (2009: 76-98), based on Martín Zorraquino and Portolés Lázaro (1999: 4143-4199), we will focus on two kinds of meta-discourse markers that are highly frequent in English conversation (Biber et al. 1999: 1096): the hesitation and self-repair markers (HSRMs) *I mean* and *well*, and the transition markers (TMs) *now* and *well*. Our specific aims are to (1) identify most frequent translation solutions, and (2) unveil whether there are statistically significant differences between translated and non-translated plays. To that end, we will use part of the contents of the parallel corpus TEATRAD, built *ad hoc*: three original plays in English and six translations into Spanish that have been performed (TEATRADp) or published for reading (TEATRADr) in the 21st century. The corpus amounts to 198,763 words and has been aligned at utterance level using TAligner 3.0 (Author 2021).

In order to identify and contrast the most frequent translation solutions used in each subcorpus (TEATRADp and TEATRADr), the mentioned English CMs are searched using TAligner 3.0, and concordances are analysed. Results show that in the TEATRADp subcorpus the omission of HSRM *well* (75%), and TMs *now* (84%) and *well* (57%) is particularly frequent, in line with the liberties taken by some theatre translators and the free nature of performance-oriented target texts (Author 2022). In both subcorpora, other solutions stand out: TMs *en fin* (over 40%) and *bueno* (over 15%). This recurrent use of the same translation solutions, seeming to be used by default, has also been found in subtitling (Mattsson 2009: 267).

The use of these frequent CMs in TEATRADp and TEATRADr is also compared with their use in the theatre subcorpus of CORPES XXI, the reference corpus of non-translated Spanish. Z-tests for independent proportions are run in order to see if there are statistically significant differences between translated and non-translated usage in each subcorpus. Findings reveal an overuse of TM *en fin* in translated Spanish in both subcorpora, which may be due either to interference with the STs or to the translators' effort to retain the style of ST authors. However, no statistically significant differences are observed in the use of TM *bueno* in the performance and reader-oriented translations, which favours the credibility of prefabricated orality in target texts.

Bibliography

Baños, R. 2009. La oralidad prefabricada en la traducción para el doblaje. Estudio descriptivo-contrastivo del español de dos comedias de situación [doctoral thesis]. http://hera.ugr.es/tesisugr/18319312.pdf.

Baños, R. and Chaume, F. 2009. Prefabricated orality: a challenge in audiovisual translation. In M. Giorgio Marrano, G. Nadiani and C. Rundle (eds.), *The Translation of Dialects in Multimedia. InTRAlinea. Special Issue: The Translation of Dialects in Multimedia.* http://www.intralinea.org/specials/article/Prefabricated_Orality.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Longman.

Chaume, F. 2004. Discourse Markers in Audiovisual Translating. *Meta,* (4), 843-855. https://doi.org/10.7202/009785ar.

Chaume, F. 2007. Quality Standards in dubbing: a proposal. *TradTerm,* (13), 71-89. https://doi.org/10.11606/issn.2317-9511.tradterm.2007.47466.

Gutiérrez Lanza, C. (in press). English-Spanish dubbese vs. natural pre-fabricated orality: a corpus-based study of conversational markers.

Martín Zorraquino, M. A. and Portolés Lázaro, J. 1999. Los marcadores del discurso. In I. Bosque and V. Demonte (eds.), *Gramática Descriptiva de la Lengua Española* (pp. 4051-4213). Espasa-Calpe.

Mattsson, J. 2009. The Subtitling of Discourse Particles. A corpus-based study of well, you know, I mean, and like, and their Swedish translations in ten American films [doctoral thesis]. http://hdl.handle.net/2077/21007.

Ramón, N. 2011. "Well" in Spanish translations: evidence from the P-ACTRES parallel corpus. In M. L. Carrío Pastor and M. A. Candel Mora (eds.), *Actas del III Congreso Internacional de Lingüística de Corpus: las tecnologías de la información y las comunicaciones: presente y futuro en el análisis de corpus* (pp. 485-493). Universitat Politècnica. http://hdl.handle.net/10612/9132.

Romero-Fresco, P. 2009. A corpus-based study on the naturalness of the Spanish dubbing language: the analysis of discourse markers in the dubbed translation of Friends [doctoral thesis]. http://hdl.handle.net/10399/2237.

"*He chose to kill. That's what terrorists do.*" Exploring how UK journalists use language to represent people with schizophrenia who kill as both *mad* and *bad*

James Balfour – *University of Glasgow*

Keywords: *health communication, critical discourse analysis, new discourse; mental health.*

Schizophrenia is a poorly understood illness and people with the disorder comprise one of the most vulnerable groups in society (Vick et al, 2012). Nevertheless, people with schizophrenia are represented in an inaccurate and intolerant way in the media, typically in the context of violent crime (e.g. Clement and Foster, 2008; Chopra and Doody, 2012). Such representations have the capacity to spread inaccurate beliefs about the illness among the wider public. Historically, they have even been shown to have impacted the decisions of court judges in the UK (cf. Bilton, 2003). In addition, commentators outside of linguistics have argued that people with schizophrenia who commit crimes are represented simultaneously as 'mad and bad' (Cross, 2014). However, how this is linguistically mediated has not to date been studied. To address this gap, this paper examines a corpus of ~15 million words containing all articles published in British national newspapers between 2000 and 2015 that make explicit reference to schizophrenia (e.g. *schizophrenia, schizophrenic*). The paper then examines consistent ways in which criminal acts are linguistically re-contextualised, with the view to identifying strategies used by the press to present individuals as both mad and bad. To do this, the analysis identifies collocates of the ten most frequent words referring to violent crime, and conducts a concordance analysis to determine those which suggest blame and responsibility. These are then classified according to categorisations offered by Malle et al (2014) in their theory of how readers cognitively process blame judgements. This paper focusses specifically on collocates relating to 'capacity', 'intentionality' and 'reasons'.

The resulting analysis identifies several relevant strategies. First, journalists reframe the agency of individuals through their choice of reporting verb when reporting on command hallucinations. For instance, while reporting verbs such as *drove* and *commanded* suggest to readers that actors had little agency, reporting verbs such as *told* or *wanted me to* suggest that individuals had the capacity to refuse. Second, the thought presentation of actors makes decontextualised references to desire and intention ('*the voices wanted me to kill everyone*'), thereby sometimes representing psychosis-induced thoughts as enduring character traits. This positions such people are more blameworthy. Third, logical (*because*) and temporal (*before, when, while, after, then*) conjunctions collocated with these references to violence. Logical conjunctions are used to position schizophrenia as a cause of violent crime ('*Jodi used to try to kill because of her condition*'). However, temporal conjunctions are also used to suggest blameworthy actors indirectly via Winter's (1977) logical-sequence relation. In some cases, people with schizophrenia are re-contextualised as having 'obliquely intended' (Bentham, 1979) their crimes by recklessly consuming cannabis or not taking prescribed medication. These examples show that while people withs schizophrenia are explicitly framed as ill (that is, as 'mad'), subtle linguistic strategies may be simultaneously used to indirectly frame them as intentional malevolents (that is, as 'bad'). The paper concludes by arguing that CDA research should be cautious when viewing semantic agency as tantamount to grammatical agency, and should instead engage with the psychological literature on how readers typically process blame judgements.

References

Bentham, J. [1789] 1948. *An introduction to the principles of morals and legislation*. New York: Hafner Pub.

Bilton, M. 2003. *Wicked Beyond Belief: The Hunt for the Yorkshire Ripper*. London: Harper Collins.

Chopra, A. K. and Doody, G. A. 2007. Schizophrenia, an Illness and a metaphor: Analysis of the use of the term 'schizophrenia' in the UK national newspapers. *Journal of the Royal Society of Medicine*, *100*(9), 423-426. doi:10.1177/014107680710000919.

Clement, S. and Foster, N. 200). Newspaper reporting on schizophrenia: A content analysis of five national newspapers at two time points. *Schizophrenia Research*, *98*(1), 178-183. doi:10.1016/j.schres.2007.09.028.

Malle, B., Guglielmo, S. & Monroe, A. 2014. A Theory of Blame. *Psychological Inquiry, 25*(2), 147-186. doi: 10.1080/1047840X.2014.877340.

Vick, B., Jones, K. & Mitra, S. 2012. Poverty and Psychiatric Diagnosis in the U.S.: Evidence from the Medical Expenditure Panel Survey. *Journal of Mental Health Policy and Economics*, Vol *15*(2), Retrieved via: https://ssrn.com/abstract=2330483.

Winter, E. 1977. A Clause Relational Approach to English Texts. *Instructional Science* (special edition), 6, 1-92.

From speaking madly to being madly curious: On the history of intensifying *madly*

Zeltia Blanco-Suárez – *University of Santiago de Compostela*

Keywords: *madly*, *intensifiers, corpora, subjectification, grammaticalization.*

Intensifiers or degree words have for long been a central topic of discussion in (socio)linguistic research. From the earliest monographs on intensifiers (cf. Stoffel 1901; Borst 1902; Fettig 1934), research has focused not only on the individual histories of some of these forms (cf. Adamson 2000; Méndez-Naya 2014; Breban and Davidse 2016; Claridge and Kytö 2020), but also on their use by different social groups (Macaulay 2002, 2006; Fuchs 2017; Núñez-Pertejo and Palacios-Martínez 2018), as well as on their distribution across different varieties and registers (Barnfield and Buchstaller 2010; Schweinberger 2020; Calle-Marín and Lorente-Sánchez 2021).

In line with other diachronic studies, the present paper sets out to explore the history of one intensifier, specifically that of the intensifying adverb *madly* (see examples (1) and (2) below), which has thus far remained largely unexplored (cf., however, Borst 1902 and Peters 1993).

(1) *Mr Williams's facial mobility is* ***madly*** *impressive.* (1974. OED, s.v. *madly* adv. 2b)

(2) *Yes, she was madly curious about his nocturnal gardening* (BNC2014)

In order to shed light on the origin and development of *madly*, this corpus-based study will analyse its function, as well as the types of collocates and their semantic prosody (Stubbs 1995). Moreover, the productivity and current usage of *madly* in the contemporary language will be examined.

According to the OED, the adverb *madly* was first documented in the first half of the thirteenth century. Back at that time, it was used literally, with the meaning 'in a mad or foolish manner', as shown in example (3):

(3) *Hwi motestu so* ***medliche*** ('why do you speak so madly?') (c. 1225. OED, s.v. *madly* adv. 1)

Over time, however, *madly* came to be used with non-literal and more subjective uses, which showed increasing subjectivity, since they were highly dependent on the speaker's/writer's subjective opinion (Traugott 2010). In such contexts *madly* was typically associated with passionate feelings, as shown in the collocation *madly in love* in (4):

(4) *I was so* ***madly*** *in love as to think of marrying her.* (1767. OED, s.v. *madly* adv. 2a)

These contexts eventually allowed for a potential degree reading of *madly* ('extremely, very') in the nineteenth century and its final reinterpretation as an intensifier, as in (1)-(2) above. Its productivity as an intensifier and, therefore, its grammaticalisation status in Present-day English, however, is not very advanced, as in fact it is rather collocationally restricted. The evolution of this adverb, therefore, is similar to that of other evaluative adverbs which have also developed intensifying uses over time, such as *luckily* and *strangely* (Lewis 2020).

Data for the present paper have been retrieved from a number of diachronic and synchronic sources, including the *OED*, the *Early English Books Online Corpus* 1.0 (*EEBOCorp* 1.0), *Eighteenth and Nineteenth Century Fiction*, the *Brigham Young University-British National Corpus* (BYU-BNC), and the *Spoken British National Corpus 2014* (Spoken BNC 2014).

References

Adamson, Sylvia. 2000. A lovely little example: Word order options and category shift in the premodifying string. In Fischer, Olga, Anette Rosenbach and Dieter Stein (eds.), *Pathways of change: Grammaticalization in English*, 39–66. Amsterdam and Philadelphia: John Benjamins.

Barnfield, Kate and Isabelle Buchstaller. 2010. Intensifiers on Tyneside: Longitudinal developments and new trends. *English World-Wide* 31(3): 252–287.

Borst, Eugene. 1902. *Die Gradadverbien im Englischen* (Anglistische Forschungen 10). Heidelberg: Winter.

Breban, Tine and Kristin Davidse. 2016. The history of *very*: The directionality of functional shift and (inter)subjectification. *English Language and Linguistics* 20(2): 221–249.

Calle-Martín, Javier and Juan Lorente-Sánchez. 2021. On the rise and diffusion of new Intensifiers: *This* and *that* in some Asian carieties of English. *ATLANTIS*: *Journal of the Spanish Association of Anglo-American Studies* 43(2): 47–67.

Claridge, Claudia and Merja Kytö. 2020. Degree and related phenomena in the history of English: Evidence of usage and pathways of change. *Journal of English Linguistics* 9(1): 3–17.

Fettig, Adolf. 1934. *Die Gradadverbien im Mittelenglischen* (Anglistische Forschungen 79). Heidelberg: Winter.

Fuchs, Robert. 2017. Do women (still) use more intensifiers than men? Recent change in the sociolinguistics of intensifiers in British English. *International Journal of Corpus Linguistics* 22(3): 345–374

Lewis, Diana. 2020. Speaker stance and evaluative *-ly* adverbs in the Modern English period. *Language Sciences 82*: 1–13.

Macaulay, Ronald. 2002. Extremely interesting, very interesting, or only quite interesting? Adverbs and social class. *Journal of Sociolinguistics* (3): 398–417.

Macaulay, Ronald. 2006. Pure grammaticalization: The development of a teenage intensifier. *Language Variation and Change* 18(3): 267–283.

Méndez-Naya, Belén. 2014. Out of the spatial domain: 'Out'-intensifiers in the history of English. *Folia Linguistica Historica 35*: 241–274.

Núñez-Pertejo, Paloma and Ignacio Palacios-Martínez. 2018. Intensifiers in Multicultural London English: New trends and developments. *Nordic Journal of English Studies* 17(2): 116–155.

Peters, Hans. 1993. *Die englischen Gradadverbien der Kategorie booster*. Tübingen: Gunter Narr Verlag.

Schweinberger, Martin. 2020. Analyzing change in the American English amplifier system in the fiction genre. In Rautionaho, Paula, Arja Nurmi and Juhani Klemola (eds.), *Corpora and the changing society: Studies in the evolution of English*, 223–250. Amsterdam and Philadelphia: John Benjamins.

Stoffel, Cornelis. 1901. *Intensives and downtoners: A study in English adverbs.* (Anglistische Forschungen 1). Heidelberg: Winter.

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of trouble with quantitative studies. *Functions of Language* 2(1): 23–55.

Traugott, Elizabeth Closs. 2010. (Inter)subjectivity and (inter)subjectification: A reassessment. In Davidse, Kristin, Lieven Vandelanotte and Hubert Cuyckens (eds.), *Subjectification, intersubjectification and grammaticalisation*, 29–71. Berlin and New York: Mouton de Gruyter.

Sources

BYU-BNC = *Brigham Young University-British National Corpus* (Based on the *British National Corpus* from Oxford University Press). Davies, Mark. 2004–. Available online at: http://corpus.byu.edu/bnc/.

ECF = *Eighteenth Century Fiction*. Chadwyck Healey. 1996–2021. Available online at: http://collections.chadwyck.co.uk/home/home_c18f.jsp.

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

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina and Tony McEnery. (2017). *The Spoken BNC2014*: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22(3): 319–344.

NCF = *Nineteenth Century Fiction*. Chadwyck Healey. 2000–2021. Available online at: http://collections.chadwyck.co.uk/marketing/home_c19f.jsp.

OED = *Oxford English Dictionary Online*. OUP. Available online at: http://www.oed.com/

Lexical complexity in L2 English dialogic speech

Raffaella Bottini – *Lancaster University*

Lexical complexity is a multidimensional construct which includes the sophistication, diversity and density of vocabulary produced in spoken or written communication. Measures of lexical complexity have applications in language assessment and language teaching and can inform research in language acquisition. Several indices and automatic tools have been developed to analyse lexical scores in written texts and speech transcripts (Kyle, 2019), and corpus linguistics has been the main method in this area. However, few studies have investigated lexical complexity in L2 spoken production since few large spoken learner corpora are available (Gablasova & Bottini, 2022). Also, the existing research on L2 speakers has mainly focused on the effect of proficiency, while task effects have not been fully investigated yet. This study aims at expanding prior research on lexical complexity in L2 speech, exploring dialogic tasks. It uses the 4.2-million-word Trinity Lancaster Corpus (Gablasova et al., 2019) based on the Graded Examination in Spoken English (GESE), which is administered by Trinity College London, a large international examination board. This study uses *Lex Complexity Tool* (Bottini, 2022) and adopts a mixed-method approach. The results show a significant effect of topic familiarity in dialogic speech on lexical density, diversity, and sophistication ($\eta^2 \leq .35$). Among the possible explanations for these findings, factors related to task design in the GESE, learners' educational background, real-time processing and social features of language use are discussed.

References

Bottini, R. 2022. Lexical complexity in L2 English speech: Evidence from the Trinity Lancaster Corpus. PhD thesis, Lancaster University.

Gablasova, D. & Bottini, R. 2022. Spoken learner corpora to inform teaching. In R. R. Jablonkai & E. Csomay (Eds.), Routledge handbook of corpora in English language teaching and learning. Routledge.

Gablasova, D., Brezina, V., & McEnery, T. 2019. The Trinity Lancaster Corpus: Development, description and application. International Journal of Learner Corpus Research, 5(2), 126–158.

Kyle, K. 2019. Measuring lexical richness. In S. Webb (Ed.), The Routledge Handbook of Vocabulary Studies (pp. 454–476). Routledge.

On the apostrophe in the history of English

Javier Calle-Martín & Marta Pacheco-Franco – *University of Málaga*

Keywords: *apostrophe, corpus linguistics, historical linguistic, prescription, punctuation.*

Derived from the virgule, the apostrophe was introduced in the English orthographic system by the mid-16th century as a printer's mark especially designed "for the eye rather than for the ear" (Sklar 1976: 175; Little 1986: 15). Whereas the uses of the apostrophe today are limited to the Saxon genitive construction (*the woman's book*) and to verbal contractions (*you'll* 'you will' or *you're* 'you are'), its early uses would include other cases of elision, such as the *-es* flourish, the nominative of *s*-plural nouns (*boy's* and *girl's*), the past forms of regular verbs (*lov'd* and *lik'd*) and some abbreviated words (*th'onely* 'the only' or *o'man* 'woman') along with metrical elisions (Beal 2010: 58; Parkes 1992: 55-56). Among this plethora of uses, perhaps one of the most distinctive functions of this symbol is the indication of the genitive construction, which has no full form in Present-day English after the progressive extinction of the genitive case affix. Such a development could have also unfolded for the regular past morpheme, where the apostrophe was similarly distributed. Nevertheless, the standardization of the functions of this orthographic mark over time led to a different outcome for the morpheme *-ed*, which continues to be spelled out today. The use of this symbol was already a tricky issue for 18th-century grammarians, then committed to provide apostrophic propriety out of its chaotic usage, an endeavor in which they might have been successful (Sklar 1976: 175-9).

The present paper is then concerned with the standardization of the apostrophe in the English orthographic system in the period 1500-1900 and pursues the following objectives: a) to study the use and omission of the apostrophe in the expression of the genitive case and the past tense in the period; and b) to evaluate the likely participation of grammarians in the adoption and the rejection of each of these phenomena in English. The source of evidence for this corpus-based study comes from the Early English Books Online Corpus (EEBOC) for the historical period 1500-1690 and A Representative Corpus of Historical English Registers (ARCHER) for the historical period 1700-1900, both of which sample language use in different genres and text types. A preliminary data analysis confirms the second half of the 17th century as the definite rise of the genitive apostrophe in English, then refuting the early assumptions which considered it to be an 18th-century development (Crystal 2003: 68; Lukac 2014: 3). The results also suggest that this phenomenon was to some extent associated with the decline of the apostrophe in the other environments, more particularly in the expression of the regular past tense forms. Moreover, there seems to be no indication that standardization emerged from linguistic prescription; instead, grammars seem to have been shaped after use.

References

Beal, Joan. 2010. "The Grocer's Apostrophe: Popular Prescriptivism in the 21st Century". *English Today* 26.2: 57-64.

Crystal, David. 2003. The Cambridge Encyclopedia of the English Language. Cambridge: CUP.

Little, Greta D. 1986. "The Ambivalent Apostrophe". *English Today* 2.4: 15-17.

Lukac, Morana. 2014. "Apostrophe(')s, Who Needs Them". *English Today* 30.3: 3-4.

Parkes, Malcolm B. 1992. Pause and Effect. An Introduction to the History of Punctuation in the West. London: Ashgate.

Sklar, Elizabeth S. 1976. "The Possessive Apostrophe: The Development and Decline of a Crooked Mark". *College English* 38.2: 175-183.

"Unfortunately for me became with child by him": Pregnant language in Late Modern British corpora

Nuria Calvo Cortés – *Complutense University of Madrid*

Keywords: *Late Modern Britain, pregnancy expressions, grammatical context, big and small corpora, midwifery books.*

The experience of having a child in Britain was transformed in the late 17th and early 18th century due to three main aspects, the opening of lying-in hospitals, the intervention of physicians, who often used instruments such as forceps, and the appearance of male midwives in the birth scene (Cody, 2004; Wilson, 2018). Also, this was a time with an increase in the amount of offspring being born out of wedlock in Britain (Griffin, 2013; Macfarlane, 1980). This meant that many women fell pregnant without having planned to, which had repercussions not only in their lives but also in the shame they felt when referring to their state. All these social aspects had an impact on the language employed to describe the whole process, from the time of pregnancy to the moment of the baby delivery.

The present study focuses on the analysis of some expressions used to refer to a state of pregnancy in Late Modern Britain such as *pregnant* or *with child*, and the grammatical context in which they were used. The terms were retrieved from two big corpora, the *Corpus of Late Modern English Texts Extended Version* (*CLMETEV*), *Eighteenth Century Collections Online* (*ECCO*) *TCP*, and from a small corpus of petitions signed by women who had recently had a child. In addition, midwifery books published at the time were also consulted so as to compare how the terms analysed were used by professional health carers. Similarly, the *Oxford English Dictionary* (*OED*) and the *Historical Thesaurus of English* were used to compare the syntactic context and the evolution of these expressions in the corpora and in the dictionaries. The main questions that this research poses are whether there was a preference for specific terms depending on the corpora, whether changes can be observed in relation to these preferences along time, and what could have motivated the expected variation.

The preliminary findings suggest that some terms were more frequent than others in general. For instance, the phrase *with child* was more common than the adjective *pregnant*, however, a tendency can also be observed of the latter becoming more popular as the texts advance in time. This is particularly observed in the midwifery texts. Also, some of the verbs used in combination with these terms are present in the corpora but not in the *OED*. The conclusion points to the need for more specialised corpora when analysing specific terminology, as the small corpus of petitions contains a wider variety and a much higher percentage of these expressions when compared to more general corpora. Also, lexical analyses can benefit from smaller corpora, as they allow a more qualitative research than bigger corpora, since the texts can be read in full more easily.

Bibliography

Cody, L. F. 2004. Living and dying in Georgian London's lying-in hospitals. *Bulletin of the History of Medicine*, *78*(2), 309-348.

Griffin, E. 2013. Sex, illegitimacy and social change in industrializing Britain. *Social History*, *38*(2), 139-161.

Macfarlane, A. 1980. Illegitimacy and illegitimates in English history, in P. Laslett, K. Oosterveen & R. M.Smith (eds.), *Bastardy and its comparative history* (pp. 71-85). Arnold.

Wilson, A. 2018. The making of man-midwifery: childbirth in England, 1660-1770. Routledge.

A corpus-based analysis on EFL learners' cultural vocabulary

Andrés Canga-Alonso & María Daniela Cifone-Ponte – *University of La Rioja*

Keywords: *EFL textbooks, PDLEX, culture, traditions, celebrations.*

Cultural and intercultural communicative competences (Byram, 1997) have gained prominence in EFL teaching programmes in Spanish compulsory education. However, research has not particularly focused on the cultural words included in 4th of ESO (10th grade) EFL textbooks, and the cultural vocabulary students at this level can elicit in response to a PdLex task based on three cultural stimulus words: CULTURE, TRADITIONS and CELEBRATIONS. These cue words are defined in the curriculum of La Rioja (Decree 5/2011) and the CEFR (Council of Europe, 2001) as relevant topics for the development of students' cultural awareness.

Hence, this study aims (i) to explore the cultural words from an EFL textbook (Marks & Scott, 2019) and the responses produced by 44 adolescent EFL learners to a PdLex task based on the three aforementioned prompts, and (ii) to measure whether there is an increase in cultural vocabulary elicitation. To address this second goal, the participants were tested at the beginning and the end of the academic year 2021-22. By cultural word we mean terms used "for special kinds of "things", "events" or "customs" […] that cannot be translated literally, because translation will distort its meaning" (Hapsari & Setyaningsih, 2013: 76).

Lexical units from the vocabulary output and input (Marks & Scott, 2019) were lemmatized adopting the same criteria as in Canga Alonso and Cifone Ponte (2021), Jiménez Catalán and Dewaele (2017) and Jiménez Catalán and Ojeda Alba (2009). Cultural meaning of the lexical units gathered from the lexical availability task and the textbook was checked using the Longman's Dictionary of English Language and Culture (2005). After being lemmatized, each file was subjected to frequency analyses by means of *WordSmith Tools* (Scot, 2016) to obtain the number of types and occurrences of each word.

Our findings emphasize the influence of the textbook on our informants' cultural word production, as most of the types elicited are included in the book. There is an increase in cultural word production in all the stimulus words at the end of the academic year. In addition, in the case of CULTURE, the number of students that exceeds the means in the second data collection is slightly higher than that obtained in the first. More than 50% of the informants are ahead of this value in regard with CELEBRATIONS. If we analyze word production in each prompt, all the words but two reported in CULTURE are in the textbook, and appear between 30 and 50 times. The same tendency was also found in TRADITIONS, all the words appeared in their textbook between 1 and 63 times. For CELEBRATIONS, we found more words that are not included in the textbook, but refer to a local festivity, the end of the school year and their 10th grade graduation. It is also noteworthy that certain words (e.g., birthday, Christmas, family, friend or party) are repeated in various stimulus words. This repetition may be due to the closeness in meaning of the three prompts. The connections the informants can establish as they take the test on the same day could be another reason to justify our findings.

In view of these results, we can conclude that there is a close relationship between the input of cultural vocabulary that students are exposed to and their responses to three stimulus words related to cultural aspects.

References

Canga Alonso, A & Cifone Ponte, M. D. 2021. Cultural terms in EFL textbooks for young learners. *Encuentro: Revista de investigación e innovación en la clase de idiomas* 29: 90-03.

Decreto 5/2011, de 28 de enero, por el que se establece el Currículo de la Educación Secundaria Obligatoria de la Comunidad Autónoma de La Rioja [From January 28th, secondary education curriculum in the autonomous community of La Rioja https://web.larioja.org/normativa?n=1420.

Jiménez Catalán, R. M. & Dewaele, J. M. 2017. Lexical availability of young Spanish EFL learners: emotion words versus non-emotion words, *Language, Culture and Curriculum*: 1-17 http://dx.doi.org/10.1080/07908-318.2017.1327540.

Jiménez Catalán, R. M, & Mancebo Francisco, R. 2008. Vocabulary input in EFL textbooks. *Revista española de lingüística aplicada* 21: 147-166.

Jiménez Catalán, R. M. & Ojeda Alba, J. 2009. Girls' and boys' lexical availability in EFL. *ITL Journal of Applied Linguistics* 158: 57-76.

Hapsari, N. D. & Setyaningsih, R. W. 2013: Cultural words and the translation in Twilight", *Anglicist,* 2/2: 75-81, [http://journal.unair.ac.id/download-fullpapers-anglicistd056474059full.pdf, retrieved 8th January 2023].

Marks, L. & Scott, A. 2019. *Think Ahead: ESO 4*. Cyprus: Burlington Books.

Council of Europe. 2001. Common European framework of reference for languages: learning, teaching, assessment. Cambridge University Press.

Longman Dictionary of English Language and Culture, 2nd edn. 2005. Pearson Education Limited.

Scott, M. 2016. *WordSmith Tools version 7*. Stroud: Lexical Analysis Software.

What words do not say during COVID-19 crisis:
An analysis of Pedro Sánchez's and Boris Johnson's tweets

María Luisa Carrió-Pastor – *Universitat Politècnica de València*

In this study, there are two aspects that have been considered. On the one hand, sentiment analysis shows us how we communicate (Moreno Ortíz and Pérez Hernández, 2013) and how we interact with readers, thus playing an important role in political discourse analysis. COVID-19 news was communicated in social media by most politicians with a hybrid style. On the other hand, political discourse on Twitter is of interest to many researchers as it shows the way politicians and citizens interact and negotiate meaning and news (Mancera and Pano, 2013; Coesemans and De Cock, 2017). Specifically, news on COVID-19 pandemic was communicated considering how to share facts and emotions. In this study, I focused on the COVID-19 news reported by two politicians on Twitter, Pedro Sánchez, and Boris Johnson, to compare the social and cultural context in which the interactions on Twitter occurred. I also studied the sentiment transmitted during the pandemic times on Twitter. The objectives of this analysis were, on the one hand, to identify the sentiment analysis of the tweets of both politicians during the pandemic times and, on the other, to contrast the different ways to share information and the comments received by both politicians. The corpus compiled for this study was composed of tweets on COVID-19 in the official Twitter accounts of Johnson and Sánchez during 2021 and 2022. Once compiled the corpus, the results were extracted and data was discussed with examples, contrasting Johnson and Sánchez's political discourse style during the COVID-19 pandemic. Furthermore, differences in the use of positive and negative discourse were identified and classified. Finally, conclusions were drawn.

References

Coesemans, Roel and De Cock, Barbara. 2017. "Self-reference by politicians on Twitter: Strategies to adapt to 140 characters". *Journal of Pragmatics*, 116: 37-50.

Mancera, Ana, and Pano, Ana. 2013. El discurso político en Twitter. Análisis de mensajes que "trinan". Barcelona: Anthropos.

Moreno Ortiz, Antonio and Pérez Hernández, Chantal. 2013. Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish. *Procesamiento del Lenguaje Natural*, 50: 93-100.

English-Spanish translation errors and register in non-fiction: A study on subject pronouns *tú* and *usted*

Sara Chamosa-Rabadán – *University of the Basque Country*

Keywords: *translation error, cross-linguistic interference, English-Spanish subject personal pronouns, informal-formal register.*

Translation as a third code (Frawley, 1984) has recurrently been viewed as the product of the opposing forces of cross-linguistic interference and standardization (Toury, 1995/2012). However, how to address these in quality assessment has hardly been explored. Research mostly focuses on the formal aspects of errors, linking them to a widespread negative view of interference (e.g., Presada & Badea, 2014; Wongranu, 2017). We believe, therefore, that errors should be redefined as misapplied techniques which fail to meet the parameters of the target text and culture (Rabadán & Fernández Nistal, 2002; Rabadán & Gutiérrez-Lanza, 2020). For this reason, it would be necessary to study anchor phenomena (Rabadán, 2010), i.e., specific language-pair and directionality grammatical elements which have proven to be problematic in the process of translation. The pioneering work of ACTRES members in third code has led them to discover over and underuse trends, outline their impact on contextual parameters (De Beaugrande & Dressler, 1981), and how anchors can help in translation quality assessment, as well as in post-editing tool development. Since their investigations into En-Es subject personal pronoun usage in translations mainly focus on fiction (Rabadán et al., 2007; Rabadán & Ramón, 2008; Ramón & Gutiérrez-Lanza, 2018), our paper will analyze subject personal pronouns in En-Es non-fiction translations, namely second person singular informal *tú* and formal *usted*, both of them equivalents of English *you*. We aim at: i) identifying over and underuse tendencies of translated pronouns, ii) examining errors in their usage, linking them to their impact on the target text, and iii) contrasting trends and effects in informal and formal registers.

With this in mind, we have carried out a comparative mixed methods study with data extracted from the non-fiction sub-corpora of CETRI (Corpus del Español Traducido del Inglés) and CORPES XXI (Corpus del Español del Siglo XXI). Data were compiled using a two-stage sampling process: First, to determine subject personal pronoun ratio; second, to obtain a representative population. Next, pronoun concordances were classified according to their pragmatic cross-linguistic labeling: neutral, (non-)optional emphasis, contrastive, formulaic, narrative discourse marker, and generic role (Ramón & Gutiérrez-Lanza, 2018). Data were processed through hypothesis testing to assess whether subject personal pronouns are used similarly or not in translated and untranslated non-fiction. Moreover, translated pronoun concordances were subject to qualitative analysis through error (Rabadán & Gutiérrez-Lanza, 2020) and technique (Molina & Hurtado Albir, 2002) classifications to tend to how they occur in translations. Their impact was examined following the framework laid out by Rabadán & Fernández-Nistal (2002) and Rabadán & Gutiérrez-Lanza (2020).

Quantitative findings have shown that contrastive and generic *tú* are overused and non-optional emphasis and narrative discourse marker *tú* are underused. No significant differences were discovered in the neutral, optional emphasis and formulaic uses. *Usted* overuses optional emphasis and underuses its non-optional emphasis, narrative discourse marker, and formulaic functions. Contrastive *usted* is not significantly different from its non-translated usage. No instances of the generic pronoun were located. Preliminary qualitative results have revealed that pronoun overuse favors redundancy, negatively affecting acceptability, informativity and situationality (including informal-formal tenor relationships), while underuse reduces cohesion, undermining acceptability and situationality. Non-significant differences in pronoun use promote the acceptability of translations. When completed, the analysis will: i) point out errors in pronoun translation, ii) identify affected parameters, and iii) highlight changes needed in the decision-making stage of translation.

References

De Beaugrande, R. & Dressler, W. 1981. *Introduction to Text Linguistics*. Longman. Frawley, W. (1984). Prolegomenon to a Theory of Translation. In W. Frawley (Ed.), *Translation: Literary, Linguistic, and Philosophical Perspectives* (pp. 159–178). Associated University Presses.

Molina, L. & Hurtado Albir, A. 2002. Translation techniques revisited: A dynamic and functionalist approach. *Meta*, *47*(4), 498–512. https://doi.org/10.7202/008033ar

Presada, D. & Badea, M. 2014. The effectiveness of error analysis in translation classes. A pilot study. *Porta Linguarum*, *22*, 49–59.

Rabadán, R. 2010. Applied Translation Studies. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of Translation Studies* (pp. 7–11). John Benjamins.

Rabadán, Rosa. & Fernández Nistal, P. 2002. *La traducción inglés-español: fundamentos, herramientas, aplicaciones*. Universidad de León & ITBYTE. http://hdl.handle.net/10612/5868

Rabadán, R., Gutiérrez, C., & Ramón, N. 2007, September 5. *Exploring Translation Research Applicability: Description for Assessment (ACTRES/TRACE)* [Conference contribution]. 5th International EST Congress "Why Translation Studies Matters", University of Ljubljana, Slovenia. http://hdl.handle.net/10612/9157

Rabadán, R. & Gutiérrez-Lanza, C. 2020. Developing awareness of interference errors in translation: An English-Spanish pilot study in popular science and audiovisual transcripts. *Lingue e Linguaggi*, *40*, 379–404. https://doi.org/10.1285/i22390359v40p379

Rabadán, R. & Ramón, N. 2008. Teaching English-Spanish Contrastive Analysis Through Translation. In F. Boers, J. Darquennes, K. Kerremans, & R. Temmerman (Eds.), *Multilingualism and Applied Comparative Linguistics* (Vol. 2, pp. 278–301). Cambridge Scholars Publishing.

Ramón, N. & Gutiérrez-Lanza, C. 2018. Translation description for assessment and post-editing: The case of personal pronouns in translated Spanish. *Target*, *30*(1), 112–136. https://doi.org/10.1075/target.15098.ram

Toury, G. 1995/2012. Descriptive Translation Studies - and beyond. John Benjamins.

Wongranu, P. 2017. Errors in translation made by English major students: A study on types and causes. *Kasetsart Journal of Social Sciences*, *38*(2), 117–122. https://doi.org/10.1016/j.kjss.2016.11.003

Narrative transformation from defendant examination to closing arguments in Chinese criminal trials

Yan Chen – *University of Leeds*

Keywords: *courtroom discourse, corpus-assisted discourse analysis, narrative.*

China is conducting a 'trial-centred' judicial reform, which highlights the critical role of trials to find facts, verify evidence and deliver justice. Against this backdrop, legal experts have had heated discussions about the questioning of defendants by prosecutors (e.g., Wang, 2017; Sun and Wang, 2017; Liu, 2017), criticising it for being a mere formality rather than a critical stage in the trial. This research aims to contribute a linguistic perspective by examining closing arguments in terms of their reference to the questioning stage. Chinese prosecutors' closing arguments initiate debate after defendant questioning and evidence verification, aiming to present their arguments regarding the facts, the charges, and sentencing. Most prior studies examine closing arguments independently (e.g., Carranza, 2003; Zhang, 2007; Felton Rosulek, 2015; Chaemsaithong, 2018; Bartley, 2020). Only a few look at the relationship between the preceding interaction and the closing arguments (e.g., Stygall, 1994; Cotterill, 2003; Carranza, 2003; Heffer, 2005), but they do not specifically compare the narratives constructed at different stages of a trial.

This comparative study focuses on how the narratives constructed in defendant questioning by prosecutors are represented in their closing arguments. As a corpus- assisted discourse study, I built a corpus of 49 transcribed criminal trials and arranged it in two subcorpora: the prosecutors' closing arguments (Corpus C) and prosecutors' questioning of the defendants (Corpus Q). The top bigram of Corpus C is '供 述' (defendant's depositions at the investigation stage and the testimony in court), which shows that one key issue in the closing argument is related to what the defendants said.

And the collocates of another frequent bigram '的 时候' (when) show that the defendants' responses in the questioning stage are referred to frequently.

The comparative analysis finds that, at the questioning stage, the defendants are encouraged to tell their version of the story, and prosecutors rarely confront them about the truthfulness of their testimony. However, the analysis reveals that in the closing arguments the prosecutors reconstruct a narrative by evaluating the truthfulness of the defendants' narrative, based on which defendants' honesty and their attitude to admit guilt, an important factor influencing the final sentence, are also evaluated. Therefore, by looking at how the narratives are reconstructed in the closing arguments, we have a better understanding of the prosecutors' questioning strategy and the function of the questioning stage. This research not only contributes a linguistic perspective to the discussion of the ongoing judicial reform in China, but also enriches the research on closing arguments as well as the intertextuality of stages in courtroom discourse. Methodologically, a corpus-assisted study of a relatively large amount of data affords new discoveries which are impossible in one case study.

Bibliography

Bartley, L.V. 2020. 'Please make your verdict speak the truth': Insights from an Appraisal analysis of the closing arguments from a rape trial. *Text & Talk*. 40(4), pp.421–442.

Carranza, I.E. 2003. Genre and institution: Narrative temporality in final arguments. *Narrative Inquiry*. 13(1), pp.41–69.

Chaemsaithong, K. 2018. Investigating audience orientation in courtroom communication: The case of the closing argument. *Pragmatics and Society*. 9(4), pp.545–570.

Cotterill, J. 2003. Language and power in court: a linguistic analysis of the O.J. Simpson trial. Basingstoke: Palgrave Macmillan.

Felton Rosulek, L. 2015. Dueling discourses: the construction of reality in closing arguments. New York, NY: Oxford University Press.

Heffer, C. 2005. The language of jury trial: a corpus-aided analysis of legal-lay discourse. Basingstoke: Palgrave Macmillan.

Liu, L. 2017. 刑事庭审中公诉人讯问环节是否必要？(Is the procedure of the prosecutors' questioning in criminal trials necessary?). *Legal Life News*.

Stygall, G. 1994. Trial Language: Differential discourse processing and discursive formation. John Benjamins Publishing.

Sun, C. and Wang, B. 2017. 论刑事庭审实质化的理念、制度和技术 (On the idea, system and technology of making criminal trials substantive). *Modern Law Science*. 39(2), pp.123–145.

Wang, H. 2017. 以庭审实质化视角对刑事庭审中公诉人讯问环节的考察、反思与建言 (On the prosecutor's questioning in criminal trials from the perspective of trial substantiation). *Journal of Law Application*. (1), pp.108–112.

Zhang, L. 2007. 以"评"说"法": 法庭辩论中的评价资源与实现手段 (Appraisal in law application: on the appraisal resources and ways for realisation in court debate). *Foreign Language Education*. 28(6), pp.29–33.

Patria: **De la novela a la serie. Apuntes para un estudio basado en corpus**

Luisa Chierichetti & Ivana Rota – *Università di Bergamo*

Palabras clave: *Patria – novela, Patria – serie televisiva, discurso literario, discurso telecinemático, reescritura cinematográfica, estilística de corpus.*

La presente propuesta se centra en el *best seller Patria* (2016) de Fernando Aramburu y en la exitosa reescritura que de esta novela se ha hecho para la televisión, con la homónima serie creada por Aitor Gabilondo (HBO, 2020). La trama, que gira en torno al afán de una viuda de lograr que el presunto autor del asesinato de su marido le pida perdón (Casas Olcoz 2018: 44), "procede del relato de la convivencia de nueve protagonistas pertenecientes a dos familias con raíces en un pueblo de Guipúzcoa" y se desarrolla en una época que abarca desde mediados de los ochenta del siglo pasado hasta el verano de 2012 (Aramburu 2017: 182). En la novela, la materia narrativa se reparte entre un narrador externo, los nueve personajes principales, que se expresan en primera persona, y, por último, el texto, provisto de facultad narrativa propia. Utilizando una técnica de puzle, "a cada una de las figuras de ficción le toca protagonizar por turno secuencias que en ningún caso debían superar los cuatro capítulos consecutivos; y los capítulos, a su vez, en ningún caso debían superar las ocho páginas de ordenador"; además, la trama no progresa con una linealidad cronológica (Aramburu 2017: 183).

Considerando la complejidad de la estructura de la novela, nos planteamos abordar las principales trasformaciones textuales que se han llevado a cabo en la reescritura audiovisual de esta obra literaria, considerando que en ella pueden implicarse tres operaciones distintas: la omisión de un elemento, su incorporación y la invención de un elemento original (Taranilla 2021: 183). Para ello, adoptamos la perspectiva de la estilística de corpus, aplicando las técnicas de la lingüística de corpus al estudio de textos literarios y *telecinemáticos* (Piazza, Bednarek, Rossi 2011). Esta metodología resulta especialmente valiosa en la medida en que aborda tanto la dimensión cuantitativa como la cualitativa, siendo posible observar cómo ambas se combinan, y en el hecho de que concentra su interés en textos concretos y no en generalizaciones (Nieto Caballero / Ruano San Segundo 2020: 28).

El corpus PATRIA, que creamos y gestionamos con el programa SketchEngine, suma 378.102 tokens y aproximadamente 307.084 palabras. Se compone de tres subcorpus distintos: *Patria novela*, con 227.050 tokens y aproximadamente 184.403 palabras, correspondiente al 60% del corpus; *Patria guiones*, con 114.982 tokens y alrededor de 93.385 palabras, correspondiente al 30,4% del corpus; *Patria subtítulos*, con 36.070 tokens y alrededor de 29.295 palabras, correspondiente al 9,5% del corpus. La articulación del corpus en tres subcorpus que coinciden con el texto literario de partida, en la dimensión de la creación (los guiones, textos "escritos para decirlos como algo no escrito" según Gregory y Carroll 1978) y en la dimensión del producto audiovisual (los subtítulos, texto escrito que pretende dar cuenta de los diálogos de los actores, así como de aquellos elementos discursivos que forman parte de la fotografía […] y de la pista sonora […] según Díaz Cintas 2003: 32) nos permitirá identificar los elementos que "se desvían" del texto inicial y que, por lo tanto, pueden considerarse distintivos de la serie televisiva.

Bibliografía

Aramburu, F. 2016. *Patria.* Barcelona:Tusquets.

Aramburu, F. 2017. Patria en el taller, *Grand Place. Pensamiento y cultura* n. 7: 181-187. Casas Olcoz, A. M. 2018. *El fenómeno* Patria*, de Fernando Aramburu: una nueva narrativa en torno al terrorismo vasco.* Theses and Dissertations. University of Wisconsin-Milwaukee 1770. https://dc.uwm.edu/etd/1770 (1.1.2023).

Diaz-Cintas, J. 2003. Teoría y práctica de la subtitulación. Inglés-Español, Barcelona: Ariel.

Gregory M. & Carroll, S. 1978. Language and Situation: Language Varieties and Their Social Contexts. London: Routledge/Kegan Paul.

Nieto Caballero, G. & Ruano San Segundo, P. 2020. Estilística de Corpus: Nuevos Enfoques en el análisis de Textos Literarios. Bern: Peter Lang.

Piazza, R., Bednarek, M. & Rossi, F. (eds.). 2011. *Telecinematic discourse: Approaches to the language of film and television series*. Amsterdam / Philadelphia: John Benjamins.

Taranilla, R. 2021. Trasformaciones textuales en la reescritura cinematográfica de obras literarias: el caso de *Zama*, *Cuadernos Aispi*, 18: 179-200.

Assessing factuality in a Spanish news corpus: Do experts and non- experts agree?

Elisabet Comelles, Hortènsia Curell, Irene Castellón & Juan Aparicio – *University of Barcelona*

Keywords: *corpus, factuality markers, polarity, news, Spanish.*

Event veridicality and factuality are key to natural language understanding. Deciding whether the events mentioned in a text are viewed as happening or not is one of the key features in fake news detection. That is one of the reasons why factuality has been widely studied and enormous efforts have been made to annotate corpora and develop tools that automatically identify factuality and veridicality in a text (Saurí & Pustejovsky, 2009/2012; Marneffe *et al.*, 2012; Soni *et al.*, 2014; Van Son *et al.*, 2014; Lee *et al.,* 2015; Prabhakaran *et al.,* 2015; White *et al.,* 2016; Stanovsky *et al.*, 2017; Rudinger *et al.*, 2018; Miller, 2020). Most of the studies carried out use English corpora annotated by non- experts; however, little has been written about factuality in Spanish (Wonsever *et al.*, 2016; Tracey *et al.*, 2019; Fernández-Montraveta *et al.*, 2020a; Rosá *et al.*, 2020b, Barrios, 2022) and on how appropriate the use of non-expert annotators is.

Based on the discussion above, the aims of this study are a) to shed some light on how non-expert Spanish speakers interpret the various factuality markers, and b) explore whether the experts' and non-experts' interpretations differ.

In order to carry out this research, several steps were taken. First, a Google Forms questionnaire was passed around a group of 97 non-experts, Spanish native or near-native speakers whose age ranged between 20 and 60. The questionnaire included a set of 29 sentences extracted from the TAGFACT corpus (Alonso *et al.*, 2018) together with an associated comprehension question. The participants had to answer the question by means of a 5-point likert scale, being 1-absolutely true and 5-absolutely false. The sentences included in the questionnaire were tweaked to reduce both their syntactic complexity and their semantic ambiguity. They were also anonymised to prevent informants from being biassed, given that some pieces of news included in the corpus mentioned well-known Spanish politicians. Secondly, a group of 4 expert linguists were asked to answer the questionnaire as well. In those cases where there was disagreement, the experts were asked to reach a consensus. Finally, the answers provided by both experts and non-experts were analysed quantitatively and qualitatively.

The quantitative analysis shows that there are 18 sentences with one dominant value, whereas the rest of the sentences range between 2 dominant values (4 sentences), 3 dominant values (5 sentences) and one sentence with a contradictory result. As for the agreement between experts and non-experts, although they generally seem to agree on their assessment of factuality, some differences have been identified and further analysed by means of a qualitative analysis. The results of this qualitative analysis seem to indicate that experts pay more attention to aspects related to the verb, such as its tense and its scope, whereas non-experts do not seem to be so aware of those aspects, but are rather highly influenced by the occurrence of polarity and modality markers.

References

Alonso, Laura, Castellón, Irene, Curell, Hortènsia, Fernández-Montraveta, Ana, Oliver, Sonia, and Vázquez, Glòria. 2018. Proyecto TAGFACT: Del texto al conocimiento. Factualidad y grados de certeza en español. *Procesamiento del Lenguaje Natural*, 61, 151-154.

Barrios, Leyre. 2022. La factualidad en las oraciones adversativas, concesivas y condicionales en español: El papel de los tiempos verbales en la anotación automática de corpus. PhD Thesis. Universitat de Lleida.

De Marneffe, Marie-Catherine, Manning, Christopher D, and Potts, Christopher. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational linguistics*, 38(2), 301-333.

Fernández-Montraveta, Ana, Curell, Hortènsia, Vázquez, Glòria, and Catellón, Irene. 2020a. The TAGFACT annotator and editor: A versatile tool. *Research in Corpus Linguistics*, 8(1), 131-146.

Lee, Kenton, Artzi, Yoav, Choi, Yejin, and Zettlemoyer, Luke. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1643-1648.

Miller, Ben. 2020. Reading Certainty: Evidence from a Large Study on NLP and Witness Testimony. *Digital Humanities 2020 Conference*. July 20-24, 2020 (online).

Prabhakaran, Vinodkumar, By, Tomas, Hirschberg, Julia, Rambow, Owen, Shaikh, Samira, Strzalkowski, Tomek, Tracey, Jennifer, Arrigo, Michael, Basu, Rupayan, Clark, Micah, Dalton, Adam, Diab, Mona, Guthrie, Louise, Prokof•ieva, Anna, Strassel, Stephanie, Werner, Gregory, Wilks. Yorick and Wiebeand, Janyce. 2015. A new dataset and evalua•tion for belief/factuality. In *Proceedings of the 4th Joint Conference on Lexical and Computational Semantic*, 82-91.

Rosá, Aiala, Alonso, Laura, Castellón, Irene, Chiruzzo, Luis, Curell, Hortènsia, Fernández-Montraveta, Ana, Góngora, Santiago, Malcouri, Marisa, Vázquez, Glòria, and Wonsever, Dina. 2020b. Overview of FACT at IBERLEF 2020b: Events detection and classification. In *Proceedings of the Iberian Languages Evaluation Forum*.

Rudinger, Rachel, White, Aaron Steven and Van Durme, Benjamin. 2018. Neural models of factuality. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (vol. 1, Long Papers), 731–744.

Saurí, Roser, and Pustejovsky, James. 2009. FactBank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3), 227-268.

Saurí, Roser, and Pustejovsky, James. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38(2), 261-299.

Stanovsky, Gabriel, Eckle-Kohler, Judith, Puzikov, Yevgeniy, Dagan, Ido, and Gurevych, Iryna. 2017. Integrating deep linguistic features in factuality prediction over uni_ed datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), 352-357.

Soni, Sandeep, Mitra, Tanushree, Gilbert, Eric and Eisenstein, Jacob. 2014. Modeling Factuality Judgments in Social Media Text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Short Papers), 415-420.

Tracey, Jennifer, Arrigo, Michael and Strassel, Stephanie. 2019. *DEFT Spanish Commited Belief Annotation*. Linguistic Data Consortium LDC2019T09. Web download. Philadelphia: Linguistic Data Consortium.

Van Son, Chantal, van Erp, Marieken, Fokkens, Antske and Vossen, Piek. 2014. Hope and fear: Interpreting perspectives by integrating sentiment and event factuality. In *Proceedings of the 9th Language Resources and Evaluation Conference,* 26-31.

White, Aaron Steven and Rawlins, Kyle. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, vol. 3, 221-234.

Wonsever, Dina, Aiala, Rosá and Malcuori, Marisa. 2016. Factuality Annotation and Learning in Spanish Texts. In *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference*, 2076-2080.

Morphosyntactic variation and change in modern Scottish Gaelic

Avelino Corral-Esteban – *Universidad Autónoma de Madrid*

Keywords: *Scottish Gaelic, morphosyntactic change and variation, corpus linguistics, standardization.*

The linguistic situation in Scotland is a complicated issue, as three languages, namely Scottish English, Scots, and Scottish Gaelic, are spoken within its national boundaries, influencing each other in certain regions and contexts. While there is no doubt about the dialectal diversity of Scottish English and Scots, there is an ongoing debate on whether contemporary Gaelic displays dialectal diversity (Dorian, 1981 & 2014; Gillies, 1987, 1988, 1989, 1992, 2008 & 2010; Lamb, 2011; among others) or, by contrast, the formerly existing dialects have become a unified whole now (McAuley, 1982; McInnes, 2006).

With the aim of confirming whether diatopic variation exists in Gaelic or not, this study attempts to examine 26 morphosyntactic properties of Scottish Gaelic that are considered to be undergoing change and variation (Lamb, 2002: 98): the first person plural conditional form, prepositions triggering genitive case, double genitive constructions, possession with inalienable nouns, synthetic passive, duplicated arguments, participial passive, progressive aspect with indefinite plural NPs, numerals + noun, embedded polar questions, reflexive pronoun in progressive aspect, levelling of case in aspectual distinctions, particle in wh-questions, tense agreement in comparatives, and emphatic pronoun with passive. These grammatical contexts will be analysed in two different in two types of corpus: 1) *Corpas na Gàidhlig* is an electronic corpus of 340 texts – written between 1694 and 2014, taken from a variety of genres (including poetry, prose, drama, and song) and belonging to different domains (folklore, religious, legal, and journalistic, among others); and 2) a collection of questionnaires taken by the author from native speakers living throughout Gaelic Scotland in 2022, which were carefully designed by the author in order to avoid altering the validity of the data for linguistic research.

The findings obtained from the examination of these grammatical contexts in the corpora show that there is sufficient evidence to claim that the more traditional forms in these grammatical contexts are more commonly found in older texts, especially religious and legal. The analysis also reveals that, mainly due to language shift and the institutionalization in broadcasting and in the education system, Scottish Gaelic is undergoing a high degree of levelling and convergence throughout the geographical area it is spoken, especially between Lewis, Harris, North Uist, South Uist, Barra, Grimsay, Benbecula, Islay, and Sky (above 90%) and between the Hebridean group, Fort William, and Inverness (between 85% and 90%). Finally, the results also appear to suggest that a more traditional variety of Gaelic continues to be spoken quite homogeneously in the Hebrides and that there are a number of minor dialectal forms in several other Gaelic-speaking areas, even in smaller Gaelic communities existing in the most important urban areas (Edinburgh, Glasgow, Aberdeen, Argyll, Ayrshire, Arran, etc.), that appear to show more recent innovations, probably due to the influence of Scottish English.

Bibliography

Corpas na Gàidhlig, Digital Archive of Scottish Gaelic (DASG). University of Glasgow. Available online at: https://dasg.ac.uk/corpus/

Dorian, N. 1984. *Language Death: The Life Cycle of a Scottish Gaelic Dialect.* Philadelphia: University of Pennsylvania Press.

Dorian, N. 2014. 'Defining the Speech Community to Include Its Working Margins'. In A. Y. Aikhenvald, R. M. W. Dixon, and N. J. Enfield (eds.) *Small-Language Fates and Prospects: Lessons of Persistence and Change from Endangered Languages: Collected Essays.* Leiden: Brill, 156 - 166.

Gillies, W. 1987. "Scottish Gaelic: the present situation". *International Conference on Minority Languages 3 (Celtic papers)*: 27–46.

Gillies, W. 1988. "The atlas of Gaelic dialects: an interim report". In *Scottish Gaelic Studies* 15: 1–5.

Gillies, W. 1989. "The future of Scottish Gaelic studies". In *Gaelic and Scotland*: 22–43.

Gillies, W. 1992. "Scottish Gaelic dialect studies". In C. J. Byrne, M. Harry & P. Ó Siadhail (eds.) *Celtic Languages, Celtic Peoples*: 315 - 329.

Gillies, W. 2008. "Scottish Gaelic". In M. J. Ball & N. Müller (eds.) *The Celtic Languages*: 230-304.

Gillies, W. 2010. "Studying Gaelic in the 21st Century". In K. E. Nilsen (ed.) *Rannsachadh na Gàidhlig 5 / Fifth Scottish Gaelic Research Conference* (Sydney, CB, 2010), 9-30.

Gillies, W. 2010. "Studying Gaelic in the 21st Century". In K. E. Nilsen (ed.) *Rannsachadh na Gàidhlig 5 / Fifth Scottish Gaelic Research Conference* (Sydney, CB, 2010), 9-30.

Lamb, W. 2002. *Scottish Gaelic*. Lincom Europa.

Lamb, W. 2011. "Is there a future for regional dialects in Scottish Gaelic?" *Paper presented to the FRLSU Colloquium,* 3 December 2011. (https://www.academia.edu/1136136/Is_there_a_future_for_regional_dialects_in_Scottish_Gaelic)

McAuley, D. 1982. 'Register Range and Choice in Scottish Gaelic'. *International Journal of the Sociology of Language* 35: 25-48.

McInnes, J. 2006. 'The Scottish Gaelic Language'. In M. Newton (ed.) *Dùthchas nan Gàidheal: Selected Essays of John MacInnes*. Edinburgh: Birlinn, 92 – 119.

McLeod, W. 2017. "Dialectal diversity in contemporary Gaelic: perceptions, discourses and responses'. In Cruickshank, Janet and Robert McColl Millar (eds.) 2017. *Before the Storm: Papers from the Forum for Research on the Languages of Scotland and Ulster triennial meeting,* Ayr 2015. Aberdeen: Forum for Research on the Languages of Scotland and Ireland, 183-211.

Demonstrative *them* in American English over two centuries (1820-2010)

Miriam Criado-Peña – *University of Granada*

Keywords: *American English, demonstrative determiners, demonstrative them, Late Modern English, Present-day English.*

The present paper examines the diachronic development of demonstrative *them* in 19th- and 20th-century American English. The English demonstrative system is composed of four determiners, i.e., the singular forms *this* and *that* and the plural ones *these* and *those*. The early Modern period witnessed the emergence of the alternative form *them* as a substitute for plural demonstratives, especially for distal *those* (e.g., *he don't say any of them things at all*). The use of *them* in demonstrative determiner constructions has been witnessed in a range of vernacular varieties of English, including those spoken in the United Kingdom, the United States and the Caribbean (see, for example, Beal 2004; Schneider 2004; Kirk, Hamilton & Vacovsky 2011). In standard British and American English, however, this construction is not as common today due to its high stigmatization, apparently enjoying greater popularity among working-class speakers (Pabst 2022: 136).

While the use of demonstrative *them* in some varieties of English has received some attention in the scholarly literature, no studies have focused on the diachronic development of the construction in American English. According to the *Dictionary of American Regional English* (Hall 2012), the first instance of the phenomenon is found in 1850, although the present study demonstrates that this variable dates back to previous decades. The Late Modern period is of paramount importance in the history of English since it represents a "transitional stage between the categorical innovations of Late Middle English and, especially, Early Modern English and the 'established' system of Present-day English" (Aarts, López-Couso & Méndez-Naya 2012: 870). This piece of research therefore presents results from an analysis of a selection of the most frequent demonstrative determiner constructions containing *them* in American English so as to shed some light on the diachronic evolution of this variable and the influence that some linguistic and extralinguistic factors may have had upon the choice of demonstrative over the two centuries under study on the basis of data from the *Corpus of Historical American English* (COHA). In this vein, a threefold objective is pursued: a) to analyze the historical development of demonstrative *them* in the period 1820-2010; b) to investigate its use and distribution across text types in order to assess whether text typology is related to the choice of demonstrative; and c) to ascertain the contribution of a series of linguistic conditioning factors in the use of this variable. The following linguistic factors are considered: animacy (inanimate vs. animate); proximity (distal, generic or proximal); recoverability of antecedent (recoverable vs. unrecoverable); direction of reference (anaphoric vs. cataphoric); and syntactic function (subject, object or object of a preposition).

References

Aarts, Bas, María J. López-Couso & Belén Méndez-Naya. 2012. "Late Modern English: Syntax". In Alexander Bergs and Laurel J. Brinton (eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 34.1*. Berlin and Boston: De Gruyter Mouton. 869-887.

Beal, Joan. 2004. "English dialects in the North of England: Morphology and syntax". In Bernd Kortmann, Kate Burridge, Rajend Mesthrie, Edgar W. Schneider & Clive Upton (eds.), *A Handbook of Varieties of English*. Volume 2: Morphology and syntax. Berlin and New York: Mouton de Gruyter. 114-141.

Hall, Joan H, ed. 2012. *Dictionary of American Regional English*. Volume 5. Cambridge: Harvard University Press.

Hazen, Kirk, Sarah Hamilton & Sarah Vacovsky. 2011. "The Fall of Demonstrative *them*: Evidence from Appalachia". *English World-Wide 32.1*: 74-103.

Pabst, Katharina. 2022. Putting 'the Other Maine' on the Map: Language Variation, Local Affiliation, and Co-occurrence in Aroostook County English. PhD thesis. University of Toronto.

Schneider, Edgar W. 2004. Synopsis: Morphological and syntactic variation in the Americas and the Caribbean. In Bernd Kortmann, Kate Burridge, Rajend Mesthrie, Edgar W. Schneider & Clive Upton (eds.), *A Handbook of Varieties of English*. Volume 2: Morphology and syntax. Berlin and New York: Mouton de Gruyter. 1104-1115.

A case study on corpus-based pre-trained language model: BERT-assisted automated evaluation of massive translation texts

Yizhuo Cui & Maocheng Liang – *Beihang University*

Keywords: *BERT, automated evaluation, massive translation texts.*

Introduction

In the Digital Era, pre-trained language models (PTMs) built on large corpora have revolutionized natural language processing (NLP), finding applications in diverse tasks such as image classification, automatic speech recognition, translation, sentence similarity, text classification, question answering, summarization, etc. To address the difficulties faced by the task of evaluating massive translation texts like time consuming, resources consuming, low internal consistency and low inter-rater consistency, this study intends to explore the feasibility of an automated evaluation system for massive translation texts, using BERT (Devlin et al., 2018), a contextual model. The proposed system is based on a case study of Han Suyin International Translation Contest, a large-scale and influential translation contest in China with over 30 years of history.

Research questions

The study aims to address two questions:

1.- How is the performance of the BERT-based system in evaluating massive translation texts?

2.- How can long translation texts be processed to meet the 512-token input constraint of BERT?

Corpus data

The corpus upon which this study is based comes from the 31st Han Suyin International Translation Contest, a total of 10,647 valid C-E and E-C translations submitted by the participants, and 8 reference translations provided by the experts. The C-E group includes 3,822 participants' translations and 4 experts' translations. The E-C group includes 6,825 participants' translations and 4 experts' translations.

Methods

Semantic similarity, as an important concept, is introduced to this study as translation is a kind of restricted writing which is expected to semantically present the same content with the source language (SL) text in the target language (TL). Namely, if a translated text is semantically similar to the SL text, it can be seen as a good translation. Semantic similarity is a method of computing the semantic distance between two concepts according to a given ontology (Slimani 2013: 1). It helps to determine the similarity between concepts that are not necessarily lexically similar (Petrakis et al. 2006:233). It assigns high values to pairs of words that are in a semantic relation (such as synonyms, hyponyms, free associations, etc.) while assigning low values to all the other unrelated pairs (Panchenko et al. 2018).

The overall semantic similarity between two texts (*text A* and *text B*) can be derived by computing the cosine value between their document embeddings:

$$similarity\ (A,B) = cos\ \theta = \frac{A\ \times B}{||A||\ \times\ ||B||} = \frac{\sum_{i=1}^{n} A_i\ \times B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\ \times\ \sqrt{\sum_{i=1}^{n} {B_i}^2}}$$

By introducing semantic similarity, translation quality evaluation is a task to calculate the degree of semantic similarity between a participant's translation and several experts' translations based on their BERT-generated

embeddings. The quality of a participant's translation is directly proportional to the value of semantic similarity; in other words, the higher the value of semantic similarity, the higher the quality of the translation. Conversely, if the value of semantic similarity is low, it indicates a lower quality of the translation.

Preliminary findings

Experimental results show that the BERT-assist system is a reliable second rater for massive translations in terms of translation quality. It can effectively sift out high- quality translations potentially shortlisted for prize with a reliability of $r = 0.9$ or higher. The consistency between the system-generated scores and the human-assigned scores is satisfactory, with a maximum accuracy 0.65+ (C-E) and 0.74+ (E-C); a maximum recall of 0.71+ (C-E) and 0.84+ (E-C); and a maximum F0.5-Score of 0.71+(C-E) and 0.84+ (E-C). The use of segmentation and summarization techniques provides possibilities to handle the BERT input constrain.

Bibliography

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Panchenko, A., Loukachevitch, N., Ustalov, D., Paperno, D., Meyer, C., & Konstantinova, N. 2018. Russe: The first workshop on Russian semantic similarity. *arXiv preprint arXiv:1803.05820*.

Petrakis, E. G., Varelas, G., Hliaoutakis, A., & Raftopoulou, P. 2006. X-similarity: Computing semantic similarity between concepts from different ontologies. *Journal of Digital Information Management*, *4*(4): 233-237.

Slimani, T. 2013. Description and evaluation of semantic similarity measures approaches. *arXiv preprint arXiv:1310.8059*.

Identification and assessment of linguistic features from students' writing patterns within a developmental education model

Miguel da Corte & Jorge Baptista – University of Algarve

Keywords: *developmental education, developmental education placement, language proficiency, annotation scheme, corpus for special purposes, machine-learning classification.*

Providing access to education and a path to literacy is at the core of the mission of higher education (Bickerstaff *et al.* 2021; Cormier & Bickerstaff 2019). For community colleges in the United States, this path is available through Developmental Education (DevEd) (Darkerwanld- DeCola 2021; Mazzariello & Edgecombe 2018). DevEd courses are designed, as far as English L1 is concerned, to support students' communication skills by strengthening their competencies in reading and writing.

Placement in DevEd is often based on the automated assessment and scoring of linguistic features extracted from standardized written assignments, administered as part of an entrance exam, using automatic systems such as ACCUPLACER [https://www.accuplacer.org/ Last access: March 1, 2023; all URLs in this paper were checked on this date] or COMPASS [https://www.compassprep.com/practice-tests/]. According to Hassel & Baird Giordano (2015) and Nazzal (2020), commonly used standardized exams, like the ones mentioned, demonstrate some limitations in the classification precision and portray a narrow conceptualization of the writing process. This study is concerned with more precise, extended annotation procedures (which include features derived from human annotation) aiming at improving the accuracy, in terms of level of placement, of students in DevEd. With a more precise annotation scheme, we seek to outline indicators that better showcase the writing skills and lexical diversity of college students within a DevEd model. While many studies investigate student s' literacy in their L2, we focus on the literacy skills of students in their L1 and how their native language proficiency can be assessed.

This paper focuses on students' writing skills (with English as their L1) enrolled in a two-level, sequential, DevEd course of study. Within these courses, students remediate linguistic deficiencies until they reach a language proficiency level to aptly participate in an academic program. This paper addresses the question of what orthographic, grammatical, lexical, semantic, and discursive patterns prevent developing writers from effectively communicating academically and participating in college-level (non-DevEd) courses. Some of the linguistic features captured build on patterns already reported as predictors of student placement (McNamara *et al.* 2006; Abba 2015; Baese-Berk *et al.* 2021), but we propose other, multilayered, linguistic aspects that, to the best of our knowledge, have not yet been explored and could offer promising enhancements to annotation guidelines within the context of DevEd. An annotation task of a learners' corpus of written assignments, obtained from this target population, is also outlined.

Students' placement has been construed here as a classification task and this study assesses the pertinence of the presented features. A corpus of 100 sample texts was randomly selected and balanced by level (50 for Level 1 and 50 for Level 2). Natural language processing techniques were combined with a machine-learning approach to produce the most accurate classification of students' placement in DevEd Level 1 or 2. A human classification task was also carried out to determine how human ratings (in terms of placement) correlate with those of machine learning tools. Several experiments were carried out from an established baseline, using full-text samples of diverse sizes. The best-performing (built-in) learning model, Neural Network (NN), achieved a suboptimal Classification Accuracy of 0.588. Additional experimental scenarios were devised, with same-sized samples (balanced by level of distribution) and adding linguistic data (106 features, drawn from Coh-Metrix3 [http://141.225.61.35/CohMetrix2017/]). These features do not overlap with the ones identified in this paper.

Still, the best-performing learning algorithm (again NN, but without feature selection), only improved by 2.5% against the dataset without those features. Within these experiments, no feature selection was performed.

The results obtained suggest that other, better-performing, features must be devised to improve the performance of placement systems (Goudas 2020; Perin & Lauterbach 2018; Qian et al. 2020), such as, but not limited to, the tokenization of multiword expressions (MWE), as suggested in the literature in line with, for instance, Kochmar et al. (2020). No information on the ACCUPLACER strategy for signaling and factoring MWE into the assessment process has been found. The impact of MWE on DevEd placement was investigated in a previous study by Da Corte & Baptista (2022), and the findings are briefly discussed here.

In this paper, the proposed additional linguistic features will be assessed, and the result will be utilized to 1) determine which linguistic features can better assess students' progression toward a path of language proficiency; and 2) gain a broader understanding of how community colleges can better align students' academic writing skills with the literacy demands of higher education.

References

Abba, K. A. 2015. *Community college students' writing: Lexical, syntactic, and cohesion differences in L1, L2, and Generation 1.5 students and examining knowledge of the writing process*. Ph.D. thesis, Texas AM University, Graduate and Professional Studies.

Baese-Berk, M. M., Drake, S., Foster, K., Lee, D -Y., Staggs, C., and Wright, J. M. 2021. Lexical diversity, lexical sophistication, and predictability for speech in multiple listening conditions. *Frontiers in Psychology*, 12:2328.

Bickerstaff, S. E., Kopko, E. M., Lewy, E. B., Raufman, J., & Rutschow, E. Z. 2021. Implementing and scaling multiple measures assessment in the context of COVID-19. *Community College Research Center, Teachers College, Columbia University*. Research brief.

Cormier, M. and Bickerstaff, S. 2019. Research on developmental education instruction for adult literacy learners. *The Wiley Handbook of Adult Literacy*, pages 541–561.

Da Corte. M. & Baptista, J. 2022, A Phraseology Approach in Developmental Education Placement, in Corpas Pastor, G., Mitkov, R., Kunilovskaya, M. and Caro Quintana, R. (eds.) Computational and Corpus-based Phraseology, Proceedings of EUROPHRAS 2022, Malaga, September 28-30, 2022. (pp. 79-86).

Goudas, A 2020. Measure twice, place once: Understanding and applying data on multiple measures for college placement. http://communitycollegedata.com/wp-content/uploads/2020/03/2020MultipleMeauresNOSS-PreconfWksp.pdf.

Hassel, H. and Baird Giordano, J. 2015. The blurry borders of college writing: Remediation and the assessment of student readiness. *College English*, 78(1):56–80.

Kochmar, E., Gooding, S., Shardlow, M. 2020. Detecting multiword expression type helps lexical complexity assessment. arXiv preprint arXiv:2005.05692.

Mazzariello, A., Ganga, E., and Edgecombe, N. 2018. Developmental education: An introduction for policymakers. *Education Commission of the States*. https://files.eric.ed.gov/fulltext/ED582926.pdf

McNamara, D., Ozuru, Y., Graesser, A., and Louwerse, M. 2006. Validating CoH-Metrix. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society,* pages 573–578.

Nazzal, J. S. 2020. *Writing Proficiency and Student Placement in Community College Composition Courses*. Ph.D. Thesis, University of California, Irvine.

Perin, D. and Lauterbach, M. 2018. Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education*, 28(1):56–78.

Perin, D., Raufman, J., and Kalamkarian, H. S. 2015. *Developmental reading and English assessment in a researcher-practitioner partnership*. Technical report.

Qian, L., Zhao, Y., Cheng, Y. 2020. Evaluating China's automated essay scoring system iWrite. *Journal of Educational Computing Research*, 58(4):771–790.

Entre ámbito y variedad: peculiaridades de un corpus de decretos traducidos automáticamente

Flavia De Camillis & Elena Chiocchetti – *Eurac Research*

Palabras claves: *traducción automática; anotación de errores; lenguaje jurídico; variedades lingüísticas.*

En Tirol del Sur alemán e italiano son idiomas cooficiales. El alemán oficial surtirolés es una variedad estándar (*Hochdeutsch*) y se diferencia de las otras variedades de alemán estándar (germánica, austríaca, suiza, etc.) especialmente por la terminología y la fraseología jurídico-administrativas (Ammon et al. 2016). Pese a emplearse sobre todo en la comunicación de las instituciones públicas de la provincia de Bolzano, que traducen diariamente entre alemán e italiano, aún no se ha desarrollado un sistema de traducción automática (TA) específico para esta combinación lingüística. La TA neuronal ofrece una calidad sin precedentes (Kenny 2022), sin embargo la oferta para las variedades de idiomas pluricéntricos, como el caso del alemán surtirolés, es todavía muy limitada (en DeepL se encuentran las variedades de tan solo dos idiomas: inglés y portugués), tanto como lo son los resultados de la TA para el lenguaje jurídico en general (Wiesmann 2019; Killman 2014).

Puesto que el principal rasgo identificativo del alemán surtirolés es la terminología jurídico- administrativa, anteriormente se llevaron a cabo experimentaciones de TA enfocadas en el ámbito jurídico-administrativo surtirolés. En concreto, se entrenó un sistema neuronal basado en textos jurídicos y los resultados demostraron como principal fallo precisamente la traducción de los términos jurídico-administrativos (Contarino 2021, Autor 1 2021). La *domain-adaptation* llevada a cabo no solventó del todo el problema terminológico, de por sí ya notorio (Heiss y Soffritti 2018). Partiendo de los estudios previos, nuestro objetivo consiste en identificar los errores del sistema neuronal entrenado siguiendo una línea de investigación ya consolidada (Popović 2018; Castilho et al. 2021). Se trata de una tarea esencial en la medida en que la coexistencia de términos homónimos, concurrentes y sinónimos en el lenguaje jurídico –a los que se añaden en Tirol del Sur términos oficiales (normados), no oficiales y obsoletos– representa una de las principales dificultades para la desambiguación semántica.

En esta comunicación, presentamos las principales categorías de errores detectadas en un corpus bilingüe (DE-IT), que consta de 52 decretos provinciales (ca. 60.000 palabras). El sistema usado para la traducción (ModernMT) se entrenó con un corpus de 200.000 segmentos bilingües, alcanzando 71,22 (DE>IT) y 74,74 (IT>DE) puntos BLEU contra los 26,65 y 27,59 respectivamente en su versión base. La anotación de los errores se ejecutó mediante una taxonomía repartida en errores de precisión (*accuracy*) y de fluidez (*fluency*) y adaptada de Tezcan et al. (2017). Los errores detectados pertenecen sobre todo al área léxica y destacan por frecuencia los errores de precisión de tipo *bilingual terminology, word sense disambiguation* y *semantically unrelated*. Estas tres categorías han creado mucho desacuerdo entre las cuatro anotadoras. En esencia, no ha sido siempre posible separar netamente los términos inadecuados para su uso en Tirol del Sur de los términos simplemente inapropiados por su significado, como en el caso de *Pflicht/Obligation/Verpflichtung* (obligación), de *Ziffer/Punkt/Nummer* (punto, como elemento inferior al coma) o de *Zuschuss/Beitrag* (contribución).

Los resultados de nuestro análisis permiten intuir que otro entrenamiento basado en listas terminológicas del sistema TA previamente entrenado no sería probablemente suficiente para solventar los problemas terminológicos detectados en el corpus, ya que la ambigüedad está directamente ligada no solo al ámbito y a la variedad lingüística, sino también al cotexto. Concluimos considerando necesario desarrollar técnicas para integrar en los sistemas de TA información detallada sobre el cotexto (a nivel de frase, párrafo y documento) junto con la información terminológica.

Bibliografía

Ammon, Ulrich, Hans Bickel, y Alexandra N. Lenz (eds). 2016. Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen. Berlin: De Gruyter.

Autor 1. 2021. [modificado para garantizar anonimato]

Castilho, Sheila, João Lucas Cavalheiro Camargo, Miguel Menezes, y Andy Way. 2021. «DELA Corpus - A Document-Level Corpus Annotated with Context-Related Issues». En Proceedings of the Sixth Conference on Machine Translation (WMT), 566-77.

Association for Computational Linguistics.

Contarino, Antonio. 2021. «Neural machine translation adaptation and automatic terminology evaluation: a case study on Italian and South Tyrolean German legal texts». Master's thesis, Bologna: Università di Bologna.

Heiss, Christine, y Marcello Soffritti. 2018. «DeepL Traduttore e didattica della traduzione dall'italiano in tedesco». inTRAlinea, 1-11.

Kenny, Dorothy. 2022. «Human and Machine Translation». En Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence, editado por Dorothy Kenny, 23-50. Berlin: Language Science Press. https://zenodo.org/record/6653406.

Killman, Jeffrey. 2014. «Vocabulary Accuracy of Statistical Machine Translation in the Legal Context». En Third Workshop on PostEditing Technology and Practice, editado por Sharon O'Brien, Michel Simard, y Lucia Specia, 85-98.

Popović, Maja. 2018. «Error Classification and Analysis for Machine Translation Quality Assessment». En Translation Quality Assessment: From Principles to Practice, editado por Joss Moorkens, Sheila Castilho, Federico Gaspari, y Stephen Doherty, 1:129-58. Machine Translation: Technologies and Applications. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7.

Tezcan, Arda, Véronique Hoste, y Lieve Macken. 2017. «SCATE Taxonomy and Corpus of Machine Translation Errors». En Trends in E-Tools and Resources for Translators and Interpreters, editado por Gloria Corpas Pastor y Isabel Duran-Munoz, 219-48. Approaches to Translation Studies, Volume 45. Leiden; Boston: Brill/Rodopi.

Wiesmann, Eva. 2019. «Machine translation in the field of law: a study of the translation of Italian legal texts into German». Comparative Legilinguistics 37: 117-53. https://doi.org/10.14746/cl.2019.37.4.

Assessing aggressive/impolite-related language toward Meghan Markle

Mª Milagros del Saz Rubio – *Polytechnic University of Valencia*

Keywords: *impolite/aggressive language, appraisal theory, technologically-mediated communication, corpus-based approach.*

The present study is concerned with a multimodal analysis of the aggressive/impolite-related discourse directed towards the controversial figure of Meghan Markle after her entry into the British Royal Family, her wedding to Prince Harry, and their decision to resign as senior royals and move to North America, all against the backdrop of a supposedly hospitable environment for Meghan as a bi-racial, divorced and self-proclaimed feminist (Duncan and Low, 2018). Considering that her media treatment went from pure adulation to utter censure (Yelin and Clancy 2021), I aim to analyze a small sample of the online hostility that she received on Twitter after the Netflix release of the *Harry & Meghan* documentary (16/12/2022). To do so, a reduced corpus of hostile responses in this multi-participant microblogging site prompted by the release of the documentary will be quantitatively and qualitatively analyzed through the lens of the existing models of impoliteness (Culpeper, 2005; 2016; Culpeper et al., 2003; Author in press) and Appraisal Theory (Martin & White, 2005) with a focus on the attitude system. For the purpose of this study, impoliteness is best defined as "a negative attitude towards specific behaviors occurring in specific contexts." In this vein, impoliteness is thought to be "sustained by expectations, desires and/or beliefs about social organization" (Culpeper, 2011: 23). On its part, Appraisal theory developed within the social semiotic paradigm of Systemic Functional Linguistics and provides a discourse semantic approach to evaluative language.

The study relies on a combination of quantitative and qualitative approaches (Baker et al., 2008). For the quantitative analysis, a corpus linguistics approach (Baker et al. 2008) is applied with the help of two software tools: *the Linguistic Inquiry Word Count2015 (LIWC)* (Pennebaker et al. 2015) for sentiment analysis and *Sketch Engine* (Kilgarriff et al. 2014*)* for the extraction and qualitative analysis of concordance lines that mention Meghan Markle through the @handle tool and other related labels. These replies were also analyzed using two functions of the five discursive strategies developed in the *Discourse Historical Approach* (DHA) (Reisigl and Wodak 2001, 2016), i.e., nomination and predication. Some tentative results point to the fact that the hostile language deployed seems to arise due to judgments towards the capacity and morality of the Duchess of Sussex, with a particular emphasis on Markle's gender and race. This study aims to contribute to the growing area of impolite phenomena in digitally-mediated communication but also to the body of literature on Violence Against Women (Bou & Garcés-Conejos, 2016) by throwing light on how interpersonal relations are enacted in social media.

References

Author, in press.

Baker, Paul, Costa Gabrielatos, Majid KhosraviNik, Anthony M. McEnery, and Ruth Wodak. 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. https://doi.org/10.1177/0957926508088962.

Bou-Franch, Patricia, and Pilar Garcés-Conejos Blitvich. 2016. "Gender Ideology and Social Identity Processes in Online Language Aggression against Women." In *Exploring Language Aggression against Women,* edited by Patricia Bou-Franch, 59–81.

Clancy, L & Yelin, H. 2021. Monarchy is a Feminist Issue: Andrew, Meghan and #MeToo Era Monarchy. *Women's Studies International Forum*, 84, 102435. https://doi.org/10.1016/j.wsif.2020.102435.

Culpeper, Jonathan, 1996. Towards and anatomy of impoliteness. *Journal of Pragmatics* 25 (3), 349–367.

Culpeper, Jonathan, 2005. Impoliteness and entertainment in the television quiz show: the weakest link. *Journal of Politeness Research* 1 (1), 35–72.

Culpeper, Jonathan, 2011. Impoliteness: Using Language to Cause Offence. Cambridge: CUP.

Culpeper, Jonathan, Bousfield, Derek, Wichmann, Anne, 2003. Impoliteness revisited: with special reference to dynamic and prosodic aspects. *Journal of Pragmatics* 35 (10–11), 1545–1579.

Duncan, E. and Low, V. 2018. Can Meghan Markle modernise the monarchy? *1843*. [Online] [Accessed on 14th May 2018] https://www.1843magazine.com/features/canmeghan-markle-modernise-the-monarchy.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years on. *Lexicography* 1: 7–36. https://doi.org/10.1007-/s40607-014-0009-9.

Reisigl, Martin, and Ruth Wodak. 2016. "The Discourse-Historical Approach." In *Methods of Critical Discourse Studies,* edited by Ruth Wodak and Michael Meyer, Third Edition, 23–61. London/CA/New Delhi: Sage.

Exploring epistemic stance in conservative newspaper opinion articles on immigration: A contrastive English and Spanish approach

Elena Domínguez-Romero & Marta Carretero – *Complutense University of Madrid*

This paper draws from previous research on stance (Marín-Arrese 2015, 2016, 2017a, 2017b, 2021; Carretero et al., 2017; Domínguez Romero and Martín de la Rosa 2023) to explore epistemic stance in conservative newspaper opinion articles on immigration-related issues in English and Spanish. The model of analysis is based on three main categories: evidentiality, which concerns the kind or source of evidence for or against the factual claim expressed in the proposition (Willett, 1988; Aikhenvald, 2004; Wiemer & Stathi, 2010; Boye, 2012, among others); epistemic modality, which pertains to the estimation of the chances for or against the truth of the proposition (Nuyts, 2001; Carretero & Zamorano-Mansilla, 2013); and factivity, which pertains to the factual status of the proposition (Kiparsky & Kiparsky, 1970). The three categories are divided into subtypes according to factors such as mode of access to the evidence, degrees of commitment to the communicated content, and subjectivity regarding (non-)explicit mention of the conceptualiser. The quantitative analysis comprises a wide range of expressions from the following syntactic categories: a) English modal auxiliary verbs and Spanish modal periphrases; b) modal adjectives; c) modal adverbs and adverbials; d) constructions with modal lexical verbs.

The focus will be on analysing the similarities and differences in the distribution and use of specific epistemic stance realisations on a 160,000-word corpus comprising general conservative newspaper opinion articles from *The Telegraph (40,000 words* on the English side) and *El Mundo (40,000 words* on the Spanish side) and a comparable corpus of specific conservative opinion newspaper articles on immigration and humanitarian crises involving refugees published in English and Spanish totalling 80,000 words approximately. The results reveal distributional differences between the epistemic expressions in the two corpora, thus shedding light on the motivations leading journalists' choices regarding these realisations to come across as reflective individuals holding personal beliefs/attitudes about or negotiating the meaning of immigration.

References

Aikhenvald, A. Y. 2004. *Evidentiality*. Oxford: OUP.

Boye, K. 2012. *Epistemic Meaning: A Crosslinguistic and Functional-cognitive Study*. (Empirical Approaches to Language Typology 43). Berlin: De Gruyter.

Carretero, M., J. I. Marín-Arrese and J. Lavid-López 2017. Adverbs as evidentials: An English-Spanish contrastive analysis of twelve adverbs in spoken and newspaper discourse. *Kalbotyra* 70, 32–59.

Carretero, M., and Zamorano-Mansilla, J. R. 2013. Annotating English adverbials for the categories of epistemic modality and evidentiality. In J. I. Marín-Arrese, M. Carretero, J. Arús Hita, and J. van der Auwera (eds.), *English Modality: Core, Periphery and Evidentiality*, pp. 317–355. Berlin: Mouton de Gruyter.

Domínguez Romero, E. and Martín de la Rosa, V. 2023. Epistemic Stance in Opinion Newspaper Articles and Political Speeches: An English/Spanish Contrastive Approach. In Marín Arrese, J. I., Hidalgo-Downing, L. and Zamorano, J. R. (eds.), *Stance, Inter/Subjectivity and Identity in Discourse*. Frankfurt: Peter Lang.

Kiparsky, P. and Kiparsky, C. 1970. Fact. In M. Bierwisch and K. E. Heidolph (eds). *Progress in Linguistics*, pp. 143–173. The Hague: Mouton.

Marín-Arrese, J. I. 2015. 'Epistemicity and stance: A cross-linguistic study of epistemic stance strategies in journalistic discourse in English and Spanish'. *Discourse Studies,* 17 (2), 210–225

Marín-Arrese, J. I. 2016. 'Epistemicidad y posicionamiento discursivo: Un estudio interlingüístico de la evidencialidad en el discurso periodístico en castellano y en inglés'. In R. González Ruiz, D. Izquierdo Alegría, and O.

Loureda Lamas (eds.), *La Evidencialidad en español: Teoría y descripción*, pp. 329–350. Iberoamericana, Madrid and Vervuert: Frankfurt am Main.

Marín-Arrese, J. I. 2017a. 'Stancetaking and Inter/Subjectivity in journalistic discourse: The Engagement system revisited'. In R. Breeze and I. Olza (eds.), *Evaluation in Media Discourse: European perspectives*, pp. 21–48. Bern: Peter Lang.

Marín-Arrese, J. I. 2017b. 'Multifunctionality of evidential expressions in discourse domains and genres: Evidence from cross-linguistic case studies'. In J. I. Marín-Arrese, G. Hassler and M. Carretero (eds.), *Evidentiality Revisited: Cognitive Grammar, Functional and Discourse-pragmatic perspectives*, pp. 195–223. Amsterdam/Philadelphia: Benjamins.

Marín-Arrese, J. I. 2021. 'Stance, emotion and persuasion: Terrorism and the Press'. *Journal of Pragmatics* 177, 135–148.

Nuyts, J. 2001. Epistemic Modality, Language and Conceptualization: A Cognitive-pragmatic Perspective. Amsterdam: John Benjamins.

Wiemer, B. & Stathi, K. 2010. The database of evidential markers in European languages. A bird's eye view of the conception of the database (the template and problems hidden beneath it). *STUF* 63(4), 275–289.

Willett, T. 1988. 'A Cross-Linguistic Survey of the Grammaticalization of Evidentiality'. *Studies in Language* 12(1), 51–97.

Estudio diacrónico comparado de esp. *certas* y fr. *certes*

Catline Dzelebdzic – *Lyon 2 University*

Palabras clave: *certes, certas, modalidad epistémica, diacronía.*

Certes y *certas* son dos adverbios de modalidad epistémica que derivan de la misma base latina CERTAS, que sustituye a CERTO, 'ciertamente' (Trésor de la Langue Française informatisé); son aún más próximos por el hecho de que la palabra entra en la lengua española pasando por el francés, o el occitano o el catalán (Corominas 1954: 795). A primera vista, la diacronía de ambos no puede ser más distinta: mientras que *certes* sigue usándose en francés, con la adición de algunos usos particulares como un empleo en contextos concesivos (Adam 1997), *certas* deja de emplearse después del siglo XV. Sin embargo, a pesar de esta diferencia evolutiva, existen similitudes en sus usos durante la Edad Media. Se comportan, así, de manera muy próxima en estos ejemplos:

(1) Et **çertas** dixo el uarones yo cuidaua que echados los pennos de Espanya que non fincaua nengun lugar ni nengun hombre que contra mi se rebellasse. (Juan Fernández de Heredia, *Gran crónica de España*, 1385, CDH XII-1975)

(2) **Certes**, si grant cop me donna / Qu'a painnes entendi son non (Renaut de Beaujeu, *Bel Inconnu*, 1207, BFM). ('Ciertamente, tan gran golpe me dio que apenas oí su nombre').

En (1) y (2), *certas* y *certas* se emplean ambos como marcador del discurso con un valor de refuerzo, que es su uso mayoritario en la Edad Media, y comparten características sintácticas: se sitúan los dos en la posición inicial de una frase, mayoritaria para ambos, con una preferencia por los contextos dialógicos en el caso de *certas*. A pesar de esta semejanza, solo se ha estudiado la diacronía de ambos adverbios de manera individual (Rodríguez Somolinos 1995 para *certes*, Estellés Arguedas 2009 para *certas*), de modo que cabe estudiar hasta qué punto se emplean de modo similar en la Edad Media y cuáles son los elementos que provocan el mantenimiento del uno y la desaparición del otro.

Para llevar a cabo el estudio, hemos trabajado, para *certas*, con el *Corpus del Diccionario histórico de la lengua española* (CDH), en su versión Nuclear y en su extensión diacrónica, y, para *certes*, con la *Base de français médiéval* (BFM). Se han manejado, en total, unas 542 ocurrencias en el CDH y unas 1.894 ocurrencias en la BFM. Además de estos datos propios, el análisis se ha apoyado en los trabajos diacrónicos existentes sobre ambos adverbios (Rodríguez Somolinos 1995 y Estellés Arguedas 2009), confirmando o completando sus observaciones.

El análisis de las ocurrencias de *certas* y *certes* en los dos corpus en la Edad Media permitirá describir y comparar los empleos y contextos de uso de cada adverbio, para evidenciar sus similitudes y diferencias. Esta comparación nos llevará a proponer hipótesis sobre la pervivencia del adverbio francés, que se enfrenta a menos competencia paradigmática que *certas*, que comparte cada vez más usos con los otros adverbios *ciertamente*, *cierto* y *por cierto*. Además, *certes* se diversifica fuera de su uso privilegiado en contextos dialógicos mientras que el adverbio español encuentra restricciones de empleo. De este modo, el estudio permitirá aportar más luces sobre la existencia compartida de esta forma en los dos idiomas, aunque sea temporánea, y sobre los factores divergentes que provocan su evolución distinta.

Bibliografía

Adam Jean-Michel 1997: «Du renforcement de l'assertion à la concession : variations d'emploi de *certes*», *L'information grammaticale*, vol. 73, nº 1, p. 3-9.

Coromines Joan 1954: Diccionario crítico etimológico de la lengua castellana, Berna: Francke.

Estellés Arguedas María 2009: Gramaticalización y gramaticalizaciones. El caso de los marcadores del discurso de digresión en español, Universitat de València.

Rodríguez Somolinos Amalia 1995: «*Certes, voire* : l'évolution sémantique de deux marqueurs assertifs de l'ancien français», *LINX*, vol. 32, nº 1, p. 51-76.

TLFi: Trésor de la langue Française informatisé, https://www.atilf.fr/tlfi, ATILF - CNRS & Université de Lorraine.

A corpus-assisted analysis of motifs in forced migration in children's picture books

Izaskun Elorza & Maria Bîrlea – *University of Salamanca*

Keywords: *corpus-assisted analysis, literary motifs, stages of forced migration, picture books, multimodal analysis.*

The social attention that migration has received in the last decades is being mirrored in children's literature, with forced migration narratives recognised as an emergent genre (Hope, 2008). This paper presents a corpus-assisted analysis (Partington, 2008) of the frequency and distribution of literary motifs of migration in a corpus of multimodal migrants' narratives. The purpose of the analysis was to find the resources employed by writers and illustrators of migration-themed picture books for representing migration as a process of forced displacement. Particularly, we were interested in finding recurrent patterns in the migrants' narratives for children. The motivation for this is that the kind of representations of migration that children's literature presents is of paramount importance, as it may condition the understanding that children will have of migration. A question we wanted to answer was which motifs were given more prominence than others in the narratives, and our assumption was that more prominence was provided by accumulation of resources. That is to say, if a motif was constructed only visually or only verbally, it was given less prominence than when it was constructed visually *and* verbally. In this sense, we were interested in finding out which motifs were constructed as *collustrations* (McGlashan, 2015) To this end, a corpus of thirty high-quality migration- themed picture books published in English in the last decades, and authored by prestigious writers and illustrators, was compiled.

As narrative picture books combine visual and verbal resources for creating narrative tension, irony and other effects on the reader (Nikolajeva & Scott, 2006), a theoretical framework has been adopted which integrates Halliday's (1978) social semiotics and Kress and Van Leeuwen's (2006, 2021) visual social grammar. The topics that have been reported as recurrent in narratives of forced migration by Arizpe (2021), Dooley et alia (2016), Evans (2015), Hope (2008) and Orgad et alia (2021) have been categorised into an annotating scheme, which has been used for scrutinising all the corpus. The annotation tags discriminate between visual and verbal resources. A pilot analysis has been carried out by two analysts on a sample of three picture books. The discrepancies in the annotation have been greater in the annotation of the visual resources than on the verbal representations of the topics, and the annotation scheme has been revised.

Together with the analysis of the relative frequencies of the motifs in the corpus, the distribution of each motif has been studied along the narrations in reference to a six-stage frame model of migration. In this way, we have not only been able to identify which motifs are present and how frequently they appear in the corpus, but also at which stage of the migrants' trajectories they appear typically. The motifs are construed in the narrations visually, verbally, or by means of resources from both modes at the same time, and correspond to two types of conceptualization of the migration process: those referring to the migrant's displacement, such as mobility, and those which are associated with the migrant's transformation along the process and connected to emotion, self, and identity.

Bibliography

Arizpe, Evelyn 2009. Sharing visual experiences of a new culture: Immigrant children's responses to picturebooks and other visual texts. In Evans, Janet (ed.) *Talking beyond the page: Reading and responding to picturebooks*. Routledge: London and New York, 134-151.

Baghban, Marcia 2007. Immigration in childhood: Using picture books to cope. *The Social Studies* 98: 2, 71-76. DOI: https://doi.org/10.3200/TSSS.98.2.71-76.

Baker, P. & McGlashan, M. 2020 'Critical Discourse Analysis'. In Adolphs, S. & Knight, D. (Eds.) *The Routledge Handbook of English Language and the Digital Humanities*. London: Routledge and New York, 220-241.

Dooley, Karen, Tait, Gordon Tait, & Zabarjadi Sar, Hora 2016. Refugee-themed picture books for ethical understanding in curriculum English. *Pedagogies: An International Journal* 11: 2, 95-108. DOI: 10.1080/1554480X.2016.1165619

Evans, Janet 2015. Could this happen to us? Children's critical responses to issues of migration in picturebooks. In Evans, Janet (ed.) *Challenging and controversial picturebooks: Creative and critical responses to visual texts*. Routledge: London and New York, 243-259.

Hope, J. 2008. "One Day We Had to Run": The Development of the Refugee Identity in Children's Literature and its Function in Education. *Children's Literature in Education*, 39, 295–304.

McGlashan, M. 2016. *The representation of same-sex parents in children's picturebooks: a corpus-assisted multimodal critical discourse analysis*. Lancaster University, UK: PhD Thesis.

Nikolajeva, Maria & Scott, Carole 2001. *How picturebooks work*. Routledge: London and New York.

Orgad, S., Lemish, D., Rahali, M., & Floegel, D. 2021. Representations of migration in U.K. and U.S. children's picture books in the Trump and Brexit era. *Journal of Children and Media*, 15 (4), 549-567.

Partington, A. 2008. The armchair and the machine: Corpus-Assisted Discourse Studies. In Taylor Torsello C, et al. (eds) *Corpora for University Language Teachers*. Bern: Peter Lang, 189–213.

El estudio diacrónico del español en contacto a través del *Corpus Mallorca*

Andrés Enrique-Arias & Ruth Miguel-Franco – *University of the Balearic Islands*

Palabras clave: *contacto de lenguas, catalán, Mallorca, lingüística histórica.*

El Corpus Mallorca (www.corpusmallorca.es) es un corpus informatizado en línea diseñado para el estudio histórico del castellano producido por catalanohablantes en la isla de Mallorca. El corpus se compone de más de 1 200 documentos (unas 715 000 palabras) fechados entre 1640 y 1909. La tipología documental es muy amplia, pero se han seleccionado preferentemente textos en los que aflora el vernáculo, como actas y declaraciones de testigos o epístolas personales. Además, el Corpus Mallorca, junto con otros corpus diacrónicos y sincrónicos del español de Mallorca, ha servido como base para el proyecto CAFECONMIEL, que aplica herramientas de inteligencia artificial al estudio del contacto de lenguas en la isla. El objetivo de esta comunicación es repasar las principales cuestiones filológicas y técnicas en la creación del Corpus Mallorca, así como presentar las funcionalidades y posibilidades del corpus en la investigación.

En primer lugar, se prestará atención a problemas como el equilibrio entre paleografía y normalización en la preparación de los textos, el diseño del motor de búsqueda y de la aplicación de descarga y análisis de resultados. El Corpus Mallorca sigue los estándares de edición documental de la Red Charta: ofrece una reproducción facsimilar, una transcripción paleográfica que refleja las grafías del original y una versión normalizada. De este modo, la transcripción paleográfica sirve para localizar fenómenos grafofonéticas y la versión normalizada, rasgos morfosintácticos y léxicos, sin que en ningún caso se pierda información. Sin embargo, la aplicación de algoritmos de agrupamiento plantea nuevos retos en la preparación de los textos, que debe combinar la exactitud filológica con la adaptación a las necesidades de la metodología de la inteligencia artificial.

En lo que respecta a las funcionalidades del corpus, el motor de búsqueda permite el empleo de expresiones regulares y la descarga de resultados en formato de hoja de cálculo, lo que amplía considerablemente las posibilidades de obtención de datos de interés sin necesidad de recurrir a la lematización. Como se ilustrará con varios ejemplos, la presentación múltiple de los textos junto a las posibilidades de la herramienta de búsqueda aporta algunas ventajas prácticas en búsquedas léxicas, morfosintácticas y fonéticas. Además, el empleo de algoritmos de clústeres ha permitido avances importantes en el estudio diacrónico del español en contacto con el catalán de Mallorca.

Taboo language and incest in the UK press (2017-2022): Finding absence in corpus linguistics

Sophie Eyssette – *University La Sapienza & University of Silesia*

Keywords: *taboo language, corpus linguistics, corpus design, media discourse, incest taboo.*

Taboos have been prevalent in all societies throughout history. Taboos prescribe behaviors related to death, food, and sexuality, and one of the primary sexual taboos is incest. It is necessary to distinguish between social taboo and linguistic taboo (Diffloth, 2014, p. 157). Incest is worth exploring from a sociolinguistic perspective as it is both a social and linguistic taboo.

Additionally, taboo language is often defined as swear words (Allan, 2019; Allan & Burridge, 2007; Bednarek, 2019; Jay et al., 2008; Madan et al., 2017; Pedraza, 2018; Reilly et al., 2020); however, this paper aims to identify taboo words related to incest that are not swear words, but rather words that are considered too taboo to be spoken (Casas Gómez, 2012; Khairullina et al., 2020).

Therefore, this research aims to analyse the unspeakable. To do this, this study addresses the methodological challenge of finding absence in a corpus. The issue of finding absence has been discussed in corpus linguistics (Duguid & Partington, 2018; Partington, 2014; Schröter & Taylor, 2018) and methods have been developed to compare corpora to see if certain elements are missing from one dataset to another (Alcántara-Plá & Ruiz-Sánchez, 2018; Strand, 2018).

However, in this study, absence is the main selection criterion for building a corpus on incest taboo. The corpus consists of eight British newspapers from October 13, 2017 to October 14, 2022, covering the period from the emergence of the #MeToo movement up to the first collection day. The goal is to find articles that discuss incest without using the word "incest." Therefore, the methodological challenge in carrying out this corpus linguistics study is to find a missing word through a search query that uses specific search terms.

To this end, I will explain why my search terms are not the lemma "incest*," but rather the broad terms "abuse" and "father." I will also introduce an iterative approach to narrow down a corpus of 23,015 articles, comprising 28,187,466 words, to a corpus of 258 articles, comprising 3,827,015 words. I will discuss limitations and solutions, and present results comparing a corpus containing the word "incest*" to a corpus deliberately omitting it.

Bibliography

Allan, K. Ed.. 2019. *The Oxford handbook of taboo words and language* First edition. Oxford University Press.

Allan, K. & Burridge, K. 2007. *Forbidden words: Taboo and the censoring of language* [Repr.]. Cambridge University Press.

Bednarek, M. 2019. 'Don't say crap. Don't use swear words.' – Negotiating the use of swear/taboo words in the narrative mass media. *Discourse, Context & Media*, *29*, 100293. https://doi.org/10.1016/j.dcm.2019.02.002.

Casas Gómez, M. 2012. The Expressive Creativity of Euphemism and Dysphemism. *Lexis*, *7*. https://doi.org/10.4000/lexis.349.

Diffloth, G. 2014. To Taboo Everything at All Times. *Annual Meeting of the Berkeley Linguistics Society*, *6*. https://doi.org/10.3765/bls.v6i0.2141.

Duguid, A. & Partington, A. 2018. Absence, you don't know what you're missing. Or do you? In C. Taylor & A. Marchi Eds., *Corpus Approaches to Discourse, A Critical Review*. Routledge.

Jay, T., Caldwell-Harris, C., & King, K. 2008. Recalling Taboo and Nontaboo Words. *The American Journal of Psychology*, *121*1, 83. https://doi.org/10.2307/20445445.

Khairullina, R. K., Fatkullina, F. G., So, Q., & Lin, Z. 2020. *Taboo As A Linguistic And Cultural Phenomenon*. 1969–1975. https://doi.org/10.15405/epsbs.2020.10.05.259.

Madan, C. R., Shafer, A. T., Chan, M., & Singhal, A. 2017. Shock and awe: Distinct effects of taboo words on lexical decision and free recall. *Quarterly Journal of Experimental Psychology*, *70*4, 793–810. https://doi.org/10.1080/17470218.2016.1167925.

Partington, A. 2014. Mind the gaps: The role of corpus linguistics in researching absences. *International Journal of Corpus Linguistics*, *19*1, 118–146. https://doi.org/10.1075/ijcl.19.1.05par.

Pedraza, A. P. Ed.. 2018. Linguistic taboo revisited: Novel insights from cognitive perspectives. De Gruyter Mouton.

Reilly, J., Kelly, A., Zuckerman, B. M., Twigg, P. P., Wells, M., Jobson, K. R., & Flurie, M. 2020. Building the perfect curse word: A psycholinguistic investigation of the form and meaning of taboo words. *Psychonomic Bulletin & Review*, *27*1, 139–148. https://doi.org/10.3758/s13423-019-01685-8.

Schröter, M. & Taylor, C. 2018. Exploring Silence and Absence in Discourse, Empirical Approaches. Palgrave.

Discursive value creation of sustainable fashion in Shanghai's high-end market: A mix-methods approach

Qin Fan – *Lancaster University*

Keywords: *distinction, sustainable fashion, Shanghai, discursive value creation, mix- methods approach.*

In this study, I examine how Bourdieu's conceptualisation of distinction manifests itself in the promotion of fashion products that are regarded as 'ethical', 'sustainable' and 'authentic'. According to Bourdieu (1984:231), the producers who are guided by thelogic of competition with other producers and by the specific interests associated with their position in the field of production, produce distinct products to meet different cultural interests that the consumers attribute to their class conditions and positions. This research investigates how the concept of 'sustainable fashion' is constructed and circulated linguistically in Shanghai's high-end market, and how added value is discursively created around their products for specific social groups. In particular, the focus is on Shanghai, one of the most affluent cities in China that is exemplary of changing consumption patterns among a growing middle- and upper-class who are geared towards consuming sustainable fashion products. Research has shown that the language used within the commodity chain process is not only limited to its descriptive function for the production, circulation, or exchange of products but can also be considered an important constitutive part of the entire process (e.g. Heller et al. 2014; Lorente, 2012; Shankar and Cavanaugh, 2012). Under this argument, the study aims to highlight the significance of language in creating taste distinction and contribute to scholarly discussions on the role of language within political economies.

To achieve methodological triangulation and provide a more comprehensive picture, a mix-methods approach that combines ethnography and corpus-assisted discourse analysis is adopted. The data under analysis consists of field notes, interview transcripts, texts collected from the field (e.g. promotional pamphlets, posters, exhibition boards) and social media texts of stakeholders within the fashion industry in Shanghai. The ethnographic part, comprised of participant observations and semi- structured interviews, aims to investigate the underlying relations of stakeholders and draw up a chain of commodities that links think tanks, recycling initiatives, garment traders, production cooperatives and fashion brands which are all engaged in the valuation of sustainable clothing. Furthermore, the ethnographic data is analysed by applying content analysis techniques (Klippendorff, 2019) to identify the discourses constructing the value of sustainable fashion.

The corpus-analytical part, informed by ethnography, examines stakeholders' Weibo and WeChat (two popular social media websites in China) texts to explore how social media contribute to the discursive creation of value and to the self-representation of stakeholders. Two specialised corpora are built: (i.) the brand corpus (collected from two local fashion brands) and (ii.) the consulting firm corpus (collected from two consulting firms that specialise in promoting sustainable fashion). For the brand corpus, both WeChat and Weibo are used to source texts. However, the text type of the two platforms is different. Considering their differences in language style, two sub-corpora are built for each brand separately, therefore, there are four sub-corpora under the brand corpus in total. Similarly, with regard to the consulting firm corpus, it also comprises four sub-corpora, whose textual data are collected from the WeChat and Weibo of two firms. This part of the data represents expert discourses around sustainability within the fashion industry, introducing the concept to a wider public. The main topics of each sub-corpora (8 in total) are identified through keywords analysis using the online corpus analysis interface, Sketch Engine (http://www.sketchengine.eu/), with the Chinese Web 2017 (zhTenTen17) Simplified chosen as the reference corpus. The keywords are then explored by investigating their contexts and mapping them onto the thematic categories based on the discourses identified previously. Additionally, the concordance analysis of keywords was conducted to look for evidence of grammatical, semantic or discourse patterns in their contextual uses, which contributes to a better understanding of discursive strategies employed to create added value. It is argued that the

added value of high-end sustainable fashion products is discursively constructed through taste distinction, which helps the stakeholders establish a niche market in Shanghai by differentiating themselves from other businesses within the fashion industry, especially fast fashion for example, which rely on the industrial-, exploitative- and delocalised forms of production.

References

Bourdieu, P., 1984. Distinction. A Social Critique of the Judgment of Taste. London: Routledge and Kegan Paul.

Heller, M., Pujolar, J. and Duchêne, A., 2014. Linguistic Commodification in Tourism. *Journal of Sociolinguistics,* 18 (4), 539–566.

Klippendorff, K. 2019. *Content Analysis: An Introduction to Its Methodology*, fourth ed. SAGE, Los Angeles.

Lorente, B. P., 2012. The Making of "Workers of the World": Language and the Labour Brokerage state. In: Duchêne, A., Heller, M. (Eds.), *Language in Late Capitalism: Pride and Profit.* New York: Routledge, 183–206.

Shankar, S. and Cavanaugh, J. R., 2012. Language Materiality in Global Capitalism. *Annual Review of Anthropology,* 41, 355–369.

Description of MedCorpus, an aligned parallel corpus of medical fictional language

Goretti Faya-Ornia[1], Carmen Quijada-Díez[2], Natalia Barranco-Izquierdo[1] & Teresa Calderón-Quindós[1]

University of Valladolid[1] – University of Oviedo[2]

Keywords: *aligned parallel corpora; foreign language learning; foreign language teaching; medical language; contrastive studies.*

MedCorpus is a corpus created by the research group "Communicative and intercultural skills in foreign language", from the University of Valladolid. It is an aligned parallel corpus (English-Spanish) of medical fiction TV programmes (Dr House, Grey's Anatomy and ER), where the source texts are the English scripts of the originally-broadcasted TV shows and the target texts are the dubbed versions broadcasted in Spanish. As of now, MedCorpus, which is expected to be open access from 2024 on, already comprises 413,000 words in English and 387,000 in Spanish and is expected to reach 12,000,000 words.

Parallel corpora have proven a useful didactic tool for foreign language learning and teaching (Aijmer 2009, Doval 2018, Sinclair 2004) and are of great interest in different fields of Translation Studies (Bernardini and Russo 2017, Doval 2016, Doval and Sánchez-Nieto 2019, Molés-Cases 2016), either to detect the translation strategies used by translators or to develop trainee translators' competences (Bernardini 2016, Gallego Hernández 2016, Liu 2013). However parallel corpora that integrate specialized languages are still scarce and so are parallel corpora that cover orality in specialized contexts. It could be argued that TV fictional shows should not be considered reliable sources regarding specialized contexts, but the artificial and made-up orality of medical language in professional environments can be proven to be effective in foreign language learning. In fact, this type of discourse integrates, if not reliable information regarding how to overcome medical conditions, at least specific medical terminology as well as common collocations, discourse markers and text conventions that are typical of the orality in healthcare settings. Moreover, it should be borne in mind that TV shows relating to particular professional fields often use specialized discourse precisely to create an impression of reality.

During our presentation we will be presenting MedCorpus and will comment on its compilation and possible applications and uses. Firstly, we will describe how source (English) and translated texts (Spanish) were aligned and then uploaded into Sketch Engine, a step in which undergraduate students participated within the frame of a teaching innovation project. We will then explore the range of possibilities that such a corpus might offer, such as performing multilingual searches, comparing the behaviour of the same string in two different languages or even observing what particular cognitive and translations strategies were used during the audio-visual translation process (dubbing).

Being this a corpus that deals with medical fictional oral language, it may also be interesting for researchers willing to study certain medical features (or at least medical features as expressed in fictional TV shows), to perform contrastive studies or to observe the features of orality in the two languages involved. Furthermore, it might be of interest for hospitals and translation companies, since an extensive aligned translation memory can be created and used as a general bilingual reference.

References

Aijmer, Karin (ed.). 2009. *Corpora and Language Teaching.* Amsterdam: John Benjamins.

Bernardini, Silvia. 2016. "Intermodal Corpora: A Novel Resource for Descriptive and Applied Translation Studies." In *Corpus-based Approaches to Translation and Interpreting*, ed. by Gloria Corpas Pastor and Míriam Seghiri, 173-194. Oxford: Peter Lang.

Bernardini, Silvia, and Mariachiara Russo. 2017. "Corpus Linguistics, Translation and Interpreting." In *The Routledge handbook of translation studies and linguistics,* ed. by Kirsten Malmkjær, 342-356. Oxford: Taylor & Francis/Routledge.

Doval Reixa, Irene. 2016. "Bilingual Parallel Corpora for Linguistic Research." In *EPiC Series in Language and Linguistics* Volume 1, CILC2016. 8th International Conference on Corpus Linguistics.

Doval Reixa, Irene. 2018. "Parallel Corpora in Foreign Language Learning and Teaching: an example of use based on the Corpus PaGeS." *Clina: An Interdisciplinary Journal of Translation*, 4 (2): 65-82.

Doval Reixa, Irene, and María Teresa Sánchez Nieto (eds.) 2019. *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications.* Amsterdam: John Benjamins.

Gallego Hernández, Daniel. 2016. "Developing Trainee Translators' Instrumental Subcompetence." In *Corpus-based Approaches to Translation and Interpreting*, ed. by Gloria Corpas Pastor, and Miriam Seghiri, 173-194. Oxford: Peter Lang.

Liu, Kanglong. 2013. "Investigating Corpus-assisted Translation Teaching: A Pilot Study." In *Conducting Research in Translation Technologies*, ed. by Pilar Sánchez-Gijón, Olga Torres-Hostench, and Bartolomé Mesa-Lao, 13: 141-162. Bern: Peter Lang.

Molés-Cases, Teresa. 2016. "Compilation and Analysis of a Parallel Corpus for Research in Translation. Project with Déjà Vu, Treetagger and IMS Open Corpus Workbench." [Compilación y análisis de un corpus paralelo para la investigación en traducción. Proyecto con Déjà VU, treetagger e IMS open corpus workbench] *RLA* 54 (1): 149-174. doi:10.4067/S0718-48832016000100008.

Sinclair, John. McH. 2004. *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins.

Towards an annotation schema of financial discourse based on functional discourse units

Javier Fernández-Cruz & Irina Muñoz-Toala – *University of Málaga*

Keywords: *sentiment analysis, textual analysis, functional discourse units, economic discourse.*

The aim of this presentation is to explore the particularities of the textual structure of economic opinion news texts. Sentiment analysis (Liu, 2015) has been one of the main tasks of natural language processing for the last two decades and its task is to automatically detect polarity (whether positive or negative) in texts of all kinds. Its main models are usually based on machine learning or large sentiment lexicons.

Lingmotif (Moreno Ortiz, 2017), is a multilingual, cross-platform, lexicon-based sentiment analysis (SA) tool that can be used with both general language and domain-specific texts. Over the last few years, it has achieved outstanding results (Moreno-Ortiz & Perez-Hernandez, 2018). However, while the lexicon and algorithms can still be refined for slight improvements, the tool has reached the point where only marginal improvements in accuracy and completeness can be obtained.

Current efforts will focus on the study of textual structure and its implications for sentiment. There have been several models to approach discourse structure. The Rhetorical Structure Theory (RST) by Mann & Thompson (1987), later implemented as an automatic parser by Marcu (1997, 1999) aims to identify the structure of a text segmented into Elementary Discourse Units which are related to each other and reflect the overall organisation and coherence of a text by postulating a hierarchical structure based on the idea that each part of a text has a function. We have drawn on the Functional Discourse Units by Egbert et al., (2021) in order to characterise the underlying structure of texts, to reflect the main discourse functions and to identify the core of the document where the main idea is expressed. This method is introduced for the segmentation of conversational texts into discourse units based on their communicative purposes. In our case, the aim is simply to identify the determining discourse segments issued by the opinion holder for the correct assignment of the polarity of the text.

To this end, we used Prodigy (Montani & Honnibal, 2018) to apply our scheme for text segments, which consists of five layers: (1) polarity, (2) discursive functions, (3) aspect, (4) entity and (5) opinion holder. In order to do this, we have identified a number of discourse units and aspects that are specific to the domain of opinion news on the economy.

This allows us to (a) generate a comprehensive dataset to serve as a basis for improving sentiment analysis at the textual level, and (b) serve as a rich source of information on the textual metafunction of opinion texts in the press. In order to understand the particularities of the domain of economics in the press, several annotators have annotated the different sections of each of the texts of a corpus of about half a hundred columns written during the autumn of 2022 by the chief economics columnists of three major English-language newspapers: *The Guardian*, *The New York Times* and *The Financial Times*.

Preliminary results of this model offers numerous opportunities for improving sentiment analysis. On the other hand, it provides a great panoramic view on textual linguistics and offers both quantitative and qualitative insights, for example, discourse markers (Dafouz-Milne, 2008) and news values (Bednarek & Caple, 2014; Caple & Bednarek, 2016).

References

Bednarek, M. & Caple, H. 2014. Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. *Discourse & Society,* 25(2), 135–158. https://doi.org/10.1177/0957926513516041

Caple, H. & Bednarek, M. 2016. Rethinking news values: What a discursive approach can tell us about the construction of news discourse and news photography. *Journalism*, 17(4), 435–455. https://doi.org/10.1177/1464884914568078

Dafouz-Milne, E. 2008. The pragmatic role of textual and interpersonal metadiscourse markers in the construction and attainment of persuasion: A cross-linguistic study of newspaper discourse. *Journal of Pragmatics,* 40(1), 95–113. https://doi.org/10.1016/j.pragma.2007.10.003

Egbert, J., Wizner, S., Keller, D., Biber, D., McEnery, T., & Baker, P. 2021. Identifying and describing functional discourse units in the BNC Spoken 2014. *Text & Talk,* 41(5–6), 715–737. https://doi.org/10.1515/text-2020-0053

Liu, B. 2015. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press.

Mann, W. C. & Thompson, S. A. 1987. *Rhetorical Structure Theory: A Theory of Text Organization* (Report ISI/RS-87-190). Information Sciences Institute.

Marcu, D. 1997. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts [Thesis]. http://citeseer.nj.nec.com/24902.html

Marcu, D. 1999. Discourse trees are good indicators of importance in texts. In I. Mani & M. Maybury (Eds.), *Advances in Automatic Text Summarization* (pp. 123–136). The MIT Press. http://www.isi.edu/~marcu-/papers/summar-book99.pdf

Montani, I. & Honnibal, M. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. Artificial Intelligence, to appear.

Moreno Ortiz, A. 2017. Lingmotif: A user-focused sentiment analysis tool. *Procesamiento Del Lenguaje Natural,* 58, 133–140.

Moreno-Ortiz, A. & Perez-Hernandez, C. 2018. Lingmotif-lex: A wide-coverage, state-of-the-art lexicon for sentiment analysis. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), 2653–2659.

Are they thematic? A systemic functional analysis of the textual role of fragments in English

Yolanda Fernández-Pena & Ana Elina Martínez-Insua – *Universidade de Vigo*

Keywords: *fragments, theme, contentfulness.*

Framed within a larger project on non-canonical syntax in written contemporary English, this study investigates the thematic nature of fragments in English. 'Fragments' are taken here as "semantically, discursively and pragmatically stand-alone constituents which are equivalent in propositional meaning, force and communicative function to a full clause" (Fernández-Pena, 2021). In terms of form, fragments are formally reduced, and syntactically and prosodically independent, as illustrated in (1)-(3):

(1) *Stupid, stupid man* <ICE-GB:W2F-008 #112:1>

(2) *Quite frightening but exciting too* <ICE-GB:W1B-014 #131:6>

(3) *In a much nicer area, away from the average back-packing German tourist!!!*

<ICE-GB:W1B-005 #126:5>

This study aims at analysing the textual role of fragmentary utterances like (1)-(3) from a systemic functional perspective (Halliday & Matthiessen, 2014). The general assumption in the framework is that only complete clauses have thematic structure, while clauses that do not have a predicator are not analysed for Theme/Rheme (Thompson, 2014). According to Thompson (2014, p. 153), either the Theme or Rheme may be missing from minor clauses (that is, those lacking a predicator) and, therefore, the remaining fragment may be either Theme (4) or Rheme (5).

(4)

Who	(would you most like to meet)
Theme	(Rheme)

(5)

(Are you)	Not sure what a special delivery is?
(Theme)	Rheme

Adapted from Thompson (2014, p. 153)

Quite on the contrary, this study will consider phrasal fragments as thematic. Accepting that Theme locates and orients the clause within its context because it has the function to "guide the addressee in developing an interpretation of the message" (Halliday & Matthiessen, 2014, p. 89), this study seeks to evince that sequences such as (1)-(3) above should be regarded as thematically prominent. It is claimed here that this type of fragment is the part chosen by the speaker as the starting point for the addressee. By using it, by making it prominent, the speaker wants to enable the addressee to process the message, just like they do when they thematise part of the message and make it prominent as Theme.

We analysed fragments from the parsed version of the British component of the *International Corpus of English* (ICE-GB) (Nelson, Wallis & Aarts, 2002), which we retrieved by means of the syntactic node PU,NONCL 'non-clausal parsing unit' (Bowie & Aarts, 2016). Once fragments have been considered as Themes, the analysis of their 'contentfulness' (Martínez-Insua, 2018; 2019) has proved helpful in characterising not only them but also the text types under analysis. The notion of contentfulness used in this study refers to content weight and is based on Berry's (2013) distinction between contentful and contentlight Subject Themes, and Prince's

(1981) assumed familiarity scale. This twofold assumption, that phrasal fragments are thematic and that their contentfulness may be measured against the same scale as that of subject Themes, allows us to investigate to what extent the two (thematic) units share features.

According to Berry's (2013, p. 259) hypothesis, most subject Themes are contentful in formal written texts, while most of them are contentlight in informal spoken texts. Our findings seem to be in partial agreement with Berry's. On the one hand, fragment Themes with heavier degrees of contentfulness are more common in written correspondence than in oral conversation, where lighter fragment Themes are more frequent. On the other hand, and quite interestingly, fragment Themes with the lightest types of contentfulness are equally scarce in both written correspondence and oral conversation. Our findings seem to suggest, then, that subjects and phrasal fragments have the same textual (thematic) role but differ in the degrees of contentfulness they bring to the thematic area.

References

Berry, Margaret. 2013. Contentful and contentlight subject themes in informal English and formal written English. In O'Grady, Gerard, Bartlett, Tom, and Fontaine, Lise (eds.). *Choice in Language, Applications in Text Analysis*. Equinox, pp. 243-268.

Bowie, Jill and Aarts, Bas. 2016. Clause fragments in English dialogue. In López-Couso, María José, Méndez-Naya, María José, Núñez-Pertejo, María José and Palacios-Martínez, Ignacio M (eds.), *Corpus Linguistics on the Move: Exploring and Understanding English through Corpora*. Leiden & Boston: Brill, pp. 259-288.

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

Halliday, M.A.K. and Matthiessen, Christian M.I.M. 2014. *Halliday's Introduction to Functional Grammar*. Routledge.

Martínez-Insua, Ana Elina. 2018. On the relevance of the textual metafunction for Spanish learners/teachers of English. In Sellami-Baklouti, Akila and Fontaine, Lise (eds.), *Perspectives from systemic Functional Linguistics*. New York and London: Routledge, pp. 269-287.

Martínez-Insua, Ana Elina. 2019. Scientific writing and the contentfulness of Subject Themes. How science was explained to (lay) audiences. *Journal of Pragmatics* 139: 216-230.

Nelson, Gerald, Wallis, Sean and Aarts, Bas. 2002. Exploring Natural Language: Working with the British Component of the International Corpus of English. John Benjamins.

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In Cole, Peter (ed.). *Radical Pragmatics*. Academic Press, pp. 223-256.

Thompson, Geoff. 2014. *Introducing Functional Grammar*. Routledge.

Why deal with *why*- and Mad-Magazine fragments? Modelling allostructional variation in contemporary English

Yolanda Fernández-Pena & Javier Pérez-Guerra – *Universidade de Vigo*

Keywords: *fragment*, *Mad Magazine sentences*, *British National Corpus*, *regression analysis*, *Construction Grammar*.

This paper focuses on two types of fragmentary constructions in English: *why*-fragments, in (1) and (2) below, and so-called Mad Magazine sentences (Akmajian 1984; Lambrecht 1990), in (3) and (4). 'Fragment' is here understood as any functionally stand-alone and syntactically and prosodically independent constituent which is formally reduced but nonetheless semantically, discursively and pragmatically equivalent to the corresponding complete clause in propositional meaning, force and communicative function (Fernández-Pena 2021).

(1) Why deal with *why*-fragments?

(2) Why *why*-fragments?

(3) Me paint the house purple?

(4) [A: I heard you went out clubbing last Saturday.] B: Me go out clubbing?

Even though these fragmentary utterances can be equivalent in meaning to their fully-fledged sentential counterparts, it is also true that the former may convey an additional nuance different from the 'orthodox' interrogative interpretation of the complete clauses. In fact, the reduced questions evoke a modal nuance (see Johnson 1975: 487; Weir 2017: 406; Zaitsu 2018, 2020) in (1) ('Why should one deal with *why*-fragments?') and (3) ('Why should I paint the house purple?'), a uniqueness interpretation (as in Weir 2014) in (2) ('Why *why*-fragments and not another type of fragment?') or serve to isolate the action from a specific time and question the possibility of such an action in general (Donaldson 2013: 13) in (4). That these 'enriched' meanings are not necessarily conveyed in the corresponding complete sentences implies that speakers and hearers conventionalise the fragments as unique constructions, that is, as special pairings of form and meaning (Goldberg 2006: 5).

This paper explores the alternation between the canonical (i.e. equivalent to that of the corresponding complete sentences) and enriched (i.e. with an added meaning or nuance) interpretations of *why*- and Mad-Magazine constructions in contemporary English based on evidence from spoken data retrieved from the BNC1994 DS (BNC Consortium 2007) and Spoken BNC2014 (Love et al. 2017) corpora. The statistical modelling of the data is based on a set of predictors compiled from the relevant literature, which includes as independent variables corpus/period, category of the fragment constituents, and the attestation of mismatches between the fragment constituents and their antecedents in the preceding text. The data are analysed using a multivariable statistical technique (binomial regression) that determines which predictors are significant in explaining the semantics of the fragments. To exemplify some of the results, our study shows that the timid (non-significant) increase in *why*-fragments revealed by the corpus data is supported by both higher probability of the so-called enriched interpretation of *why*-fragments in the Spoken BNC2014 data than in examples from BNC1994 DS, and the preference for novel expressions in the fragmentary construction (i.e. expressions not attested as such in the antecedent clauses) when the fragments have enriched meanings.

Based on Cognitive Construction Grammar theory (Goldberg 2019), this analysis suggests that *why*-fragments and complete *why*-sentences, on the one hand, and Mad Magazine sentences and conventional questions, on the other, may be interpreted as competing allostructional pairs (Cappelle 2006; Perek 2021). The choice of the fragmentary construction over the fully-fledged other depends on the enriched meaning intended by the speaker in a given situational context. Also, the findings confirm that linguistic variation and, in particular, processes of constructional change may be shaped and demonstrated by the application of statistical methods.

References

Akmajian, Adrian. 1984. Sentence types and the form-function fit. *Natural Language & Linguistic Theory* 2: 1–23.

BNC Consortium. 2007. *British National Corpus: XML edition*. Oxford: Oxford Text Archive.

Cappelle, Bert. 2006. Particle placement and the case for 'allostructions'. *Constructions* SV1-7: 1–28. urn:nbn:de:-0009-4-6839. <http://www.constructions-online.de>

Donaldson, James. 2013. On elliptical *why*-questions. MSc dissertation. Edinburgh: The University of Edinburgh.

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

Goldberg, Adele E. 2006. Constructions at work: The nature of generalization in language. Oxford: Oxford University Press.

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

Johnson, David E. 1975. Why delete Tense? *Linguistic Inquiry* 6/3: 481–489.

Lambrecht, Knud. 1990. 'What, me worry?' – 'Mad Magazine sentences' revisited. In *Proceedings of the Sixteenth Annual Meeting of the Berkeley Linguistics Society 16*, 215–228.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina and Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22/3: 319–344.

Perek, Florent. 2021. Alternation-based generalizations are stored in the mental grammar: Evidence from a sorting task experiment. *Cognitive Linguistics* 23/3: 601–635.

Weir, Andrew. 2014. *Why*-stripping targets Voice Phrase. *Proceedings of NELS 43*, 235–248.

Weir, Andrew. 2017. But write what? In Nicholas LaCara, Keir Moulton and Anne-Michelle Tessier eds. *A Schrift to Fest Kyle Johnson*. Amherst, MA: University of Massachusetts, 401–408.

Zaitsu, Anissa. 2018. Why make sense of silence? The clausal syntax of a reduced why-question. PhD, UC Santa Cruz.

Zaitsu, Anissa. 2020. Modality force and syntax in an understudied class of reduced *why*-questions in English. *Glossa: A Journal of General Linguistics* 5/1: 1–37.

MexLeC: A spoken and longitudinal corpus of Mexican beginner to advanced learners of English

Ana Abigahil Flores-Hernández & Pauline Moore – *Universidad Autónoma del Estado de México*

Keywords: *learner corpus, English L2, longitudinal, design.*

The present study reports the process of designing and collection a spoken and longitudinal corpus of Mexican university learners as well as a brief description of the data obtained throughout three years of work as part of a Postdoctoral research project. The aim of this project is to collect a national database to be used in the development of learner-centred tools and materials for ELT and as an empirical database for SLA research in Mexico (Meunier, 2021; Guilquin, 2015).

An interview eliciting monological and non-interactive tasks was designed to collect learner interlanguage in extended turns. This semi-guided interview lasting 10-16 minutes will be applied once a year, following each cohort of learners' acquisition process through 4-5 years. The rationale for these tasks are descriptors from CEFR (Council of Europe, 2018) matched with an analysis of tasks used in examinations and 140 currently available English learner corpora, and the text types in (Biber, 2004) as well as internal and external considerations of linguistic representativeness.

Interviews were videorecorded using the Zoom app and the transcription guidelines have been adapted from those used by the Trinity Lancaster Corpus and The International Corpus of Learner English. By May 2023 the corpus will include cohorts from three different state universities holding proficiency levels from A1 to B2 (CEFR). The first university (UAEMéx), with three years tracking time (3 samples), the second (UAEH), with two years tracking time (2 samples) and the third (UAQ), with a first sample collected. This provides coverage of ELT students in a variety of programs with distinct characteristics allowing for research into a wide range of variables.

The current size of the corpus is approximately 200,000 tokens and some of the most interesting preliminary findings are the (expected) low type/token ratio scores; the wide use of fillers and pauses followed by elaborated chunks; and the dissimilar features produced in the narrative task from those expected to be distinctive of this text-type.

References

Biber, D. 2004. Conversation text types: A multi-dimensional analysis. 7es Journées internationales d'Analyse statistique des Données Textuelles.

Council of Europe 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment; Companion Volume with New Descriptors. Strasbourg, Language Policy Division: Cambridge University.

Gablasova, D., Brezina, V., y McEnery, T. 2019. The Trinity Lancaster Corpus: development, description and application. International Journal of Learner Corpus Research, 5(2), 126-158.

Guilquin, G. 2015. From design to collection of learner corpora. In Granger, Sylviane, Gilquin.

Meuner, F. 2021. Introduction to learner corpus research. In Tracy-Ventura, N. and Paquot, M., eds. The Routledge Handbook of Language Acquisition and Corpora. London: Routledge.

TeCoPhy: A Text Corpus of German Physics Texts

Vitor Lécio Lacerda Fontanella[1,2], Tom Bleckmann[2], Lukas Dieckhoff[2],
Gunnar Friege[2] and Christian Wartena[1]

Hannover University of Applied Science and Arts[1] – Leibniz University Hannover[2]

Keywords: *corpus construction, German, physics, textbooks.*

To learn a subject, the acquisition of the associated technical language is important (Diethelm & Goschler, 2014; Pineker-Fischer, 2017; Poupova, 2018). Despite this widely accepted importance of learning the technical language, hardly any studies are published that describe the characteristics of most technical languages that students are supposed to learn. This might largely be due to the absence of specialized text corpora to study such languages at lexical, syntactical and textual level. In the present paper we describe a corpus of German physics text that can be used to study the language used in physics. The composition of such a corpus faces three major challenges:

1. We have to deal with OCR and the complicated layout of textbooks;
2. Physics texts contain a large number of symbols and formula.
3. Due to copyright restrictions, a corpus of texts from textbooks cannot be published.

Our primary goal was to have a large collection of German texts on physics covering various topics and different levels of proficiency, including at least some texts intended for secondary school students. Thus we included Wikipedia articles from the category physics (excluding articles about institutions and biographies of physicists), articles on school physics from the website https://www.leifiphysik.de/, as well as many (printed) textbooks (at secondary school and university level) and a few scientific books. Initially 264 books were scanned. Books with severe OCR problems were just removed from the collection. 221 books could be used for further processing.

We extracted the text from the scanned books using PDFMiner (https://github.com/pdfminer/pdfminer.six). To avoid problems with footers, page numbers, captions, etc., we determined the main fonts used in each book and extracted only text blocks using these fonts. After sentence splitting, only sentences having at least 50% alphabetical characters are kept. Finally, we removed English sentences, appearing e.g. in quotes. Thus, the corpus is a collection of sentences rather than a collection of coherent texts. On average 47% of the text could be extracted, resulting in 2.36 /$\cdot 10^5$ sentences or 5.3 /$\cdot 10^6$ tokens.

According to the German copyright laws we are allowed to distribute (still with restrictions) at most 15% of the text of each book. Thus we have to make a subselection of the texts. To guarantee the presence of enough terminology, we extracted a list of nouns occurring more than 5 times in the corpus and having a higher relative frequency than the word has in the German Reference Corpus DeReKo (Kupietz & Lüngen, 2014). We added moreover around 600 words that occur in typical collocations. This results in a list of 30.681 words. Half of the small corpus was constructed by selecting between 5 and 10 example sentences for each noun. The other half was selected by random sampling from the remaining sentences. It was guaranteed that at most 14% of each book was included.

The selection consists of 2.36 / 10^5 sentences and 5.32 / 10^6 words. The composition of this selection from different types of sources is given in Figure 1.

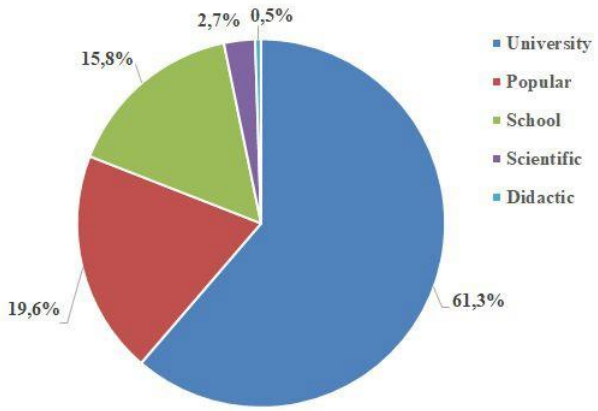

Figure 1: Composition of the corpus.

In order to see how representative the selection is for the large corpus, we extracted from the large corpus a list of words that occur at least 10 times. This list contains 45, 332 word forms, 39, 317 of which are also found in the small corpus. The Pearson correlation of the frequency values in both corpora is 0,999. If we extract from both corpora the 1,000 word forms with the highest relative frequency (compared with the DeReKo data), 904 words are included in both lists.

The large corpus cannot be distributed. The smaller variant of TeCoPhy can only be distributed for research on text and data mining and is available on request from the authors.

References

Diethelm, I. & Goschler, J. 2014. On human language and terminology used for teaching and learning cs/informatics. In *Proceedings of the 9th workshop in primary and secondary computing education* (p. 122-123). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2670757.2670765 doi: 10.1145/2670757.2670765

Kupietz, M. & Lüngen, H. 2014. Recent developments in DeReKo. In (p. 2378- 2385). Reykjavik: European Language Resources Association (ELRA). Retrieved from https://nbn-resolving.org/urn:nbn:de:bsz:mh39-31353

Pineker-Fischer, A. 2017. Von der Alltags- zur Bildungs- und Fachsprache. In Sprach- und fachlernen im naturwissenschaftlichen unterricht: Umgang von lehrpersonen in soziokulturell heterogenen klassen mit bildungssprache (pp. 41–82). Wiesbaden: Springer Fachmedien Wiesbaden. Retrieved from https://doi.org/10.1007-/978-3-658-16353-2_5 doi: 10.1007/978-3-658-16353-2_5

Poupova, J. 2018. Biological terminology: an opportunity for teaching in tandem. In *International conference new perspectives in science education* (pp. 382–385).

The grammatical complexity of film dialogue as input for L2 learning: A corpus-based study

Maicol Formentelli, Liviana Galiano & Maria Pavesi – *Università degli Studi di Pavia*

Keywords: *grammatical complexity, film dialogue, L2 input, L2 learning, English.*

Audiovisual dialogue in traditional and new media is a preferred means of accessing English extramurally and informally (Sockett 2014; Arnbjörnsdóttir and Ingvarsdóttir 2018; Pavesi and Ghia 2020). Although it may represent an optimal input for the L2 acquisition (Alvarez-Pereyre 2011; Forchini 2012; Pavesi 2015), it poses the paradox of making the language widely available but hardly accessible to language learners (Vanderplank 2010: 9). Complexity in film has been associated with its dual functionality at the realistic and diegetic level (Zago 2019), narrative and multimodal architecture (Perego et al. 2018), sociolinguistic variation and fast rate of speech delivery (Vanderplank 2020: 187-188). Linguistically, it has been extensively investigated at the lexical level against corpora of spoken English to identify the most common word families and define the vocabulary coverage learners need for the comprehension of texts (Webb and Rodgers 2009a, b; Jones 2017; Scheffler et al. 2020). Little or no research, however, has been conducted on the grammatical complexity of English audiovisual dialogue, a major component of L2 input and one that can be used in predicting acquisitional paths of development, to be distinguished from cognitive complexity or difficulty (see DeKeyser 2005; Pallotti 2015; Housen and Simoeans 2016).

The present corpus-based study aims to describe the complexity of film dialogue from a structural and register-functional approach perspective (Biber et al. 2022), hence combining structural elaboration with an understanding of complexity as a register-specific dimension of texts. A corpus of film dialogue will be compared to corpora of spontaneous conversation, the register it simulates on-screen, with a view to identifying main features of the input contemporary learner-users of L2 English are frequently and often predominantly exposed to in out-of-the classroom settings.

The following research questions will be addressed:

1) Which syntactic features contribute to the expression of grammatical complexity in film dialogue?
2) Through which syntactic features and to what extent does film dialogue approximate spontaneous conversation in grammatical complexity?

The quantitative analysis relies on 34 POS-tagged orthographic transcriptions of films from the Pavia Corpus of Film Dialogue (https://studiumanistici.unipv.it/?pagina=p&titolo=pcfd) divided into British and American film dialogue (ca. 176,000 and 190,000 tokens respectively) and combines POS-tag searches for word classes and selected syntactic patterns. The Spoken BNC2014 and the Longman Spoken American Corpus have been chosen as reference corpora to guarantee comparability in the tagging system (Claws 6 and 7) and across varieties of English. The following features of complexity were selected (Bulté and Housen 2012; Biber et al. 2021; Biber et al. 2022): index of subordination - finite and non-finite dependent clauses, clausal vs. phrasal coordination, length and type of NP premodification and postmodification through prepositional phrases. POS-tag patterns (e.g., VV* that, for verb controlled *that*-clauses; AT* JJ* NN*, for premodified noun phrases) were retrieved semi-automatically and the data were manually checked to exclude false positives.

Initial results show that film dialogue approximates spontaneous conversation in terms of complexity of both subordination and NP premodification, which corroborates the hypothesis that film dialogue represents a rich and complex input for L2 learners of English. The findings on the grammatical complexity of film dialogue will be discussed in light of their implications for naturalistic second language acquisition in out-of-the-classroom settings.

References

Alvarez-Pereyre, M. 2011. Using film as linguistic specimen: Theoretical and practical issues. In Piazza, R., Bednarek, M., Rossi, F. (eds), *Telecinematic Discourse: Approaches to the Language of Films and Television Series*. Amsterdam: John Benjamins, pp. 47-67.

Arnbjörnsdóttir, B. and Ingvarsdóttir, H. (eds) 2018. Language Development across the Life Span. The Impact of English on Education and Work in Iceland. Cham: Springer International Publishing.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. 2021. *Grammar of Spoken and Written English*. Amsterdam/Philadelphia: John Benjamins.

Biber, D., Gray, B., Staples, S., Egbert, J. 2022. The Register-functional Approach to Grammatical Complexity: Theoretical Foundation, Descriptive Research Findings, Application. London: Routledge.

Bulté, B., Housen, A. 2012. Defining and operationalising L2 complexity. In Housen A., Kuiken, F., Vedder, I. (eds), *Dimensions of L2 Performance and Proficiency – Investigating Complexity, Accuracy and Fluency in SLA*. Amsterdam/Philadelphia: Benjamins. pp. 21-46.

DeKeyser, R. M. 2005. What makes learning second-language grammar difficult? A review of issues. *Language Learning*, 55(S1): 1–25.

Forchini, P. 2012. Movie Language Revisited. Evidence from Multi-Dimensional Analysis and Corpora. Bern: Peter Lang.

Housen, A., Simoens, H. 2016. Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition*, 38(2): 163–175.

Jones, C. 2017. Soap operas as models of authentic conversations: Implications for materials design. In B. Tomlinson, Maley, A. (Eds), *Authenticity in Materials Development for Language Learning*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp.158-175.

Pallotti, G. 2015. A simple view of linguistic complexity. *Second Language Research*, 31 (1): 117- 134.

Pavesi, M. 2015. From the screen to the viewer-learner. Audiovisual input as a context for second language acquisition. In Campagna, S., Ochse, E., Pulcini, V., Solly, M. (eds) *Languaging in and across Communities: New Voices, New Identities. Studies in Honour of Giuseppina Cortese*. Bern: Peter Lang, pp. 83-104.

Pavesi, M., Ghia, E. 2020. Informal Contact with English: A Case-Study of Italian Postgraduate Students. Pisa: Edizioni ETS.

Perego, E., Del Missier, F., Stragà, M. 2018. Dubbing vs. subtitling: Complexity matters. *Target. International Journal of Translation Studies*, 30 (1): 137-157

Scheffler, P., Jones, C., Dominska, A. 2020. The Peppa Pig television series as input in pre-primary EFL instruction: a corpus-based study. *International Journal of Applied Linguistics*, 31: 3-17.

Sockett, G. 2014. *The Online Informal Learning of English*. London: Palgrave MacMillan.

Vanderplank, R. 2010. Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching*, 43 (1): 1-37.

Vanderplank, R. 2020. Video and informal language learning. In Dressman, M., Sadler, R. W. (eds), *The Handbook of Informal Language Learning*, Hoboken: Wiley-Blackwell, pp. 183-201.

Webb, S., Rodgers, M. 2009a. The lexical coverage of movies. *Applied Linguistics*, 30 (3): 407- 427.

Webb, S., Rodgers, M. 2009b. Vocabulary demands of television programs. *Language Learning*, 59 (2): 335-366.

Zago, R. 2019. Complexity in film dialogue. *Le Forme e la Storia*, 12 (1): 123-134.

Traducción automática para lenguas pobres de recursos: el caso del sardo

Gianfranco Fronteddu – *Autonomous University of Barcelona*

Palabras clave: *corpus paralelo, corpus monolingüe, traducción automática, traducción automática estadística, lenguas pobres de recursos, sardo.*

Este trabajo forma parte de un proyecto más grande, objeto de la tesis doctoral del autor y tiene por objetivo describir el proceso de compilación de corpus para entrenar un traductor automático (TA) estadístico genérico sardo-italiano.

Actualmente, existen motores para TA basados en reglas (TABR) para las combinaciones italiano-sardo (Tyers et al. 2017) y catalán-sardo (Alòs i Font, Fronteddu, y Tyers 2017). En la TABR, unos expertos intervienen directamente para crear reglas e introducir datos lingüísticos (Forcada et al. 2011). Siendo el sardo una lengua pobre de recursos, la TABR se consideró, en su momento, la tecnología más adecuada para lenguas de la misma familia (Tyers et al. 2017). Estudios recientes se han centrado en optimizar la calidad de los TA basados en corpus para lenguas con una cantidad limitada de recursos disponibles (Doğru, Martín-Mor, y Aguilar-Amat 2018, Koehn y Schroeder 2007). Si bien la traducción automática neuronal (TAN) es la tecnología de referencia actualmente, esta es accesible sólo para unas cuantas lenguas en el mundo (Melby 2019). La creación de un motor TAE, en cualquier caso, podría ser un pasaje intermedio para alcanzar una tecnología más avanzada como la TAN en el futuro (Sen et al. 2021).

Siendo el objetivo final de este proyecto el entrenamiento de un TAE genérico (*out-of-domain*), el corpus paralelo it-sc puede ser de textos de cualquier tipología. Aunque la cantidad de recursos disponibles en red para el sardo es notable, la existencia de modelos ortográficos diferentes dificulta la reutilización de muchos de los textos existentes para crear un corpus unitario. Además, cabe añadir que la búsqueda y la recogida de recursos textuales requiere un esfuerzo considerable, ya que están esparcidos por la web y en formatos no siempre editables. Por lo tanto, se ha decidido, por razones de coherencia, elegir como modelo la Limba Sarda Comuna (LSC), propuesta ortográfica adoptada por el Gobierno sardo en 2006. En paralelo, para sistematizar y agilizar el proceso de búsqueda y compilación, se recurrió a técnicas de *web scraping* y a la descarga masiva de textos. El procesamiento de los textos incluye la conversión en formatos procesables (desde PDF a ODT, TXT etc.), la limpieza (por ejemplo, quitando elementos no textuales) y, por último, la alineación con programas de traducción asistida por ordenador (TAO).

Siguiendo esta metodología se consiguió un corpus paralelo it-sc que cuenta con 1.010.634 palabras, de textos de fuentes distintas: proyectos de traducción institucional (Floris 2020); tesis de máster bilingüe y productos de localización.

La presentación describirá las características del corpus y se cerrará definiendo las líneas futuras de trabajo: la elección de un motor para entrenar la TAE (Martín-Mor 2017), y la evaluación del rendimiento con métricas automáticas de calidad (Tomás, Mas, y Casacuberta 2003, Papineni et al. 2002). De este modo, esperamos poder formular conclusiones en términos de rendimiento en comparación con los traductores existentes.

Bibliografía

Alòs i Font, Hèctor, Gianfranco Fronteddu, y Francis Tyers. 2017. «Una eina per a una llengua en procés d'estandardització: el traductor automàtic català-sard». *Linguamática* 9 (diciembre): 3-20. https://doi.org/10.21814/-lm.9.2.255.

Dogru, Gokhan, Adrià Martín-Mor, y Anna Aguilar-Amat. 2018. «Parallel Corpora Preparation for Machine Translation of Low-Resource Languages: Turkish to English Cardiology Corpora».

Floris, Flavia Eva 2020. *Tecnologias pro sa tradutzione e limbas minorizadas: is deliberas de su Comunu de Bauladu in sardu*. Facoltà di Studi Umanistici, Università degli studi di Cagliari

Forcada, Mikel L. 2009. «Apertium: traducció automàtica de codi obert per a les llengües romàniques». *Linguamática* 1 (1): 13-23.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz- Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, y Francis M. Tyers. 2011. «Apertium: A Free/Open-Source Platform for Rule-Based Machine Translation». *Machine Translation* 25 (2): 127-44. https://doi.org/10.1007/s10590-011-9090-0.

Koehn, Philipp, y Josh Schroeder. 2007. «Experiments in Domain Adaptation for Statistical Machine Translation». En *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, 224-27. Prague, Czech Republic: Association for Computational Linguistics. https://doi.org/10.3115/1626355.1626388.

Martín-Mor, Adrià. 2017. «MTradumàtica: Statistical machine translation customisation for translators». *Skase Journal of Translation and Interpretation* 10 (1): 25-39.

Melby, Alan K. 2019 «Future of Machine Translation: Musing on Weavers's memo». *The Routledge Handbook of Translation and Technology*. 25: 418-420. Routledge.

Papineni, Kishore, Salim Roukos, Todd Ward, y Wei-Jing Zhu. 2002. «Bleu: a Method for Automatic Evaluation of Machine Translation». En *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-18. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135.

Regione Autonoma della Sardegna. *Limba Sarda Comuna. Norme linguistiche di riferimento a carattere sperimentale per la lingua scritta dell'Amministrazione regionale.* Cagliari: 2006 http://www.regione.sardegna.it/documenti/1_72_-20060418160308.pdf

Sen, Sukanta, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, y Andy Way. 2021. «Neural Machine Translation of Low-Resource Languages Using SMT Phrase Pair Injection». *Natural Language Engineering* 27 (3): 271-92. https://doi.org/10.1017/S1351324920000303.

Tomás, Jesús, Josep Àngel Mas, y Francisco Casacuberta. 2003. «A Quantitative Method for Machine Translation Evaluation». En *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?*, 27-34. Columbus, Ohio: Association for Computational Linguistics. https://aclanthology.org/W03-2804.

Tyers, Francis M., Hèctor Alòs i Font, Gianfranco Fronteddu, y Adrià Martín-Mor. 2017. «Rule-Based Machine Translation for the Italian–Sardinian Language Pair». *The Prague Bulletin of Mathematical Linguistics* 108 (1): 221-32. https://doi.org/10.1515/pralin-2017-0022.

Tweeting in tongues: A multilingual religious corpus on social media

Anna Beatriz Dimas Furtado & Anne O'Connor – *University of Galway*

Keywords: *corpus building, multilingual corpus, corpus of tweets, religious translation.*

Since December 2012, the Pope and his communications team have been producing tweets in 9 languages on an almost daily basis using the @Pontifex accounts, reaching out to millions of followers globally. In maintaining multilingual Twitter accounts for the words the Pope, the platform presents a modern form of the Pentecostal myth of linguistic accessibility, with associated implications for equivalence and symbolic importance of speaking to followers in their own language. In order to understand how the Catholic Church, an extensive global institution, has adapted its commitment to multilingual communication under these on-going technological advances and media, a large corpus is needed. This paper will discuss the process of building the first multilingual corpus of the Catholic Church's tweets under the lens of the Corpus-based Translation Studies (Baker 1993; Oakes and Ji 2012; Li and Hu 2018): it is a medium-size, parallel, multilingual corpus aligned on the document (tweet) level, which contains 35,684 tweets (931,853 words) in English, French, Spanish, Italian, Portuguese, Arabic, German, Polish, and Latin. It is a corpus of translations created by humans and none of the content is machine generated. As it is published on Twitter, the corpus has distinct characteristics in terms of sentence length and the use of linking apparatus such as hashtags. The paper will discuss how the corpus has been created and how the tweets have been aligned. It will further discuss the challenges faced when building a multilingual corpus from Twitter and tools used to work and align multilingual tweets. The second part of the presentation will explore the motives behind the gathering and creation of this corpus and the insights it can provide on institutional use of social media. We will present elements of the methodology to be implemented for the interrogation and analysis of the corpus, and the questions to be posed regarding, for example, metaphorical language, register, readability, accessibility, sentiment analysis and persuasive language.

References

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. Text and technology: *In honour of John Sinclair.* M. Baker, F. Gill and E. Tognini-Bonelli. Amsterdam, John Benjamins. 250**:** 233-248.

Li, X. and K. Hu 2018. Corpus-based Critical Translation Studies: Research Areas and Approaches. *Meta* 63(3): 583-603.

Oakes, M. P. and M. Ji, Eds. 2012. *Quantitative Methods in Corpus-Based Translation Studies.* Amsterdam, John Benjamins.

A corpus approach to the construction of Violence Against Women in the US press during 2015–2020

Miguel Fuster Márquez & Carmen Gregori Signes – *University of Valencia*

Keywords: *Gender Violence, Discursive News Values Analysis, Corpus Linguistics, Critical Discourse Analysis.*

Lazar (2018: 375-7) argues that Violence Against Women (VAW, one of "the most overt forms of dominance in relation to (hetero)sexuality" according to Motschenbacher (2018: 388) is a key research area in Feminist Critical Discourse Studies. Tabbert (2015: 1) claims that "the media mirrors and at the same time perpetuates predominant perceptions of crime in society." The aim of this paper is to analyse how journalists report on cases of VAW, since arguments about victims and perpetrators in cases of gender violence and how these are construed by journalists can have an impact on readers' perceptions. As Easteal et al. (2022) claim, journalists themselves may be biased when reporting on VAW, for example, in their treatment of the attribution of blame. This contribution explores how the relationship between victims and their perpetrators is reported by the press in the NEWGEN-US Corpus (2015-2020). NEWSGEN-US contains 4,5 million words of American news published between 2015 and 2020 in three prestigious American papers: *The New York Times*, *The Boston Globe* and *The Washington Post*. This is part of a larger corpus on VAW which contains samples of journalistic production from Spain, the UK and the US in the last four decades. Thanks to the corpus metadata, researchers can carry out multiple kinds of crosslinguistic and diachronic corpus research. For this analysis, the annotated subcorpus was uploaded onto the online corpus software platform Sketch Engine. Various techniques were needed to explore this aspect, namely word frequency, word sketch and concordancing, which allowed for closer contextualized qualitative analysis. Collocational in- built statistics offered precise quantification of relevant linguistic patterns which required more in-depth inspection. In fact, we took as a point of departure the break-down of the very frequent lemma WOMAN. The analysis paid attention to the collocational behaviour of this social actor in order to see what kind of discursive relationships are established in journalistic news on VAW. To do so we have adopted a Discursive News Values Analytical (DNVA) approach in combination with a corpus methodology (see also Bednarek & Caple 2014, 2017; Potts et al. 2015, AUTHOR & AUTHOR 2019). According to Cotter (2010:73), news values such as Timeliness, Eliteness, Impact, Negativity, Personalisation, etc. govern each stage of the reporting and editing process. Two different frequencies and collocational behaviours were revealed around the lemma WOMAN: while singular *woman* can be used to single out VAW cases and represents the news value of personalisation, the far more frequent *women* is often discursively construed by journalists to highlight general VAW patterns, as in "One in three women is sexually assaulted on the long journey north." (The Washington Post, 2019).

References

Bednarek, M. & Caple, H. 2014. Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. *Discourse & Society* 25(2): 136-158.

Bednarek, M. & Caple, H. Potts 2017. *The Discourse of News Values: How News Organizations create Newsworthiness.* O.U.P.

Cotter, C. 2010. *News Talk: Investigating the Language of Journalism* DOI:10.1017/CBO9780511811975. CUP.

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

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

Motschenbacher, H. 2018. Sexuality in critical discourse studies. In Flowerdew, J. & E. Richardson (eds.) *The Routledge Handbook of Critical Discourse Studies*, Routledge, pp. 388- 402.

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.

Tabbert, U. 2015. *Crime and Corpus. The linguistic representation of crime in the press*. John Benjamins.

Desde el Reino de las Españas hasta el estado de las autonomías: análisis diacrónico del lenguaje constitucional español

Giovanni Garofalo – *Università di Bergamo*

Palabras clave: *constitucionalismo español, lenguaje constitucional, diacronía del español jurídico, tendencias, estudios del discurso asistidos por corpus.*

A partir del Estatuto de Bayona (1808), primer embrión de Constitución, la historia del constitucionalismo en España "no ha sido ni homogénea ni estable" (González-Ares 2010: 14). El Estado liberal español se fue configurando bajo la hegemonía de una oligarquía integrada por la nobleza y la alta burguesía que, con la connivencia de la Corona y del Ejército, se opuso a las crecientes demandas de democracia de la pequeña burguesía y de la clase obrera. La profunda inestabilidad de los regímenes constitucionales instituidos en España y los múltiples movimientos revolucionarios y pronunciamientos militares que se produjeron a lo largo del tiempo han llevado a algunos historiadores a afirmar que "nuestro pasado constitucional […] es la sucesión de los intentos de los sectores progresivos por reformar el sistema institucional de la oligarquía […] y forjar un sistema estatal de signo liberal democrático" (Solé Turá y Aja 2009: 136).

Este trasfondo histórico es el necesario punto de anclaje de la presente investigación, basada en un corpus que reúne las ocho constituciones que entraron en vigor en España a lo largo de las dos últimas centurias (1812, 1834, 1837, 1845, 1869, 1876, 1931, 1978), a las que se han añadido otros seis textos, como el mencionado Estatuto de Bayona y los demás proyectos constitucionales que no llegaron a aplicarse, a saber, el de Bravo Murillo de 1852, el de 1856 o 'Constitución non nata', el de 1873 o 'Constitución de la Primera República', el de Primo de Rivera de 1929. Por último, por su relevancia constitucional se han considerado también las Leyes Fundamentales del Reino, promulgadas por Franco entre 1936 y 1967. En total, el corpus compilado contiene catorce textos y 103.762 palabras y ofrece una interesante ventana a la evolución del constitucionalismo español, de sus conceptos medulares y de su lenguaje.

Tras la necesaria anotación diacrónica del corpus (*timestamps*), el análisis se ha llevado a cabo acudiendo a la herramienta *Tendencias (Trends)* de la plataforma Sketch Engine (Kilgarriff *et al.* 2015), realizando búsquedas con un umbral de frecuencia mínima igual a 2 y un *P-value* máximo de 0,05. De esta manera, se han extraído 120 lemas que manifiestan una frecuencia creciente a lo largo del intervalo temporal de referencia (1808-1978) y 57 lemas cuyo uso mengua o desaparece en el tiempo.

El análisis de concordancia de los lemas con tendencia creciente acusada (p. ej., *profesional*, *derecho*, *jurídico, competencia*, *garantizar*) refleja el difícil camino hacia la modernización y la democratización del Estado y la importancia cada vez mayor atribuida por los constituyentes a la formación académica y profesional de la ciudadanía y a la salvaguarda de las garantías y de los derechos fundamentales, cuya tutela empieza a asomar en la Constitución democrática de 1869 y va afianzándose en la Constitución de la Segunda República y en la de 1978. En cambio, observando el contexto de uso de los lemas con tendencia decreciente (p. ej., *Españas*, *potencia*, *pluralidad [de votos], contribuciones, ramos [de la Administración]*) es posible apreciar los cambios diacrónicos de algunos conceptos fundamentales del orden constitucional –p. ej., las nociones de 'territorio', 'ciudadanía' o 'país vecino'– y el progresivo abandono del centralismo administrativo borbónico. Asimismo, entre mediados del siglo XIX y principios del XX, se observa un interesante proceso de especialización léxica y una gradual sustitución de palabras vagas y ambiguas o de verbos 'comodín' polisémicos (*llegar*, *hacer, volver, dar, tomar*) por términos y locuciones verbales con un grado de especialización cada vez mayor (Ullmann 1983: 194), acordes con las convenciones del lenguaje constitucional moderno.

Referencias

González-Ares, José Agustín. 2010. Las constituciones de la España contemporánea. Del Estatuto de Bayona a las Leyes Fundamentales del franquismo. Santiago de Compostela: Andavira.

Kilgarriff, Adam / Herman, Ondřej / Bušta, Jan / Rychlý, Pavel / Jakubíček, Miloš. 2015. "DIACRAN: a framework for diachronic analysis". In *Corpus Linguistics (CL2015)*, United Kingdom, July 2015.

Solé Turá, Jordi / Aja, Eliseo 2009. *Constituciones y períodos constituyentes en España (1808-1936)*. Madrid: Siglo XXI de España Editores.

Ullmann, Stephen. 1983. Semantics. An Introduction to the Science of Meaning. Oxford: Basil Blackwell.

The use of nominalizations in noun-noun phrases by L1 Spanish EFL teachers

Roger Gee[1], M. Karen Jogan[2] & Kathleen Jogan[3]

Holy Family University[1] – Albright College[2] – University of Arkansas[3]

Keywords: *nominalizations, noun phrases, academic writing, premodifiers, EFL teachers.*

Nominalizations are an important feature of academic language and are a "distinctive characteristic" of academic writing (Biber & Gray, 2022, p. 139). Nominalizations "allow authors both to pack information into fewer words and, more importantly, to linguistically reorganize everyday experience into new categories that can be further discussed and elaborated" (Hyland & Jiang, 2021, p. 2).

One English structure where nominalizations appear is in noun-noun phrases (NNPs). Nominalizations are used in both the first noun and the second noun, the head noun modified by the first noun (Biber & Gray, 2022). NNPs, like nominalizations, allow writers to compress information, an important element in a register-functional approach in which linguistic units are embedded in phrases that may be "characterized as 'structurally compressed'" (Biber et al, 2022, p. 3). The purpose of the present research is to investigate the use of nominalizations as prenominal noun modifiers.

Nominalizations and NNPs have usually been considered separately, not in the context of nominalizations as constituents of NNPs, and much of the extant research has been carried out with student writing and specialist expert writing (Díez-Bedmar & Pérez-Paredes, 2020; Fang et al., 2021; Gourlart, 2021; Hyland & Jiang, 2021). In contrast, the present research focuses on nominalizations in NNPs produced by Spanish L1 EFL teachers (presumably advanced learners of English) who were divided into three advanced proficiency levels. As Spanish has a more complex system of affixes than does English, it would be expected that the teachers in this study were proficient with both the form and the use of nominalizations. On the other hand, NNPs develop at a late stage in Biber et al.'s (2022) description of academic language development. Thus, the present research offers insight into the hypothesis of a threshold level "beyond which variation in morphological complexity is no longer related to learners' linguistic ability in their L2" (Brezina & Pallotti, 2019, p. 115).

The corpus used in this research contains all 48 essays about teaching English during the pandemic that were submitted prior to a national conference, with 48,187 total words tagged with the TreeTagger tagset. The essays were divided equally into three significantly different groups, low-, medium-, and high-advanced proficiency, based on use of vocabulary beyond the 3000-word level of the BNC/COCA wordlists. The three distinct groups provided a cross-sectional design that allowed developmental insights. NNPs were retrieved using AntConc, and nominalizations were manually identified. There were three research questions:

1. To what extent do Spanish L1 EFL teachers use nominalizations in NNPs?
2. Are verb or adjective stems more frequent in forming nominalizations?
3. Is there a difference in raw frequency use of nominalizations among proficiency groups?

Preliminary results suggest that

1. Over 50% of the head nouns in NNPs were nominalizations for all proficiency groups and approximately 45% of the premodifying nouns were nominalizations for the low- and high- advanced proficiency groups.
2. Verb stems were more frequent as only a total of eight adjective stems were used by the mid- and high-advanced proficiency groups while there were none in the low-advanced group.

3. Raw frequencies of nominalizations were progressively greater by proficiency group with the mid-advanced using more nominalizations than the low-advanced and the high- advanced using more than the mid-advanced.

Results will be related to usage-based construction grammar. Implications for instruction will conclude the presentation.

Bibliography

Biber, D., Gray, B., Staples, S., & Egbert, J. 2022. The register-functional approach to grammatical complexity: Theoretical foundation, descriptive research findings, application. Routledge.

Biber, D. & Gray, B. 2022. Nominalizing the verb phrase in academic science writing. In The register-functional approach to grammatical complexity: Theoretical foundation, descriptive research findings, application (pp. 176-198). Routledge.

Brezina, V. & Pallotti, G. 2019. Morphological complexity in written L2 texts. *Second language research,* 35(1), 99-119.

Díez-Bedmar, M. B. & Pérez-Paredes, P. 2020. Noun phrase complexity in young Spanish EFL learners' writing: Complementing syntactic complexity indices with corpus-driven analyses. *International Journal of Corpus Linguistics,* 25(1), 4-35.

Fang, Z., Gresser, V., Cao, P., & Zheng, J. 2021. Nominal complexities in school children's informational writing. *Journal of English for Academic Purposes,* 50, 100958.

Goulart, L. 2021. Register variation in L1 and L2 student writing: A multidimensional analysis. *Register Studies,* 3(1), 115-143.

Hyland, K. & Jiang, F. 2021. Academic naming: Changing patterns of noun use in research writing. *Journal of English Linguistics,* 49(3), 255-282.

Nation, I. S. P. 2012. *The BNC/COCA word family lists*. Retrieved from https://www.laurenceanthony.net/software/antwordprofiler/

Creativity in popular music criticism: A diachronic corpus-based analysis of *Rolling Stone* album reviews

Gilberto Giannacchi – *Università degli Studi dell'Insubria*

Keywords: *corpus linguistics, critical genre analysis, popular music criticism.*

Popular music criticism represents a culturally significant artifact which has generated interest in popular culture studies and musicology (Shuker 2001, Jones & Baker 2002). A key role in the establishment of such practice as a pop culture staple has been played by the American magazine *Rolling Stone*. This magazine's album reviews and 'top 100… of all time' lists can be considered a prototypical outcome of popular music criticism (Schmutz 2005). Despite being influent and creatively stimulating, these texts have never been systematically investigated from a linguistic perspective.

This study aims to offer a diachronic corpus-based overview on the language used in *Rolling Stone* album reviews to highlight creative uses of vocabulary and ubiquitous stylistic strategies which may help to define contemporary music reviews as a textual genre and provide insight into the genre's professional culture (Bhatia 2004). To this end, a sample corpus of *Rolling Stone* album reviews published over a 32-year time-span (1990-2022) was investigated – 15394 word tokens, 5037 word types. To account for balance and variability, reviews by different authors were selected to make up the corpus. Moreover, the texts deal with different music genres to obtain a multifaceted and as objective as possible lexical overview. The corpus analysis was carried out with AntConc (Anthony 2022), mainly focusing on concordance strings and key words in context (KWIC).

Findings show a slow, yet consistent, decrease in the use of creative imagery to evaluate albums. Creativity in language is mainly observable in the reviews published in the 1990s and the 2000s. Reviewers recontextualize adjectives that are not commonly relevant to the semantic field of music (Semino et al. 2013). The widespread adoption of this stylistic strategy creates intertextuality and conveys a sense of familiarity to the reader. Music-related terms often form synesthesias – e.g., 'pumped *guitars*' (Gold 1994), 'sputtering *guitar*', 'angular *pop* melodies' (Ganz 2008) – or contribute to creating evocative imagery – e.g., 'layers of warped harmonic *guitar* noise' (Robbins, 1992), 'the clattering *drums, trombones* and impasto of underwater *guitar fuzz*' (Ratliff 1998). These instances of semantic recontextualization stem from Hunter S. Thompson's gonzo journalism style, which openly rejected objectivity in favor of the author's own sensorial experience (Hoover 2009). Gonzo journalism played a key role in the development of popular music criticism (Jacke *et al.* 2014), although it has partly lost its influence in latter years. The most recently published reviews in the corpus are presented with a more traditional journalistic approach. They feature fewer dashing uses of language – e.g., in both collocations and syntax – and privilege 'track to track' and background descriptions. This could signal a change in the professional and cultural environment in music journalism – particularly as for the authors' authority and roles.

This study's framework might possibly be replicated on larger music reviews corpora to obtain more statistically relevant results. It would also be interesting to compare the genre's stylistic conventions between different publishing platforms – e.g., newspapers, music magazines and webzines. Further attention can also be paid to intertextual relationships with other genres – e.g., newspaper articles, academic writing, advertising.

Bibliography – Primary Sources

Drozdowski, T. 1995, March 8. The Bends. *Rolling Stone*. https://www.rollingstone.com/music/ music-album-reviews/the-bends-2-126433/

Ganz, C. 2008, October 30. Deerhunter: Microcastle. *Rolling Stone*. https://web.archive.org/web/-20090303162951 + http://rollingstone.com/reviews/album/22190741/review/23589022/microcastle

Gold, J. 1994, March 24. The Downward Spiral. *Rolling Stone.* https://www.rollingstone.com/ music/music-album-reviews/the-downward-spiral-187149/

Ratliff, B. 1998, February 13. In The Aeroplane Over The Sea. *Rolling Stone. https:// www.rollingstone.com/music/music-album-reviews/in-the-aeroplane-over-the-sea-112548/*

Robbins, I. 1992, March 25. Loveless. *Rolling Stone. https://www.rollingstone.com/music/music- album-reviews/loveless-251215/*

Secondary Sources

Anthony, L. 2022. AntConc (Version 4.1.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Bhatia, V. K. 2004. Worlds of written discourse: A genre-based view. A&C Black.

Jacke, C. & James, M. & Montano, E. 2014. Editorial Introduction: Music Journalism. 4. 1-6. 10.5429/2079-3871-v4i2.1en.

Jones, S. & Baker, S. S. (Eds.). 2002. *Pop music and the press* (Vol. 12). Temple University Press.

Hoover, S. 2009. Hunter S. Thompson and Gonzo Journalism: a guide to the research. *Reference services review.*

Schmutz, V. 2005. Retrospective cultural consecration in popular music: Rolling Stone's greatest albums of all time. *American Behavioral Scientist*, *48*(11), 1510-1523.

Semino, E., Deignan, A., & Littlemore, J. 2013. Metaphor, genre, and recontextualization. *Metaphor and Symbol*, *28*(1), 41-59.

Shuker, R. 2001. *Understanding Popular Music* (second edition). Abingdon, Oxon: Routledge.

Using corpus linguistics to assess the evolution of Plain English in institutional language: The case of the Scottish Ombudsman

Gabriel González-Delgado – *University of Alicante*

Keywords: *institutional language, Plain English, comprehensibility, Ombudsman, Corpus Linguistics, diachronic variation.*

The language produced by the Public Administration is usually referred to as 'officialese', 'bureaucratic language' (Anderson, 2003, p. 11), or —in a more pejorative sense— 'gobbledygook' because of the comprehension difficulties that it poses to citizens in general. The fact is that preeminent values in legal language such as objectivity and accuracy have some pragmatic implications that on many occasions obscure the overall meaning (Crystal & Davy, 2013). As a consequence, Plain Language campaigns appeared to advocate for the simplification of legal language, and especially, of institutional language. While the language of the courtroom or the language of the Law have been deeply studied since the mid-late 20th century (Bhatia, 1987; Maley, 1987; Mellinkoff, 1963), little is known about the language of the Public Administration. A thorough analysis of this technical language is the threshold for defining and applying effective simplification mechanisms. Parallel to the development of the Plain English movement, Ombudsman offices were established in Anglo-Saxon countries as guarantors of citizens' rights against maladministration. Even though they lack coercive powers to impose their decisions like judges do, Omdusmen act as whistle-blowers by raising concerns and holding public organisations accountable for their actions or inactivity (Remac & Langbroek, 2011). This also includes the right to be addressed in a readable, understandable way by official bodies, including Ombudsman offices themselves.

The aim of this paper is twofold. On the one hand, I will define the linguistic characteristics of the language employed in its case reports by the Scottish Public Services Ombudsman, as a sample of a Human Rights Agency (HRA). On the other hand, I will analyse its diachronic variation within an 18-year spam to measure the degree of application of plain language precepts since the beginnings of the Scottish Ombudsman up to date. Within English for Specific Purposes (ESP), Corpus Linguistics provides both the theoretical framework and the tools to achieve such type of analyses (Bowker & Pearson, 2002; Flowerdew, 2005; McEnery & Gabrielatos, 2006; Nesi, 2013). Not only are lexical units relevant in regard to comprehensibility, but syntactic structures play a central role in this respect as well (Bhatia, 1983). Consequently, the complexity of lexicon will be closely related to keywords (among other variables), while the syntactic analysis will focus on complex structures such as subordination or passivity. The corpus will be proportionally built out (Biber, 1993) of about 100.000 words from three different thematic domains: Education, Health, and Housing. Besides, each text included in the corpus will be marked-up in reference to its year of production, so that its variability through time may be determined. Results make it possible to conclude that the language of the Scottish Public Services Ombudsman may be categorised within legal language because of its linguistic features. Nevertheless, there are some characteristics that may only be attributed to this type of institutions. Finally, results also suggest that there exists a slight tendency towards linguistic simplification, both syntactically and lexically.

Bibliography

Anderson, W. J. 2003. A corpus linguistic analysis of phraseology and collocation in the register of current European Union administrative French. https://research- repository.st-andrews.ac.uk/handle/10023/4909.

Bhatia, V. K. 1983. Simplification v. easification - the case of legal texts. Applied Linguistics, 4(1), 42–54. https://doi.org/10.1093/applin/4.1.42.

Bhatia, V. K. 1987. Language of the law. Language Teaching, 20(4), 227–234. https://doi.org/10.1017/-S026144480000464X.

Biber, D. 1993. Representativeness in Corpus Design. Literary and Linguistic Computing, 8(4).

Bowker, L. & Pearson, J. 2002. Working with Specialized Language. In Working with Specialized Language. https://doi.org/10.4324/9780203469255.

Crystal, D. & Davy, D. 2013. Investigating English Style (R. Quirk (ed.)). Routledge.

Flowerdew, L. 2005. An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. English for Specific Purposes, 24(3), 321–332. https://doi.org/10.1016/j.esp.2004.09.002.

Maley, Y. 1987. The Language of Legislation. 16(1), 25–48.

McEnery, T. & Gabrielatos, C. 2006. English Corpus Linguistics. In B. Aarts & A. McMahon (Eds.), The Handbook of English Linguistics (pp. 33–71). Blackwell Publishing. https://doi.org/10.1017/CBO-9781107415324.004.

Mellinkoff, D. 1963. The Language of the Law. Little, Brown.

Nesi, H. 2013. ESP and Corpus Studies. In B. Paltridge & S. Starfield (Eds.), The Handbook of English for Specific Purposes (pp. 407–426). Wiley-Blackwell. https://doi.org/10.1002/9781118339855.ch16.

Remac, M. & Langbroek, P. M. 2011. Ombudsman' Assessments of Public Administration Conduct: Between Legal and Good Administration Norms. NISPAcee Journal of Public Administration and Policy, 4(2), 87–115. https://doi.org/10.2478/v10110-011-0005-5.

El uso de adverbiales en textos instructivos del siglo XIX

Carolina González-Quintana & Ivalla Ortega-Barrera – *University of Las Palmas de Gran Canaria*

Palabras clave: *metadiscurso, adverbios, género, evidencialidad, perspectiva.*

La investigación llevada a cabo se ha centrado en muestras de textos instructivos ingleses escritos en el siglo XIX con el fin de evaluar los usos y funciones de las formas adverbiales como mecanismos metadiscursivos en estos textos, según la variable de género. Existen estudios previos sobre los rasgos del metadiscurso en textos de diversos periodos de la lengua inglesa (cf. Gray, Biber y Hiltunen 2011; Alonso-Almeida y Mele-Marrero 2014; Álvarez-Gil 2017). Siguiendo esta tradición, nos centramos en dichas formas adverbiales en el sentido desarrollado en Hyland (2005).

Se estudiarán también algunos aspectos que se relacionan con el metadiscurso, como es el significado evidencial codificado en estas formas adverbiales. Mientras que, para algunos investigadores, la evidencialidad representa un subdominio de la modalidad epistémica, hay quienes consideran que la evidencialidad representa una categoría independiente, o incluso yuxtapuesta (cf. Dendale y Tasmowski 2001). En este trabajo, sin embargo, seguimos el enfoque disyuntivo en consonancia con Cornillie (2009), quien sustenta que el modo de conocer no debe asociarse con el grado de compromiso de los autores hacia sus textos, lo que también propone Alonso-Almeida (2015).

La razón para elegir los adverbios como recursos lingüísticos objeto de este análisis reside en el hecho de que estos son una de las categorías gramaticales que más claramente contribuyen a la expresión de significados interactivos (Biber y Finegan, 1988). Se describirá, por eso, su uso por parte de autores de textos instructivos del siglo XIX con el fin de caracterizarlos en términos de impronta y comprobar, por lo tanto, cómo los utilizan para intercambiar significados interaccionales con sus lectores potenciales. El uso de los textos instructivos responde a que, cuantitativamente, hay mayor número de estos escritos por mujeres en los repositorios dada la concepción que se tenía de la mujer fuera del ámbito doméstico. El siglo XIX presenta en este sentido una fuente importante de estos textos, pues las mujeres podían publicar con cierta libertad acerca de temas asociados a su labor, sobre todo, en la gestión doméstica.

Nuestro interés es, en definitiva, explorar el uso de estos adverbios y las diferentes funciones pragmáticas que cumplen en los textos instructivos analizados. Para ello, hemos empleado el subcorpus del siglo XIX del *Corpus of Women's Instructive Writing in English*, es decir, Co-WITE19, que contiene extractos de varios textos con finalidad didáctica escritos entre 1800 y 1899, utilizando herramientas de lingüística de corpus para recuperar formas adverbiales de manera computarizada. Los mecanismos adverbiales se cuantifican y agrupan según los grados de certeza y probabilidad. Esperamos que los resultados muestren que, según el contexto, estas estructuras pueden cumplir varias funciones pragmáticas, como la indicación de distintos grados de compromiso o distanciamiento del autor hacia la información presentada, la persuasión y la cortesía, entre otras, y que exista distinción según el género de autoría.

References

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

Alonso-Almeida, Francisco. 2015b. Introduction to stance language. *Research in Corpus Linguistics* 3: 1-5.

Alonso-Almeida, Francisco and Margarita Mele-Marrero. 2014. 'Stancetaking in seventeenth-century prefaces on obstetrics', *Journal of Historical Pragmatics* 15(1): 1–35.

Álvarez-Gil, F. J. 2017. Apparently, fairly and possibly in The Corpus of Modern English History Texts (1700-1900). In F. Alonso-Almeida (ed.). In Stancetaking in Late Modern English Scientific Writing. Evidence from

the Coruña Corpus. Colección Scientia [Applied Linguistics]. Valencia: Servicio de Publicaciones de la Universidad Politécnica de Valencia.

Biber, Douglas and Edward Finegan. 1988. Adverbial stance types in English. *Discourse Processes* 11.1, 1-34.

Cornillie, Bert. 2009. Evidentiality and epistemic modality. On the close relationship between two different categories. *Functions of Language* 16.1, 44-62.

Dendale, Patrick and Liliane Tasmowski. 2001. Introduction: Evidentiality and related notions. *Journal of Pragmatics* 33, 339-348.

Diewald, G., M. Kresic and E. Smirnova. 2009. "The grammaticalization channels of evidentials and modal particles in German: Integration in textual structures as a common feature" in Hansen, M.M. and Visconti J. (eds.) *Current Trends in Diachronic Semantics and Pragmatics.* UK: Emerald, 189-209.

Gray, B., Douglas, B. and T. Hiltunen. 2011. The expression of stance (1665-1712) in early publications of the Philosophical Transactions and other contemporary medical prose: Innovations in a pioneering discourse. In Irma Taavitsainen and Päivi Pahta (eds.). *Medical writing in Early Modern English*. Cambridge: Cambridge University Press, 221-247.

Hyland, Ken. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies* 7(2): 173–192.

Diphthong Shift: The representation of a typical southern-English trait in the Lancashire dialect

Nadia Hamade-Almeida – *Camilo José Cela University*

Keywords: *diphthong shift, Lancashire dialect, literary-dialect texts, dialect representation, non-standard spellings.*

Diphthong Shift is a phonological change that originated in the beginning of the nineteenth century in London, and later expanded to southern England and the Midlands. This shift affected several RP monophthongs and diphthongs: [iː], [uː], [ɛɪ], [aɪ], [ɔɪ], [aʊ] and [əʊ]. Considering that Diphthong Shift is a southern-English linguistic trait and that the dialect of Lancashire preserves much of the North Midlands (Ellis: 1889), the aim of this paper is to observe the reach and effect of this shift in the dialect of the nineteenth-century Lancashire. As evidence of this phonological change at that time is attained via written texts, this paper manually examines several nineteenth-century Lancashire literary-dialect texts composed by five distinct authors who were born in Lancashire, except two of them who were not born in this county but were believed to have mastered the Lancashire dialect.

Literary-dialect texts are distinguished by the portrayal of traditional dialect. This concept refers to the conservative varieties spoken in certain areas of England, such as northern England. These works, which are characterized by the presence of semi-phonetic spellings based on Standard English, are considered a useful tool in dialect study. A meticulous analysis of the non-standard spellings represented may give an insight into the dialect features and sounds of a particular vernacular variety at a specific period of time. This paper attempts to analyze these non-standard spellings and subsequently attribute them to their corresponding pronunciations in the dialect. As a complete examination of sounds and spellings is beyond the scope of this paper, this study focuses on the group of words related to the aforementioned RP sounds, those involved in Diphthong Shift.

When obtaining phonological data of a particular sound change via the examination of literary-dialect works, two distinct issues may arise. On the one hand, since the traditional dialect is the variety depicted in the selected works, there might be a proneness to use old or archaic sounds that were probably regressive during the nineteenth century. As a result, more recent realizations, as those resulting from Diphthong Shift, may be less coveted. On the other hand, drawn on the writers' absence of linguistic accuracy, there may appear sounds that do not relate to the Lancashire dialect. These pronunciations could respond to the authors' whimsical preference or, on the contrary, they would be associated with other geographical areas. This is particularly relevant considering that two of the authors of the corpus were not born in Lancashire and would have adopted pronunciations usual of their original birthplace.

The results show that merely the group of words connected with RP [aɪ] and [aʊ] or the PRICE and MOUTH lexical sets, respectively, according to the classification Wells (1982) provides for both groups, are affected by Diphthong Shift. In addition, the results reveal that these two lexical sets exhibit in the dialect a coexistence of old pronunciations and recent realizations, which, in the latter case, are the result of this phonological sound change.

References

Ellis, John Alexander, 1889. *On Early English Pronunciation*. London: Asher & Co. Print.

Wells, John Christopher, 1982. *Accents of English*. Vol. 1. Cambridge: Cambridge University Press.

Alternative second person pronouns in English: A corpus-based study of their number reference

David Hernández-Coalla – *University of Vigo*

Keywords: *non-standard pronouns, number, second person, varieties of English.*

Traditionally, the English pronominal paradigm only lists a single second person pronoun: *you*. With the infrequent form *thou* limited to archaic speech and fossilized expressions in religious contexts, it has been argued that English native speakers do not feel that the absence of distinctive forms to mark singular and plural reference constitutes a gap in the system. Nevertheless, the emergence of alternative second person pronominal forms in several varieties of the language calls into question statements such as this and demands further study backed by empirical data.

Wales (1996) provides a comprehensive list of the different pronouns found across the English-speaking world, including several forms with a second person reference; many of these have actually been included in the OED, such as *oonu*, *du*, *you-uns*, *yinz*, etc. Although these non-standard pronouns can be used both with a singular and a plural reference, it seems that they could be on their way to losing their singular-marking capacity, retaining only their plural connotation in contrast with the traditional *you*. A prospective search on the corpus of *Global Web-based English* (GloWbE; Davies 2013) has retrieved a considerable number of instances where *you* is followed by a verb which appears to contain the third person singular ending of the simple present, which could indicate the increasing singular connotation of the pronoun. This trend could also be supported by the spread of periphrastic pronominal forms such as *you guys* or the more common *you all*, especially in informal contexts.

This paper aims to provide an overview of the varieties in which these alternative second person pronouns are attested and study their reference to conclude whether they are used in opposition to *you*. Periphrastic combinations will be analysed in detail to elucidate their status and determine whether they can actually be classified as pronouns (Valentínyová 2015) or instead they are the result of a combination of the standard form *you* and a quantificational adjunct (Huddleston and Pullum 2002). To this purpose, data will be extracted from the GloWbE, given its broad coverage of geographical varieties and word count. The composition of the selected corpus will also allow for the study of any potential influence of a particular variety on another one. Since the GloWbE consists of thousands of webpages, and due to the role of the internet in our globalized world as the meeting point of individuals of all nationalities, it will be possible to determine whether a certain pronoun has extended beyond its place of origin and has been incorporated into more varieties. This question is of paramount importance in the case of *you all*, which in spite of being ascribed to varieties of the American South, seems to have expanded across the English-speaking countries, perhaps due to U.S. influence. The answers to these questions will help to refine our current knowledge of the English pronominal system.

Bibliography

Davies, Mark. 2013. *Global Web-Based English Corpus* (GloWbE). Available at: https://english-corpora.org/glowbe/.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Oxford English Dictionary Online. 2022. https://www.oed.com.

Valentínyová, Kristína. 2015. *Non-standard forms of the pronoun "you" in English*. Prague: Univerzita Karlova [Unpublished Bachelor's Dissertation]. https://dspace.cuni.cz/bitstream/handle/20.500.11956/66219/BPTX-2013_2_11210_0_345119_0_154697.pdf?sequence=1&isAllowed=y

Wales, Katie. 1996. *Personal Pronouns in Present-Day English*. Cambridge: Cambridge University Press.

Super **as a cross-linguistic intensifier**

Chad Howe, Camila Lívio & Katherine Ireland – *University of Georgia*

Keywords: *cross-linguistic intensifiers, language variation and change, corpus linguistics.*

The study of intensifiers is of great interest, especially due to their capacity for rapid change (Ito & Tagliamonte 2003). In English, related work has focused primarily on the interplay between high frequency variants, such as *very athletic* and *really athletic*, and the emergence of newer variants like *so* (Ito & Tagliamonte 2003, Tagliamonte & Roberts 2005, Brown & Tagliamonte 2012). Additional English-focused studies have analyzed the use of intensifiers by particular language varieties (Canadian and New Zealand English) or groups of speakers (Bauer & Bauer 2002; Macaulay 2006; Tagliamonte 2008). Tagliamonte argues that the rapid changes involved in intensifiers, as shown in Canadian English speakers, suggest that the linguistic mechanisms involved are highly "complex". Waksler (2012) further documents the utility of intensifiers for highlighting subjectivity in discourse and contexts, with a range of formalities and ages of users. Other work on intensifiers has focused on Spanish, displaying similar findings, with an interplay of sociolinguistic in addition to linguistic variables affecting changes in use (Kanwit et al. 2018; Brown & Cortés-Torres 2013). Despite this, more recent changes in intensifiers are understudied, especially in their tendencies for cross-linguistic generalization. This proposed paper examines the distribution of *super* as an intensifier, providing comparative corpus evidence from English, Portuguese, and Spanish. We find that *super* exhibits parallel patterning across these languages.

The data used in this analysis were extracted from three of the corpora available through SketchEngine (Kilgarriff et al. 2014): English Web 2020 (enTenTen20), Portuguese Web 2018 (ptTenTen2018), and Spanish Web 2018 (esTenTen18). The part of speech tagging in the corpus facilitated the process of extracting collocates by focusing on the structure **SUPER** + Adjective.

(1) Woah, this looks **super** interesting and I love the art as well (enTenTen20)

(2) *Achei* ***super*** *interessante o projecto* (ptTenTen18) 'I find the project super interesting'

(3) *La oferta laboral me parece* ***super*** *interesante* (esTenTen20) 'The work offer seems super interesting to me'

The comparison of the data from English (1), Portuguese (2), and Spanish (3) reveals several key patterns. First, examination of the individual adjectives suggests a high rate of cross-linguistic similarity between specific cognates, as shown above with *interesting/interessante/interesante* in (1-3). We argue that this reflects characteristic parallel extension of intensifiers across languages, with specific lexical items serving as a vector for change. The data from Portuguese and Spanish suggest a variable morphological suffixation that is not observed in the English data (e.g., *superbueno* 'super good' in Spanish; see Foltran & Nóbrega 2016). The use of large-scale corpus data in this paper provides a cross-linguistic view for the distribution of novel intensifiers. Finally, we argue that this work underscores the necessity of additional cross-linguistic corpus- based studies.

Bibliography

Bauer, Laurie and Winifred Bauer. 2002. Adjective Boosters in the English of Young New Zealanders. *Journal of English Linguistics.* 244-257.

Brown, X and Cortés-Torres. 2013. Puerto Rican intensifiers: Bien/Muy Variables. *Selected Proceedings of the 6th Workshop on Spanish Sociolinguistics,* edited by Ana M. Carvalho and Sara Beaudri, Cascillada Proceedings Project. 11-19.

Brown, L. & Tagliamonte, S. A. 2012. A Really Interesting Story: The Influence of Narrative in Linguistic Change. University of Pennsylvania *Working Papers in Linguistics.* 18(2): 1–10.

Foltran, M. J. & Nóbrega, V. A. 2016. Intensifier adjectives in Brazilian Portuguese: Properties, distribution, and morphological reflexes/Adjetivos intensificadores no Português Brasileiro: Propriedades, distribuição e reflexos morfológicos. *Alfa: Revista de Linguística,* Vol, 2, 319–340.

Ito, R. & Tagliamonte, S. A. 2003. Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society*. 32:257–279.

Kanwit, Matthew, Vanessa Elias, and Rebecca Clay. 2018. Acquiring intensifier variation abroad: Exploring *muy* and *bien* in Spain and Mexico. *Foreign Language Annals.* 455-471.

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.

Tagliamonte, S. & Roberts, C. 2005. So weird; so cool; so innovation: The use of intensifiers in the television series Friends. *American Speech*. 80: 280–300.

Tagliamonte, S. 2008. So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics* 12:2. 361-194.

Waksler, R. 2012. Super, uber, so, and totally: Over-the-top intensification to mark subjectivity in colloquial discourse. In N. Baumgarten, I. Du Bois, & J. House (Eds.), *Subjectivity in Language and Discourse,* 17–31. Netherlands: Brill.

The circulation of endometriosis terms: Towards the analysis of the appropriation of terms by laypeople in a comparable corpus

Julie Humbert-Droz – *Université Lumières Lyon 2*

Keywords: *term circulation, endometriosis, comparable corpus, terminological appropriation, corpus compilation.*

The circulation of medical terms raises important issues of understanding and appropriation of terms by laypeople, especially patients (Gill & Maynard 2006, León-Araúz 2015, Delavigne et al. 2022). Many studies show the necessity of making medical knowledge accessible to patients in different text types, such as science popularisation articles and forum posts (Rouillard 2016, Delavigne 2020, Estopà & Montané 2020). In this context, my presentation aims at addressing issues related to the circulation of the terminology of endometriosis in different speech communities, in French. The analysis is based on the exploration of a comparable corpus (Sinclair 1996), which is designed and compiled precisely for this purpose, within the scope of Textual Terminology (Condamines & Picton 2022) and Socioterminology (Gaudin 2003, Humbley 2018).

Endometriosis affects 1 to 2 out of 10 women worldwide and is still incurable (Johnston et al. 2015, Ilschner et al. 2022). It is known that the disease is poorly understood by the general public and that certain misrepresentations keep being conveyed by the media (Young et al. 2015). Therefore, proper diagnosis and treatment can be considerably delayed (Bullo 2020). In this context, my main goal is to better understand the impact of the circulation of endometriosis terms on their reception and appropriation by laypeople and patients through a corpus-based study.

The goal of the presentation is twofold: first, I will discuss the challenges of building a corpus that intends to represent the ways endometriosis terms circulate among different communities, i.e., the ways they are used by experts, patients, and laypeople in different text types. These challenges are related to the inclusion of heterogeneous texts, from sources of different nature (e.g. scientific articles, forum posts, popularisation). This heterogeneity is necessary to reflect the dynamics of term circulation but questions the ideals of corpus balance and representativeness (Biber 1993, Leech 2007). The organisation of the data in different sub- corpora is another challenge of corpus compilation in this case. Indeed, the sub-corpora are meant to represent key stages of term circulation but the number of stages to include in the corpus, hence the number of sub-corpora, should remain manageable. This is necessary to ensure that the corpus is comparable and that the differences in usage and meaning are observable when comparing the sub-corpora.

Second, I will illustrate some of the main differences in meaning between experts and laypeople with examples from the corpus. Preliminary results show that diverging definitions found in the corpus for the same terms, whether in different sub-corpora or in the same sub- corpus, have a strong impact on laypeople's understanding of the terms. I will argue that the diverging definitions that keep being disseminated through various media contribute to the phenomenon of conceptual indeterminacy (e.g. Péraldi 2012). I will also discuss the consequences of indeterminacy on the perception of endometriosis and of its severity.

These results are a first step towards a better understanding of the ways in which the circulation of endometriosis terms affects their reception and appropriation by laypeople and patients. Future work will then focus on refining these first observations in the corpus, which will eventually serve to enhance communication with the public and raise awareness about the disease.

Bibliography

Biber, D. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4), 243-257.

Bullo, S. 2020. "I feel like I'm being stabbed by a thousand tiny men": The challenges of communicating endometriosis pain. *Health*, *24*(5), 476-492.

Condamines, A. & Picton, A. 2022. Textual Terminology: Origins, principles and new challenges. In P. Faber & M.-C. L'Homme (Eds.), *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge* (p. 219-236). John Benjamins.

Delavigne, V. 2020. De l'(in)constance du métalinguistique dans un corpus de vulgarisation médicale. *Corela. Cognition, représentation, langage*, HS-31.

Delavigne, V., Picton, A., & Thibert, E. 2022. Socioterminologie et terminologie textuelle: L'expertise en questions. *Actes du Congrès Mondial de Linguistique Française (CMLF 2022).*

Estopà, R. & Montané, M. A. 2020. Terminology in medical reports: Textual parameters and their lexical indicators that hinder patient understanding. *Terminology*, *26*(2), 213-236.

Gaudin, F. 2003. Socioterminologie: Une approche sociolinguistique de la terminologie. De Boeck-Duculot.

Gill, V. T. & Maynard, D. W. 2006. Explaining illness: Patients' proposals and physicians' responses. In D. W. Maynard & J. Heritage (Eds.), *Communication in Medical Care: Interaction between Primary Care Physicians and Patients* (p. 115-150). Cambridge University Press.

Humbley, J. 2018. Socioterminology. In J. Humbley, G. Budin, & C. Laurén (Eds.), *Languages for Special Purposes: An International Handbook* (p. 469-488). De Gruyter.

Ilschner, S., Neeman, T., Parker, M., & Phillips, C. 2022. Communicating Endometriosis Pain in France and Australia: An Interview Study. *Frontiers in Global Women's Health*, *3*.

Johnston, J. L., Reid, H., & Hunter, D. 2015. Diagnosing endometriosis in primary care: Clinical update. *British Journal of General Practice*, *65*(631), 101-102.

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* (p. 133-149). Rodopi.

León-Araúz, P. 2015. Term Variation in the Psychiatric Domain: Transparency and Multidimensionality. In P. Hacken & R. Panocová (Eds.), *Word Formation and Transparency in Medical English* (p. 33-54). Cambridge Scholars Publishing.

Péraldi, S. 2012. L'anglais de spécialité en chimie organique: Entre indétermination terminologique et multidimensionnalité. *ASp*, *62*, Art. 62.

Rouillard, C.-A. 2016. Usage spécialisé ou usage courant des désignations de la maladie mentale dans le discours des non-spécialistes: Le cas des termes autisme et hystérie. *ScriptUM: la revue du colloque VocUM*, *2*.

Sinclair, J. 1996. *Preliminary Recommendations on Corpus Typology*. EAGLES (Expert Advisory Group on Language Engineering Standards).

Young, K., Fisher, J., & Kirkman, M. 2015. Women's experiences of endometriosis: A systematic review and synthesis of qualitative research. *Journal of Family Planning and Reproductive Health Care*, *41*(3), 225-234.

Syntactic Complexity in Smartphone Application Contracts

Katherine Ireland & Tim Samples – *University of Georgia*

Keywords: *legal linguistics, syntactic complexity, corpus-based analysis, python programming.*

Smartphone application contracts govern user rights and user data worldwide. Although legal scholars and researchers have analyzed the readability of application contracts, this area of work remains understudied through linguistic approaches, most particularly in analyzing their grammatical and syntactic complexity (Becher & Benoliel 2019, 2022; Wagner 2022). Related corpus-based studies on syntactic complexity have generally focused specific genres, such as Biber and Gray's work on the noun phrase in written genres of English over time (2011) and grammatical and syntactic complexity in Academic English contexts (2016). Other studies have addressed implications for grammatical complexity in educational and language acquisition contexts (Larrson & Kaatari 2020; Ortega 2015; Kyle 2016; Kyle & Crossley 2018). We combine these areas and expand on previous research on the law and linguistics of mobile application contracts to focus on the syntactic complexity and linguistic patterns utilized in these contexts. The dataset analyzed in this work includes smartphone application contracts from different categories that mirror the average smartphone. These categories are social, finance and tech, games, education, and shopping applications with over 2 million tokens in total. Each application was selected due to its popularity and current usage on Android and Apple downloads. The corpus also includes multiple contract types: terms-of-use (TOUs) and privacy policies.

This study uses corpus-based methods in conjunction with the R and python programming languages to understand what grammatical constructions and linguistic patterns are present within these non-transparent, virtually unreadable contracts. We also utilize python with the Tool for the Automated Analysis of Syntactic Sophistication and Complexity (TAASSC, Kyle 2016; Kyle & Crossley 2018) to gather detailed information on embedding and syntactic structures. TAASSC crucially utilizes grammatical relationships within sentences to calculate syntactic complexity, counting the number of dependents per phrase type (Kyle 2016). TAASSC results are calculated for over thirty indices of clausal and phrasal types and constructions within the data, with composite scores for noun phrase and verb phrase complexity. We further compare this dataset with other notable genres of English writing, specifically using the Brown corpus (Francis and Kučera 1979) to understand the differences in grammatical constructions.

Preliminary results show several key findings. First, the highest use of complex noun phrase types occurs in Terms-of-Use contracts, with more dependents per noun phrase such as those occurring in the sentence: (*Your user contract for the purposes of our agreement must not violate the following purposes, and your violation of the rights of any third party and any violation by you of these Terms of Service will result in termination of your account*). Secondly, less complex noun phrase types occur in privacy policies. Additional findings across specific TOU categories show that gaming applications include the most prevalent use of complex noun phrase types. One specific noun phrase type highly used by gaming platforms is that of passive nominal subjects like (*Your gaming user account was terminated*). Finally, privacy policies are distinctive, with the greatest frequency of possessives overall.

The specific syntactic phrases and clauses used by these important contracts contribute to their lack of understanding by consumers. This has key legal implications for both application companies and for smartphone users across the globe. Further, this interdisciplinary work underscores the value that corpus-based methods bring to diverse contexts, like law and policy.

Bibliography

Biber, Douglas and Bethany Gray. 2011. Grammatical change in the noun phrase: the influence of written language use. *English Language and Linguistics* 15:2: 223-250.

Biber, Douglas and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing. Studies in English Language.* Cambridge University Press. Becher, Shmuel I. 2008. Asymmetric Information in Consumer Contracts: The Challenge That is Yet to Be Met. *American Business Law Journal* 45(4), 723-774.

Becher, Shmuel I. and Uri Benoliel. 2019. The Duty to Read the Unreadable.

Becher, Shmuel I. and Uri Benoliel. 2022. Dark Contracts. *Boston College Law Review 64* (forthcoming 2023).

Biber, Douglas and Bethany Gray. 2016. Grammatical Complexity in Academic English: Linguistic Change in Writing. Cambridge University Press.

Biber, Douglas and Bethany Gray. 2011

Kyle, Kristopher and Scott A. Crossley. 2018. Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices, *Journal of Modern Language.*

Kyle, Kristopher. 2016. Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. Ph.D. dissertation, Georgia State University, https://doi.org/10.57709/8501051.

Larsson, Tove and Henrik Kaatari. 2020. Syntactic complexity across registers: Investigating (in)formality in second-language writing. *Journal of English for Academic Purposes* 45. 1475- 1485.

Martinc, Matej, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Journal of Computational Linguistics.*

Martínez, Eric, Francis Mollica, and Edward Gibson. 2022. Poor Writing, Not Specialized Concepts, Drives Processing Difficulty in Legal Language. *Cognition.*

R Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Wagner, Isabel. 2022. Privacy Policies Across the Ages: Content and Readability of Privacy Policies 1996–2021, https://doi.org/10.48550/arXiv.2201.08739.

Wauters, Ellen, Eva Lievens, and Peggy Valcke. 2014. Towards a Better Protection of Social Media Users: A Legal Perspective on the Terms of Use of Social Networking Sites, 22 *Int'l J. L. Info. Tech.*

Wickham et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software, 4(43), 1686,* https://doi.org/10.21105/joss.01686.

Transforming identities in Chauvin's criminal trial: Prosecution and defence strategies from opening speech to closing argument

Natalie Jones – *University of Leeds*

Keywords: *forensic linguistics, positioning, courtroom discourse, lawyer talk, opening statement, closing argument.*

With a specific focus on the *State of Minnesota v. Derek Michael Chauvin* trial, this study uses the prosecution and defence's opening statements as a backdrop to the closing arguments to examine how the key social actors' identities are transformed. Critical discourse analysis (CDA), corpus linguistics (CL) using *AntConc* 3.5.9 (2020), and positioning theory (Davies and Harré, 1990) are used to investigate how Chauvin, Floyd, and the jury are positioned at the beginning and end of the trial. The quantitative results produced using CL methods were used to inform the CDA, working collectively to inform the findings.

This research concentrates on the opposing lawyers' strategic use of nomination/categorization and the surrounding collocations. According to van Leeuwen (2008), nomination/categorization is used to create unique identities, drawing on an individual's characteristics, group membership, occupation, and/or role. It is argued that when 'a speaker has many options as to what to call a person and chooses one systematically over the others […]', they are 'discursively creating' a specific identity and this 'shows what aspects of the person the speaker is highlighting in the discourse at that particular time' (Felton-Rosulek, 2009, p.9). For example, the prosecution initially refers to Chauvin as *Mr. Chauvin* in the opening but this shifts in the closing, with the predominant use of *the defendant* plus negative action. This was uncovered using the keyword and collocation function in *AntConc.* Using *the defendant* over *Mr. Chauvin* is a constant reminder to the jury of Chauvin's legal role in the trial, his alleged criminal actions and his stigmatised identity, while silencing his non-culpable respected identity outside of the trial's context. This subtle but dominant nuance is used to mold the jury's perception of the defendant in line with the prosecution and defence's desired crime narrative roles, through the discursive creation of their shifting identities. Similarly, this technique is used when responsibilizing the social actors in the closing arguments, for example, to assign blame, avoid responsibility, or reduce doubt about their culpability.

Additionally, the positioning of the jury within the courtroom is explored through the lawyers' strategic use of grammatical patterns. The N-gram clusters revealed that when referring to the jury in the opening statements, both the prosecution and defence repeatedly use the bigram *you will*, followed by the perception verbs *see* and *hear* or the cognitive verb *learn.* In comparison with N-grams used in the closing arguments, there is depletion in this bigram*,* as the most frequent bigram is *you can*, also followed by the perception verbs *see* and *hear.* Rather than the preemptive suggestion of what the jury *will see* and *hear,* the lawyers focus on what they *can see* and *hear.* The jury's identity is transformed from the position of observers in the opening to one of decision makers with knowledge and power.

While the transformation of the defendant's identity foregrounds his legal identity and position and suppresses his position in the wider world, the transformation of the jury's position is from one of possibility and prediction to one of ability and action as epistemic modality is replaced by dynamic, positioning them as powerful agents with clear perspectives for decision-making.

Bibliography

Davies, B. and Harré, R. 1990. Positioning: The Discursive Production of Selves. *Journal for the Theory of Social Behaviour.* 20(1), pp.43–63.

Felton Rosulek, L. 2009. The sociolinguistic creation of opposing representations of defendants and victims. *International Journal of Speech, Language & the Law.* 16(1), pp. 139–30.

Lawrence Anthony. 2020. *AntConc (3.5.9)* [Software]. [Accessed 25 June 2022].

van Leeuwen, T. 2008. Discourse and Practice: New Tools for Critical Discourse Analysis. Oxford University Press.

Wright, D. 2021. Positioning and responsibility in the opening statements of the Grenfell Tower inquiry: a corpus-assisted analysis In: *15th International Association of Forensic Linguistics conference,* Aston University. [Accessed 1 March 2022]. Available from: http://irep.ntu.ac.uk/id/eprint/44464/.

El *Corpus Diacrónico Andalusí del Árabe* (CORDANA): una puesta a prueba con el marcador de auto-referencia *nafs-ī* 'mi alma'

Laila M. Jreis-Navarro – *University of Zaragoza*

Palabras clave: *corpus; árabe clásico, al-Andalus, subjetividad.*

La construcción del Corpus Diacrónico Andalusí del Árabe (CORDANA) tiene como objetivo el estudio de la evolución de la expresión subjetiva premoderna que tuvo lugar en árabe en la Península Ibérica a través de una serie de marcadores (Jreis, 2022 y 2023), que definen la auto-expresión como auto-referencia, acción, emoción y evaluación. La lingüística de corpus ha sido pionera en esta materia, desde el trabajo seminal de Biber y Finegan (1989) sobre el punto de vista (stance), que fue decodificado en una serie de marcadores gramaticales y semánticos en lengua inglesa pertenecientes a las dos áreas de la evidencialidad y el afecto.

El CORDANA está conformado por una selección de 255 textos de 74 autores andalusíes (entre los siglos IX y XV) extraídos del macro corpus OpenITI de textos árabes premodernos (Versión 2022.1.6; Nigst *et al.*, 2022). El corpus ha sido procesado con AntConc y su tamaño es de 37.316.948 unidades lingüísticas (tokens). Los textos han sido limpiados de metadatos y anotación y clasificados en función de una serie de dominios (derecho, filosofía, historia, lengua, literatura, ciencia, religión, etc.) y géneros (*Adab* [miscelánea didáctica de bellas letras], Diccionario, Hadiz [tradición del profeta Muḥammad], *Maqāla* [tratado], *Mulaḫḫaṣ* [resumen], *Riḥla* [relato de viaje], *Risāla* [epístola], *Šarḥ* [explicación], *Šiʿr* [poesía], *Tafsīr* [exégesis], *Tariḫ* [crónica], *Tarǧama* [biografía], entre otros). En la determinación de los géneros se ha seguido un criterio externo (Bibers, 1988: 170), fundamentalmente cultural, basado en los títulos que los propios autores dieron a sus obras, en las referencias a estos por parte de los conformantes de su tradición o en su contenido.

Para poner a prueba la utilidad del corpus en el estudio de la auto-expresión andalusí, se ha tomado el marcador de auto-referencia *nafs* 'alma' con sufijo pronominal de primera persona del singular (*nafs-ī*). La importancia de este marcador reside en su representación de la complejidad del sujeto hablante cuyas creencias, emociones y juicios se hallan en debate a través de esta forma reflexiva (Lakoff, 1992). La búsqueda de las concordancias de este marcador para identificar sus patrones de uso y su evolución en al-Andalus ha hecho palpables las fortalezas y debilidades del CORDANA en este sentido, señalando vías de acción para su mejora. Las debilidades que han de abordarse son las siguientes: la abundancia de producción de autores orientales en las obras misceláneas, antologías y diccionarios bio-bibliográficos, que dificulta la circunscripción geográfica y la perspectiva diacrónica; la presencia reiterada de tradiciones del profeta (hadices) en las obras de derecho cuyo contenido dialogado contamina el análisis con una falsa primera persona; y el desequilibrio entre los distintos autores, en relación con la extensión y número de sus obras así como de sus épocas de producción.

Los primeros resultados que lanza AntConc en sus diversas herramientas no permiten identificar patrones claros; estos mejoran en un sub-corpus de textos del que se han excluido las obras dentro del dominio del derecho, o las pertenecientes a géneros como los diccionarios o las antologías, pero la necesidad de un mayor control del contenido del CORDANA se hace evidente, porque se pierden instancias valiosas del fenómeno. El análisis mejoraría con un nivel de anotación más profundo, identificando los prólogos de las obras, las biografías y la producción andalusí en los diccionarios, así como la voz autorial en los comentarios a obras ajenas.

Bibliografía

Biber, Douglas. 1988. *Variation across speech and writing*, Cambridge University Press.

Biber, Douglas y Finegan, Edward. 1989. «Styles of stance in English: Lexical and grammatical marking of evidentiality and affect», *Text*, 9:1, 93-124.

Jreis Navarro, Laila M. 2022. «La codificación lingüística de la subjetividad en la *Nufāḍat al-ǧirāb* de Ibn al-Ḫaṭīb: verbos, emociones y adjetivos evaluativos', *Al- Qanṭara*, 43:1. https://doi.org/10.3989/alqantara.2022.015.

Jreis Navarro, Laila M. 2023. «Ibn Khaldūn in his subjective lexicon. The emotional constellation of an intellectual in transition», *Miscelánea de Estudios Árabes y Hebráicos. Sección Árabe-Islam*, 72, 87-115.

Lakoff, George. 1992. «Multiple selves. The Metaphorical Models of the Self Inherent In Our Conceptual System», en *The Conceptual Self In Context*, *a Conference of the Mellon Colloquium on the Self at the Emory Cognition Project*. (Emory University) https://escholarship.org/uc/item/53g1n5b2 [consultada el 16/12/2022].

Nigst, Lorenz, Romanov, Maxim, Savant, Sarah Bowen, Seydi, Masoumeh y Verkinderen, Peter. 2022. *OpenITI: a Machine-Readable Corpus of Islamicate Texts* (2022.1.6) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6808108.

An overview of empirical and quantitative approaches to corpus translation studies

Hannu Kemppanen – *University of Eastern Finland*

Keywords: *corpus translation studies, translation universals, corpus-based critical translation studies, corpus-driven approach, corpus-assisted approach, corpus-based approach.*

This study provides a brief overview of empirical and quantitative approaches to corpus translation studies. It describes the development of this research field starting from the "dawn" of corpus translation studies during the years 1993–1996, when the idea of using corpus methods for descriptive translation research was introduced by Mona Baker (1993, 1995, 1996) and Sara Laviosa-Braithwaite (1996). The use of large electronic corpora played a relevant role in the further development of translation studies. The focus of the research shifted from the prescriptive study of equivalence between the source and target text to the analysis of translated texts in their host context (Kenny 2006: 43–44).

The description of the "dawn" of corpus translation studies is followed by an overview of research on translation universals during the early years of the millennium (e.g. Olohan 2000, Mauranen & Kujamäki 2004). Research focused initially on four translation universals: simplification, explicitation, the law of growing standardization and the law of interference (Laviosa et al. 2017). In addition to the basic four hypotheses on translation universals, Sonja Tirkkonen-Condit (2002, 2004) proposed *a unique items hypothesis* according to which translations tend to contain fewer unique items than comparable non-translated texts. Further, the paper introduces how the range of the research field expanded after the first years of the new millennium, and it provides a summary of new areas of corpus research, such as *corpus-based critical translation studies* (see Laviosa 2004).

The overview presents corpus-driven (statistical machine translation; exploratory corpus statistics), corpus-assisted (translation stylistics; parallel corpus comparison) and corpus-based approaches (translation universal features or general translation tendencies) (see Ji et al. 2017). In addition, the paper introduces standard corpus design practices of building corpora for different purposes, and it outlines how the basic corpus analysis tools — word-frequency lists, type-token ratio, keyword lists, collocations, clusters and concordancers — have been used in empirical translation studies.

The analysis of the historical development of corpus linguistic translation research shows that this field of study was ahead of its time in terms of using extensive digital materials in research. Many of the basic tools of traditional corpus translation studies are suitable for research in other disciplines as well. For example, keyword analysis provides a good opportunity to study the linguistic properties and content of extensive textual material. In particular, aboutness studies allow the use of corpus methods for linguistic analysis of any kind of texts, not just source or target texts. Analysing concordances and keyword lists provides good research methods for content and discourse analysis, which can be utilized in any field of science when studying, for example, the concepts used in the texts of the field under study. The current area of research, called *digital humanities*, continues the tradition of linguists and translation studies scholars by utilizing extensive text materials.

References

Baker, M. 1993. Corpus linguistics and translation studies. Implications and applications. In M. Baker, Francis, G. & Tognini-Bonelli, E. (Eds) *Text and Technology. In Honour of John Sinclair*. John Benjamins, 233–250.

Baker, M. 1995. Corpora in Translation Studies. An Overview and Some Suggestions for Future Research. *Target* 7(2), 223–243.

Baker, M. 1996. Corpus-based translation studies: The challenges that lie ahead. In Somers, H. (Ed.) *Terminology, LSP and translation. Studies in language engineering in Honour of Juan C. Sager*, (pp. 175–187). John Benjamins.

Ji, M., Hareide, L., Li, D. & Oakes, M. 2017. Corpus Methodologies Explained. An Empirical Approach to Translation Studies. Routledge.

Kenny, D. 2006. Corpus-based Translation Studies: A Quantitative or Qualitative Development? *Journal of Translation Studies* 9(1), 43–58.

Laviosa, S. 2004. Corpus-based translation studies: Where does it come from? Where is it going? *Language Matters* 35(1), 6–27.

Laviosa-Braithwaite, S. 1996. *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*. Department of Language Engineering, Volume I. A thesis submitted to the University of Manchester Institute of Science and technology for the degree of Doctor of Philosophy.

Laviosa, S., Pagano, A., Kemppanen, H. & Ji, M. 2017. *Textual and Contextual Analysis in Empirical Translation Studies*. Springer.

Mauranen, A. & Kujamäki, P. (Eds.) 2004. *Translation Universals – Do They Exist?* John Benjamins.

Olohan, M. (Ed.) 2000 Intercultural Faultlines. Research Models in Translation Studies 1. Textual and Cognitive Aspects. St. Jerome Publishing,

Tirkkonen-Condit, S. 2002. Translationese: A myth or an empirical fact? A study into the linguistic identifiability of translated language. *Target* 14(2), 207–220.

Tirkkonen-Condit, S. 2004. Unique items — over- or under-represented in translated language? In Mauranen, A. & Kujamäki, P. (Eds.), *Translation Universals: Do they exist?* (pp. 177–184). John Benjamins Publishing Company.

Construction and analysis of Tamazight (Berber) text corpus

Zayd Khayi – *University of Belgrade*

Keywords: *Tamazight (Berber) language, corpus linguistics, grammar rules, statistical methods.*

First of all, the grammatical structure of the Tamazight language remains poorly understood and lack of comparative grammar of Tamazight language leads to linguistic issues. In order to fill this gap even small, we have constructed the diachronic corpus of the Tamazight language and we have elaborated the program tool and this work is devoted to constructing that tool to analyze the different aspects of Tamazight or Berber language with its different dialects using in the north of Africa specifically in Morocco. At this corpus, I have worked only on three Moroccan dialects: Tamazight, Tarifiyt and Tachlhit. As you may know Tifinagh (Tuareg Berber language: ⵜⵉⴼⵉⵏⴰⵖ Berber pronunciation: [tifinaɣ] - *Wikipedia*) for Tamazight language, an official language in Morocco and Algeria. However, I deliberately worked on Latin version because of more sources it has. The corpus based on grammatical parameters and features of that language. The text collection contains more than 500 texts that cover long historic period. The corpus is free available and it will be useful for further investigations on Tamazight language. The texts were transformed into xml-format standardization goal. The corpus counts more than 200,000 words.

Based on the linguistic rules and statistic methods, original user interface and software prototype were developed by combining the technologies of web-design and object programming in Python.

In my presentation, I would like to present more details and features about how this corpus provides the users to distinguish between feminine/masculine nouns, verbs etc. The interface I use has three languages TMZ/FR/EN.

When selecting texts, we were guided by the idea that the best way to integrally represent the language is to collect the texts not only in different domains of knowledge but also written in different literary styles. Selected texts were not initially categorized. This work was made in a manual way. Within corpus linguistics, there is currently no commonly accepted approach to the classification of texts. We distinguish 10 categories of texts.

To describe and represent the texts in the corpus we elaborated the XML-structure according to the TEI recommendations. Using the search function may provide us with type of words we would search for like feminine/masculine nouns and verbs.

Nouns are divided into two parts. The gender in corpus has two forms. The neutral form of word corresponds to masculine while the feminine is indicated by a double *t- -t* affix (the prefix *t-* and the suffix *-t*). Ex: *Tarbat* ('girl'), *Tamtut* ('woman'), *Taxamt* ('tent') and *Tislit* ('bride'). However, there are some words whose feminine form contains only the prefix *t-* and the suffix *-a*. Ex: *Tasa* ('liver'), *tanja* ('family'), *tarwa* (progenitors).

Generally, Tamazight masculine words have such prefixes that distinguish them from other words. For instance, *a-*, *u-*, *i-*. Ex: *Asklu* ('tree') *udi* ('cheese'), *ighef* ('head').

As differentiated from the rule for feminine nouns, the rule for the defining the masculine nouns has a fair bit of exceptions. Verbs the in corpus are for the first person singular and plural that has suffixes *-agh*, *-ex*, *- egh*. Ex: *ghrex* ('I study'), *-fegh* ('I go out'), *nadagh* ('I call').

The program tool permits to obtain such characteristics of this corpus:

- list of all tokens;
- list of unique words;
- lexical diversity;

– realize different grammatical requests/

Currently, we are working on a method that enables to group automatically the words in grammatical classes like NOUN, VERB and ADJECTIVE using n-gram method.

References

Brenier-Estrine, C., Institut de recherches et d'études sur le monde arabe et musulman. 1994. *Bibliographie berbére annotée: 1992-1993*. Aix-en-Provence: Institut de recherches et d'études sur le monde arabe et musulman.

Sinclair, J. 2004. *Trust the text: Language, corpus and discourse*. 1-212. 10.4324/9780203594070.

An automatic syntax-based genre classification of Czech texts

Miroslav Kubát[1], Radek Čech[1], Ján Mačutek[2] & Michaela Nogolová[1]

University of Ostrava[1] *– Mathematical Institute, Slovak Academy of Sciences,* and *Constantine the Philosopher University*[2]

Keywords: *stylometry, syntax, genre, text classification, Czech.*

Corpus linguistics has undergone significant development both in terms of the number and size of built corpora as well as the quality of the automatic lemmatization and morphological annotation. However, an automatic annotation on the syntactic level is still quite challenging. A lack of large balanced corpora with syntactic annotation has made it difficult to conduct corpus-based syntactically oriented stylometric research on registers, styles, and genres. However, a number of syntactically annotated corpora have been recently released.

The study uses text material from the Czech National Corpus, namely the balanced corpus of contemporary written Czech SYN2020 (Křen et al. 2020). The corpus consists of three main text types (fiction, non-fiction, journalism) which are divided into subgroups (e.g. novel, short story, poetry, drama, administrative texts, humanities, social sciences, newspapers, leisure magazines, etc.). The total size is 100 million words and more than 120 million tokens (including punctuation). The corpus is lemmatized, morphologically and syntactically annotated (Jelínek et al. 2021). The syntactic annotation was performed using a parser from the NeuroNLP toolkit trained on the data from the Prague Dependency Treebank (Bejček et al 2012) and the fiction corpus FicTree (Jelínek 2017). The accuracy rates of SYN2020: UAS (unlabeled attachment score) = 92.39%, LAS (labeled attachment score) = 88.73%. Despite some errors, such an accuracy rate in a large balanced corpus is outstanding. We therefore consider the quality of the syntactic annotation sufficient for our research.

The aim of this research is to analyze different types of texts (styles and genres) from the viewpoint of frequencies of syntactic functions. More specifically, relative frequencies of syntactic functions are used to measure distances among groups of texts. The distances are calculated using the Cosine Delta method (Smith and Aldridge 2011) and a hierarchical cluster analysis is then performed. Using syntactic functions appeared to be efficient in authorship attribution in a small corpus in previous research (Čech et al. 2022; Soler-Company and Wanner 2016). We apply this approach to a genre analysis based on a large balanced corpus. We conduct two cluster analyses. The first focuses on more general style groups (txtype): NOV: novels; COL: short stories; VER: poetry; SCR: drama, screenplays; SCI: scientific literature; PRO: professional literature; POP: popular literature; MEM: memoirs, autobiographies; ADM: administrative; NEW: traditional journalistic texts; LEI: leisure magazines. The second approach provides a finer style analysis based on groups defined as genre in SYN2020 such as ECO: economy; POL: politics; PSY: psychology; NTW: nationwide newspapers; REG: regional newspapers; LIF: lifestyle; HOU: home, garden, hobbies; SPO: sports; etc.

Obtained results indicate that this approach can distinguish between different types of texts quite effectively. The use of syntactic functions seems to be an important characteristic for the classification of genres and styles.

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.

Čech, R., Mačutek, J., Kubát, M., Koščová, M. 2022. Does an author leave a syntactic footprint? In: Misuraca, M., Scepi, G., Spano, M. (eds.), *Proceedings of the 16th International Conference on Statistical Analysis of Textual Data*. Naples: Vadistat Press, pp. 221-228.

Jelínek, T. 2017: FicTree: a Manually Annotated Treebank of Czech Fiction. In: Hlaváčová, J. (ed.), *Proceedings of the 17th Conference on Information Technologies - Applications and Theory (ITAT 2017)*, pp. 181–185. http://ceur-ws.org/Vol-1885/181.pdf.

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. Cham: Springer, pp. 48–59.

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. Available at http://www.korpus.cz.

Smith P., Aldridge W. 2011. Improving authorship attribution: Optimizing Burrows' delta method. *Journal of Quantitative Linguistics*, 18(1), pp. 63-88.

Soler-Company, J. and Wanner, L. 2016. Authorship attribution using syntactic dependencies. In: Angulo, C., Godo, L. (eds.), *Artificial Intelligence Research and Development*, pp. 303-308. IOS Press.

Debunking perceptions in ERPP: a case study of research paper drafts

Natalia Judith Laso Martín & Elisabet Comelles – *University of Barcelona*

Keywords: *ERPP, academic writing, corpus-based resources and tools, EFL researchers' Perceptions.*

Research writing competence in English is essential for EFL communities, who experience increasing pressure to publish their research studies in peer-reviewed international journals. As already stated in the literature (Cargill & Burgess, 2008; Villagrán & Harris, 2009; Lillis & Curry, 2010; Cargill & O'Connor, 2013; Cotos, 2016; Burgess *et al.*, 2019), the purpose of the scientific community is to disseminate their results among practitioners in their field. Thus, it seems reasonable that in order to gain well-recognised access to a discourse community and, most importantly, acceptance within it, researchers must become familiar with the 'game strategies' (Etherington, 2008) involved in domain-specific research writing (Pérez-Llantada, 2014; Laso & John, 2017).

Hence, any EFL writer who aspires to get their manuscript accepted for publication needs to master the stylistic conventions characteristic of their research discipline (Cotos, 2016). To this respect, there is mounting evidence in ESP writing research (Nesselhauf, 2005; Ellis 2007; Hyland, 2008; Granger & Meunier, 2008; Paquot, 2008, and Laso & John, 2013, among others) that points to the fact that EFL specialized discourse communities already have a good command of domain-specific terminology. However, they find it tantalizingly challenging to acquire phraseological competence and become familiar with the most prototypical domain-specific collocational patterns in English (Boulton & Cobb, 2017; Chen & Flowerdew, 2018; Lee *et al.*, 2019; Li & Flowerdew, 2020; Charles & Hadley, 2022).

Based on the discussion above, we decided to explore the main challenges that English for Research Publication Purposes (ERPP) poses for a group of six EFL university lecturers and researchers, as well as train them in the use of some corpus-management tools, which may enhance their written production in English. Participants in the study were from a variety of domains, including psychology, earth sciences, economy, climatology, and anthropology. They were administered a pre-intervention questionnaire regarding their research writing process in English. Additionally, they were asked to submit an 800-word draft of a research paper (including the abstract) they were currently working on. All the collected data were used in the intervention stage, which consisted of a workshop on "English for Research Publication Purposes".

The aim of this study is twofold: a) identify what areas of research writing seem to be most problematic for participants in the study, and b) train participants in the use of corpus-based resources and tools to help EFL writers produce a phraseologically competent academic research article. Results from the questionnaire show that participants perceive word combinations and grammatical issues as the most challenging aspects in their writing process, followed by misuse of similar pairs of words and, to a much lesser extent, text-level issues regarding punctuation, linking words, and paragraph distribution. Interestingly, the analysis of their drafts revealed that text-level issues and word combinations pose the most salient difficulties in their research papers, whereas grammar and misuse of similar pairs of words are less problematic in their written production.

Based on these findings, a set of academic writing resources (i.e., the Academic Phrasebank and the Academic Collocation List) and corpus-based tools (i.e., AntConc 4.1.2 and Writefull) were presented and used during the workshop, which aimed at helping participants improve their writing as well as become independent users of such tools.

Bibliography

Laso, N. J. & John, S. 2013. An exploratory study of NNS medical writers' awareness of the collocational patterning of abstract nouns in medical discourse. *Revista Española de Lingüística Aplicada*, 307-331.

Laso, N.J. & John, S. 2017. The Pedagogical benefits of a lexical database (SciE-Lex) to assist the production of publishable biomedical texts by EAL writers. *Iberica*, 33, 147 - 172.

Boulton, A. & Cobb, T. 2017. Corpus use in language learning: A meta-analysis. *Language learning*, 67(2), 348-393.

Burgess, S., Martín, P. & Balasanyan, D. 2019. English or Spanish for Research Publication Purposes?: Reflections on a Critical Pragmatic Pedagogy. In Corcoran, J. N., Englander, K, & Muresan, L. (Eds.) *Pedagogies and Policies for Publishing Research in English. Local Initiatives Supporting International Scholars.* (pp. 128-140). New York: 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.

Cargill, M. & O'Connor, P. 2013. *Writing Scientific Research Articles*. Oxford: Wiley- Blackwell.

Charles, M. & Hadley, G. 2022. Autonomous corpus use by graduate students: A long-term trend study (2009–2017). *Journal of English for Academic Purposes*, 56, Article 101095.

Chen, M. & Flowerdew, J. 2018. A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics*, 23(3), 335-369.

Cotos, E. 2016. Computer-assisted research writing in the disciplines. In *Adaptive educational technologies for literacy instruction* (pp. 225-242). New York: Routledge.

Ellis, N. C. 2007. Learned attention in language acquisition: Blocking salience, and cue competition. Paper presented at the *EuroCogSci07, the Second European Cognitive Science Conference*, May 23-27, Delphi, Greece.

Etherington, S. 2008. Academic writing and the disciplines. In Friedrich, P. (Ed.) *Teaching Academic Writing.* (pp. 26-58). London: Continuum.

Granger, S. & Meunier, F. (Eds.) 2008. *Phraseology in foreign language learning and teaching.* Amsterdam/Philadelphia: John Benjamins.

Hyland, K. 2008. As can be seen: lexical bundles and disciplinary variation" *English for Specific Purposes*, 27, 4-21. doi:10.1016/j.esp.2007.06.001

Lee, H., Warschauer, M. & Lee, J. H. 2019. The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, 40(5), 721-753.

Li, Y. & Flowerdew, J. 2020. Teaching English for Research Publication Purposes (ERPP): A review of language teachers' pedagogical initiatives. *English for Specific Purposes*, 59, 29-41.

Lillis, T. & Curry, M. L. 2010. Academic writing in a global context: The politics and practices of publishing in English. London and New York: Routledge.

Nesselhauf, N. 2005. *Collocations in a learner corpus*. Amsterdam/Philadelphia: John Benjamins.

Paquot, M. (2008). Exemplification in learner writing: A cross-linguistic perspective. In Granger, S. & Meunier, F. (Eds.) *Phraseology in Foreign Language Learning and Teaching.* (pp. 100-120). Amsterdam/Philadelphia: John Benjamins.

Pérez-Llantada, C. 2014. Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, 14, 84-94.

Villagran, T., Harris, D., & Paul, R. 2009. Some key factors in medical writing. *Revista Chilena de Pediatría-Chile*, *80*(1), 70-78.

Lexical diversity of nouns in a learner corpus of Spanish EFL learners' B1 and C1 email writing. How does it correlate with the noun database recorded in the *English Vocabulary Profile* (EVP)?

Natalia Judith Laso – *University of Barcelona*

Keywords: *lexical diversity, learner email writing, EVP, CEFR B1, CEFR C1.*

Much of the past and current literature on the noun phrase in EFL writing focuses on linguistic complexity and accuracy (Ortega, 2003; Biber *et al.*, 2011; Lu, 2011; Ai & Lu, 2013; Bulté & Housen, 2014; Crossley & McNamara, 2014; Parkinson & Musgrave, 2014; Liu & Li, 2016; Xu, 2019; Díez-Bedmar & Pérez-Paredes, 2020; Kim, 2021), as well as lexical richness and phraseological competence (Howarth, 1998; Biber & Conrad, 1999; Nation, 2001; Hyland, 2008; Šišková, 2012; Peters, 2016; Vedder & Benigno, 2016; Paquot, 2019; Du *et al.*, 2022). Additionally, the linguistic characteristics of EFL writing have been widely discussed in relation to learners' L1, topic and genre effects, task complexity, and learners' L2 level of proficiency (Ellis & Yuan, 2004; Ong & Zhang, 2010; Díez-Bedmar, 2015; Mazgutova & Kormos, 2015; Liu & Li, 2016; Yoon, 2017, among others). However, little is known about the lexical diversity of nouns through the lens of CEFR proficiency levels in (Spanish) EFL written production.

The purpose of this research is twofold. First, to describe the lexical diversity of nouns both in the B1 and C1 learner subcorpora of email production. Second, to investigate if the nouns found in our B1 and C1 learner corpora are aligned with the nouns recorded in the English Vocabulary Profile (EVP), which provides the lexical items used by learners at each CEFR level.

This study examines a total of 680 noun lexemes extracted from a subcorpus of B1 (44 texts) and C1 (46 texts) emails taken from the XXX corpus, which contains the written production by Spanish learners who have passed the high-stakes XXX exam in University Language Centres in Spain. The sample consisted of 90 emails, which were POS tagged using Freeling (Padró *et al.*, 2010; Padró & Stanilovsky, 2012). All noun lexemes were first disambiguated by means of the UKB option (Agirre *et al.*, 2018) in Freeling and were later annotated using WordNet (Fellbaum, 1998). A total of three (direct and inherited) hypernyms were retrieved and then manually annotated, according to the corresponding sense in each given context. Each noun lexeme was then classified as a hyponym of the selected superordinate terms and two databases of semantic fields (i.e., B1 and B2, respectively) were created.

The results confirm the expectation that there is a lower range of nouns in the B1 sample in comparison to the C1 learners' sample. However, these results need to be interpreted with caution: the use of nouns from different semantic fields tends to be generally influenced by the topic and text type of given tasks. As evidenced by Du *et al.* (2022:8), the lexical choices made by language learners may be influenced by several variables, such as text types, genres, and registers (Biber & Conrad, 1999; Hyland, 2008). So may EFL learners' written production. On the contrary, the comparison of the noun databases explored reveals unexpected results. In our sample, both at B1 and C1 levels, the percentage of nouns that, following the information in the EVP, belong to the target level is significantly low.

This corpus-based study has several implications for teachers, materials writers, and test developers: a) our results highlight the need to align the lexicon of nouns used by Spanish B1 and C1 EFL learners to the EVP, and b) this research provides an insight into the most prototypical nouns, classified into semantic fields, at different CEFR proficiency levels, which may inform the writing analytical scales of high-stakes exams.

Bibliography

Agirre, E., López de Lacalle, O., & Soroa, A. 2018. The risk of sub-optimal use of Open Source NLP Software: UKB is inadvertently state-of-the-art in knowledge-based WSD. NLP-OSS workshop at ACL (arXiv:1805.04277).

Ai, H. & Lu, X. 2013. A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In Díaz-Negrillo, A., Ballier, N., & Thompson, P. (Eds.) *Automatic treatment and analysis of learner corpus data* (pp. 249–264). Amsterdam: John Benjamins Publishing Company.

Biber, D. & Conrad, S. 1999. Lexical bundles in conversation and academic prose. *Language and Computers*, 26, 181–190.

Biber, D., Gray, B., & Poonpon, K. 2011. Characteristics of conversation to measure complexity in L2 writing development. *TESOL Quarterly*, 45(1), 5–35.

Bulté, B. & Housen, A. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26, 42–65.

Crossley, S. A. & McNamara, D. S. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26, 66–79.

Díez-Bedmar, M. B. 2015. Article use and criterial features in Spanish EFL writing. *Learner corpora in language testing and assessment*, 70, 163-190.

Díez-Bedmar, M. B. & Pérez-Paredes, P. 2020. Noun phrase complexity in young Spanish EFL learners' writing. Complementing syntactic complexity indices with corpus-driven analyses. *International Journal of Corpus Linguistics*, 25/1, 4-35.

Du, X., Afzaal, M. & Al Fadda, H. 2022. Collocation Use in EFL Learners' Writing Across Multiple Language Proficiencies: A Corpus-Driven Study. *Frontiers in Psychology*, 13:752134.

Ellis, R. & Yuan, F. 2004. The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, 26, pp. 59- 84.

Fellbaum, Ch. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Howarth, P. 1998. The phraseology of learners' academic writing. In Cowie, A. P. (Ed.) *Phraseology: Theory, analysis, and applications* (pp.161–186). Oxford: Oxford University Press.

Hyland, K. 2008. As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27, 4–21.

Kim, J. 2021. Measuring NP complexity in Korean EFL writing across CEFR levels A2, B1 and B2. *Korean Journal of English Language and Linguistics*, 21, 341-358.

Liu, L. & Li, L. 2016. Oun Phrase Complexity in EFL Academic Writing: A Corpus-Based Study of Postgraduate Academic Writing. *The Journal of Asia TEFL*, 13(1), 48-65.

Lu, X. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1), 36-62.

Mazgutova, D. & Kormos, J. 2015. Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, 29, 3- 15.

Nation, I.S.P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Ong, J. & Zhang, L. J. 2010. Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing*, 19(4), 218-233.

Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 Proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 4(4), 492–518.

Padró, Ll., Reese, S., Agirre, E., & Soroa, A. 2010. Semantic Services in FreeLing 2.1: WordNet and UKB. In Bhattacharyya, P., Fellbaum, C., & Vossen, P. (Eds.) *Principles, Construction, and Application of Multilingual Wordnets* (99-105). Mumbai, India: Narosa Publishing House.

Padró, Ll. & Stanilovsky, E. 2012. FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*. Istanbul, Turkey.

Paquot, M. 2019. The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1), 121–145.

Parkinson, J. & Musgrave, J. 2014. Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes*, 14, 48–59.

Peters, E. 2016. The learning burden of collocations: the role of interlexical and intralexical factors. *Language Teaching Research*, 20, 113–138.

Šišková, Z. 2012. Lexical Richness in EFL Students' Narratives. *Language Studies Working Papers*, 4, 26-36.

Vedder, I. & Benigno, V. 2016. Lexical richness and collocational competence in second- language writing. *International Review of Applied Linguistics in Language Teaching*, 54(1), 23-42.

Xu, L. 2019. Noun phrase complexity in integrated writing produced by advanced Chinese EFL learners. *Papers in Language Testing and Assessment*, 8(1), 31-51.

Yoon, H.-J. 2017. Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality, *System*, 66, 130-141.

El léxico del arte en textos museísticos: aproximaciones a partir de un corpus paralelo y alineado

Jorge Leiva-Rojo – *University of Málaga*

Palabras clave: *traducción inglés-español, estudios basados en corpus, textos museísticos.*

Las visitas a los museos y centros de arte son una actividad cultural de ocio y cultura de primer orden. Así lo señalan Falk y Dierking, quienes además estiman que «more than a billion people, young and old, alone or in groups, visit a museum of some kind every year» (2016, p. 23). Si se tiene en cuenta que la mayoría de las visitas a los museos son por parte de personas que viven fuera de las ciudades donde radica el museo en cuestión (Brida et al., 2016, p. 216), es fácil concluir que las necesidades de traducción en estos espacios serán más que ocasionales. La propuesta que se presenta para el *XIV Congreso Internacional de Lingüística de Corpus* tiene como objetivo profundizar en un ámbito poco estudiado hasta la fecha, a pesar de su relevancia (Liao, 2018, p. 46), y analizar la forma en que se traducen al español algunos de los términos relacionados con el arte que aparecen con mayor frecuencia en textos en inglés y español procedentes de museos de los Estados Unidos. Se pretende con esta aportación, en primer lugar, describir las principales características que presenta el corpus que se empleará para este trabajo. En segundo lugar, mediante un análisis de los textos de que se compone el corpus —y de los elementos que lo integran, tales como palabras más frecuentes, palabras clave y unidades fraseológicas, entre otros—, se tiene la intención de comprobar cómo es el español que aparece en las traducciones de textos de museos de los Estados Unidos hacia esa lengua. Para ello, se recurrirá en gran medida al contraste con un corpus comparable de textos museísticos escritos originariamente en lengua española.

En lo relativo al marco teórico en que se inscribe este trabajo, es preciso indicar que en los últimos años son relativamente numerosas las contribuciones que se han centrado en el texto museístico traducido como objeto de estudio. Sirvan como ejemplo de ello los trabajos de Liao (2018) y Neather (2018), o los de Guillot (2014) y Valdeón (2015). Estos dos últimos, junto con algunos trabajos recientes del proponente de la comunicación, incorporan el estudio basado en corpus para la realización de la investigación, lo que permite llegar a resultados más sistemáticos y exhaustivos, en la línea de lo promulgado por Neather (2012: 207).

En lo que respecta a los resultados que se espera lograr de este trabajo, es preciso indicar que se partirá de la base de un corpus compilado en 2016 y sucesivamente ampliado hasta su versión actual, finalizada en el año 2021. El corpus, que contiene textos museísticos escritos originariamente en inglés y traducidos a la lengua española, cuenta con un total de 832 bitextos, procedentes de 65 museos distintos de toda la geografía estadounidense; por último, entre los dos subcorpus suman más de 2,7 millones de palabras.

Referencias bibliográficas

Brida, J. G., Dalle Nogare, C., y Scuderi, R. 2016. Frequency of museum attendance: motivation matters. *Journal of Cultural Economics*, *40*(3), 261–283. https://doi.org/10.1007/s10824-015-9254-5.

Falk, J. H. & Dierking, L. D. 2016. *The Museum Experience Revisited.* Routledge. https://doi.org/10.4324/-9781315417851.

Guillot, M.-N. 2014. Cross-cultural pragmatics and translation: the case of museum texts as interlingual representation. En J. House (Ed.), *Translation: A Multidisciplinary Approach* (pp. 73–95). Palgrave Macmillan. https://doi.org/10.1057/9781137025487_5.

Liao, M.-H. 2018. Museums and creative industries: The contribution of Translation Studies. *The Journal of Specialised Translation*, *29*, 45–62.

Neather, R. 2012. Intertextuality, translation, and the semiotics of museum presentation: The case of bilingual texts in Chinese museums. *Semiotica*, *192*, 197–218. https://doi.org/10.1515/sem-2012-0082.

Neather, R. 2018. Museums, material culture, and cultural representation. En S.-A. Harding y O. Carbonell Cortés (Eds.), *The Routledge Handbook of Translation and Culture* (pp. 361–378). Routledge. https://doi.org/-10.4324/9781315670898.

Valdeón, R. A. 2015. Colonial museums in the US (un)translated. *Language and Intercultural Communication*, *15*(3), 362–375. https://doi.org/10.1080/14708477.2015.1015351.

Enlarging the inventory of evidential expressions in English: A look from COHA and COCA

María José López Couso & Belén Méndez Naya – *University of Santiago de Compostela*

Keywords: *evidentiality, parenthetical, grammaticalization, American English, COHA, COCA.*

In contrast to other languages (e.g. Amerindian languages; see Mithun 1986; Babel 2009), English has never expressed evidentiality, i.e. the category encoding the speaker's information source for his/her statement, by means of dedicated morphological devices. However, over history it has made use of various evidential strategies (Aikhenvald 2004), including modal auxiliaries (e.g. Old English hearsay *sceolde*), adverbs (e.g. *evidently*), and parentheticals (e.g. *it seems*) (see, among others, Chafe 1986: 261).

In addition to these more classic ways of conveying evidentiality, this paper examines an evidential parenthetical type which seems to have developed in the course of the twentieth century and which (to our knowledge) has gone unnoticed in the extensive literature on evidentials and on parentheticals. Examples (1) and (2) illustrate this pattern:

(1) Several factors could abort a real-estate turnaround before it even begins**, experts warn.** (COCA, 1993, NEWS)

(2) Four to eight people grieve intensely for each suicide**, studies show**. (COCA, 2013, NEWS)

As seen here, the evidential parenthetical pattern under analysis shows a third person subject featuring either an animate noun such as *experts* or an inanimate noun of the type *study*, *evidence*, etc., and a predicate of utterance (e.g. *say*, *warn*), demonstration (e.g. *show*, *demonstrate*), knowledge or acquisition of knowledge (e.g. *find*, *conclude*), and the like. As illustrated in examples (3) and (4), the parentheticals at issue here have a main clause counterpart as matrices in complementation structures, from which they are taken to derive historically, in accordance with Thompson & Mulac's (1991) matrix clause hypothesis. The parentheticals in (1) and (2) differ from the complementation constructions in (3) and (4) in showing a reversal in terms of syntactic (main vs. subordinate) and discourse prominence (primary vs. secondary information) (Boye & Harder 2007).

(3) **Experts warn that** the debt-paying process differs with every financial situation (COCA, 2012, WEB)

(4) **Studies show that** 90% of people with hypothyroidism are producing antibodies to thyroid tissue (COCA, BLOG, 2012)

Focusing on parentheticals showing the nouns *study* and *expert* as heads of the subject noun phrase, in our talk we address the following questions: (i) when did this parenthetical type emerge?; (ii) is this parenthetical type favoured in particular text-types?; (iii) which are the most common predicates occurring in the pattern under analysis?; (iv) which is the preferred position (initial, medial, final) occupied by the parenthetical in the sentence?; (v) does the parenthetical show an increase in morphosyntactic fixation (loss of variability in the subject NP; TAM restrictions in the VP) over time?; and (vi) does subject animacy play a role in the development and distribution of the parenthetical?

Evidence is drawn from two multigenre mega-corpora of American English: the *Corpus of Historical American English* (COHA; Davies 2010) and the *Corpus of Contemporary American English* (COCA; Davies 2008).

Bibliography

Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford: Oxford University Press.

Babel, Anna M. 2009. *Dizque*, evidentiality, and stance in Valley Spanish. *Language in Society* 38: 487-511.

Boye, Kasper & Peter Harder. 2007. Complement-taking predicates. Usage and linguistic structure. *Studies in Language* 31(3): 569-606.

Chafe, Wallace. 1986. Evidentiality in English conversation and academic writing. In *Evidentiality: The Linguistic Coding of Epistemology*, Wallace Chafe & Johanna Nichols (eds.), 261-272. Norwood, NY: Ablex.

Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/.

Davies, Mark. 2010. *The Corpus of Historical American English (COHA)*. Available online at https://www.english-corpora.org/coha/.

Mithun, Marianne. 1986. Evidential diachrony in Northern Iroquoian. In *Evidentiality: The Linguistic Coding of Epistemology*, Wallace Chafe & Johanna Nichols (eds.), 89-112. Norwood, New Jersey: Ablex.

Thompson, Sandra & Anthony Mulac. 1991. A quantitative perspective on the grammaticization of epistemic parentheticals in English. In Elizabeth C. Traugott & Bernd Heine (eds.) *Approaches to Grammaticalization*. Vol. II. Amsterdam: John Benjamins, 313-339.

Past participle forms in competition: -(*e*)*d* vs -(*e*)*n* in historical British and American English

Juan Lorente-Sánchez – *University of Málaga*

Compared with other stages of the history of the language, the verbal morphology of Late Modern English (henceforth LModE) is characterised by considerable levels of variation in terms of the use and distribution of participial forms (Mondorf 2012: 843), especially of those verbs falling into the category of weak in Old English (OE). Ascribed within class 2 of OE weak verbs, some of these adopted the strong past participle ending -(*e*)*n* in Middle English (ME) alongside the original weak -(*e*)*d*, leading to an alternation between the two forms that has continued in the following centuries until the present day (e.g. 'show' > OE *scēawod;* ME *sheu(e)de*, *shauen*; ModE *showed, shown*) (cf. Lass 1987: 175, 1992: 127; Welna 2012: 425).

Even though -(*e*)*n* is today the predominant option to the detriment of -(*e*)*d* in both British English (BrE) and American English (AmE), these participial suffixes present a variable historical distribution, particularly in the seventeenth century and the whole LModE period. This regarded, the present contribution seeks to address the -(*e*)*d*/-(*e*)*n* alternation in BrE and AmE in 1600-1900 and, to the purpose, it considers six verbs belonging to the category of OE weak class 2 which later developed the strong -(*e*)*n* variant, namely *awake, saw, sew, show, strew* and *wake* (cf. Welna 2012). The paper is therefore conceived with a twofold objective: 1) to analyse the distribution of these alternatives in the period under scrutiny; and 2) to shed some light on their use in the two varieties over time, determining its contrasts and similarities.

The evidence stem from a number of sizeable corpora representative of early British and American English. As far as BrE is concerned, the data come from three different sources, including the seventeenth-century component of the *Early English Books Online* corpus (*EEBO*) (Davies 2017), the *Eighteenth Century Collections Online* (*EECO*) and the *Corpus of Late Modern English Texts* (*CLMET*) (De Smet et al. 2015). When it comes to AmE, the analysis is based on the information supplied by the *Evans Early American Imprints* collection (*EVANS*) as well as the nineteenth-century component of the *Corpus of Historical American English* (*COHA*) (Davies 2010). All together amount to up to 979,889,648 words from a wide variety of documents, thus making them the appropriate input for both a synchronic and a diachronic study of these two participial forms in competition in historical BrE and AmE.

While preliminary results denote a somewhat variable diffusion of the variants depending on the specific verb at hand, they show a preference for the –(*e*)*d* participial ending in the seventeenth and eighteenth centuries in the two varieties of English considered. In addition to this, the data also reveal an inversion of the trend in nineteenth- century BrE, whereas in AmE the picture remains unaltered, which leads us to conclude –at least tentatively– that in the second of these varieties the change to the current state- of-affairs took place later in time.

References

Davies, Mark. 2010. *The Corpus of Historical American English (COHA).* Available online at https://www.english-corpora.org/coha/.

Davies, Mark. 2017. *Early English Books Online Corpus.* Available online at https://www.english-corpora.org/eebo/.

De Smet, Hendrik, Susanne Flach, Jukka Tyrkkö and Hans-Jürgen Diller. 2015. *The Corpus of Late Modern English (CLMET), Version 3.1: Improved Tokenization and Linguistic Annotation.* Katholieke Universiteit Leuven, Freie Universität Berlin, University of Tampere, Ruhr-Universität Bochum. Available from https://fedora.clarin-d.uni-saarland.de/clmet/clmet.html.

Eighteenth Century Collections Online (ECCO). Available online at https://quod.lib.umich.edu/e/ecco/.

Evans Early American Imprint Collection (EVANS). Available online at https://quod.lib.umich.edu/e/evans/.

Lass, Roger. 1987. *The Shape of English: Structure and History*. London: J. M. Dent & Sons Ltd.

Lass, Roger. 1992. "Phonology and Morphology". In Norman Blake (ed.), *The Cambridge History of the English Language. Volume II: 1066-1476*. Cambridge: Cambridge University Press. 23-155.

Mondorf, Britta. 2012. "Late Modern English: Morphology". In Alexander Bergs and Laurel J. Brinton (eds.), *English Historical Linguistics: An International Handbook. Volume 1*. Berlin and Boston: Mouton de Gruyter. 842-869.

Welna, Jerzy. 2012. "Middle English: Morphology". In Alexander Bergs and Laurel J. Brinton (eds.), *English Historical Linguistics: An International Handbook. Volume 1*. Berlin and Boston: Mouton de Gruyter. 415-434.

Inteligencia artificial para el estudio de la variación de género textual en corpus históricos de español en contacto

Thomas Louf, Ruth Miguel-Franco, Bàrbara Montoya-Boix & David Sánchez

University of the Balearic Islands

El objetivo de esta comunicación es aplicar metodologías de inteligencia artificial, en particular, algoritmos de clasificación, para estudiar patrones de variación lingüística en documentación notarial y epístolas privadas producidas en una situación de contacto de lenguas en la Mallorca del siglo XVIII. Para ello, se utilizarán, por una parte, documentos notariales y públicos del Corpus Mallorca (www.corpusmallorca.es) y, por otra, epístolas privadas pertenecientes al epistolario de Pedro de Santacilia. Todos los textos fueron redactados por mallorquines entre 1700 y 1799, esto es, tienen la misma procedencia geográfica pero pertenecen a diferentes tradiciones discursivas. Asimismo, se utilizará un corpus de control de documentación y otro de epístolas redactadas en español de zonas monolingües de la Península Ibérica, con la misma cronología y las mismas características discursivas. Los resultados de los algoritmos empleados permiten agrupar los documentos según su origen geográfico y, además, nos proporcionan una serie de rasgos que caracterizan el sistema gráfico, el vocabulario y la ortografía léxica de cada corpus, lo que posibilitará, entre otras cosas, detectar las diferencias en la aparición de rasgos de contacto en los diferentes géneros.

En resumen, los resultados del estudio dan fe de la importancia de la aproximación cuantitativa en el estudio de corpus, sobre todo desde el punto de vista de la sociolingüística histórica. Las herramientas de inteligencia artificial nos permiten no solo analizar la variación lingüística, sino además, mediante un estudio objetivo, asociar diferentes rasgos con una u otra tradición discursiva.

Democratization, colloquialization and informalization in NYT editorials (1860–1979)

Lucía Loureiro-Porto[1] & Elena Seoane[2]

University of the Balearic Islands[1] – University of Vigo[2]

Keywords: *democratization, colloquialization, informalization, (semi)modals, contractions, passives.*

Across an extensive literature it has been shown that socio-historical events and trends can have a notable impact on language use (Fairclough 2003: 3). These include, among others, democratization (Fairclough 1992: 201-207), colloquialization (Mair 1997: 203-205), and popularization (Biber and Gray 2012), the latter also known as informalization (Farrelly and Seoane 2012: 393-396; see also Leech 2004: 75; Mair 2006: 183-193; Mair and Leech 2006: 336–337; Leech et al 2009: 259 Hiltunen and Loureiro-Porto 2020b; Loureiro-Porto 2020, 2021; Loureiro-Porto and Hiltunen 2020b). Previous research on these three trends (see also Biber and Gray 2012; Hiltunen and Loureiro-Porto 2020a; Loureiro-Porto and Hiltunen 2020a) has concentrated largely on their linguistic expression, identifying the linguistic features associated with them and describing their frequency in different registers. This paper focuses on the role of the three trends as discourse- pragmatic processes entailing the "reflection, through language, of changing norms in personal relations" (Leech et al. 2009: 259). It therefore adopts a pragmatic approach to explore how the pragmatic negotiation of power relations evolves as a consequence of (i) democratization, the avoidance of overt markers of power asymmetry with the aim of expressing more equal and solidary power relations, (ii) colloquialization, which introduces speech-like features to reproduce a communicative setting that is more relaxed and spontaneous and thus reduces the social distance between interlocutors, and (iii) informalization, whereby traditionally formal texts written by experts (scientific and newspaper discourse, for example) use linguistic devices that make them less distant, elaborate and formal, and thus closer to the reader (Mair 2006; Seoane 2006; Leech et al. 2009; Farrelly and Seoane 2012: 393). For this purpose, we analyze diachronic corpus data from Davies' (2010) *Corpus of Historical American English* (COHA), which hence situates the study at the intersection of three fields: pragmatics, historical linguistics and corpus linguistics. In other words, we wish to provide an "application of corpus linguistic methods to research questions in pragmatics applied to historical data" (Jucker and Taavitsainen 2014: 3). Through vertical and horizontal analyses, our data allow us to answer two broad research questions: (1) what is the quantitative evolution of (a) modal vs semi-modal verbs, (a) full vs contracted forms and (c) passives?; and (2) what do these pragmatic variables indicate about the evolution of power relations? Results show (i) a pronounced decrease of deontic modal *must* in favor of deontic semi- modals NEED TO, HAVE TO and HAVE GOT TO (democratization), (ii) a slight decrease in the number of full forms in favor of contracted forms (colloquialization), and (iii) a pronounced decrease of passives (informalization). The three trends intersect in editorials, with a predominance of democratization and informalization and a weaker role of colloquialization. This difference is interpreted as a sign that the three trends, though interrelated, are distinct and their effects are register-dependent. As for the second research question, we find that power relations in editorials are shown to be modelled by the long-term shift towards more democratic relations together (with specific socio-historical events which may temporarily halt the effects of democratization), just like we observe an increasing tabloidization (which reinforces the tendency towards informalization) and a reduction of the distance between writers and readers, contributing to a more symmetrical power distribution.

References

Biber, Douglas, and Bethany Gray. 2012. "The Competing Demands of Popularization vs. Economy: Written Language in the Age of Mass Literacy." In *The Oxford Handbook of the History of English*, ed. by Terttu Nevalainen, and Elizabeth Closs Traugott, 314-328. New York: Oxford University Press.

Davies, Mark. 2010. *The Corpus of Historical American English* (COHA). Available online at https://www.english-corpora.org/coha/.

Fairclough, Norman. 1992. *Discourse and Social Change.* Cambridge: Polity.

Fairclough, Norman. 2003. Analysing Discourse: Textual Analysis for Social Research. London: Routledge.

Farrelly, Michael, and Elena Seoane. 2012. "Democratization." In *The Oxford Handbook of the History of English*, ed. by Terttu Nevalainen, and Elizabeth Closs Traugott, 392-401. New York: Oxford University Press.

Hiltunen, Turo, and Lucía Loureiro-Porto. 2020a. "Democratization of Englishes: Synchronic and Diachronic Approaches." *Language Sciences* 79, May 2020, 101275.

Hiltunen, Turo, and Lucía Loureiro-Porto (eds). 2020b. *New Perspectives on Democratization: Evidence from English(es).* Special issue of *Language Sciences* 79, May 2020, 101275.

Jucker, Andreas H., and Irma Taavitsainen. 2014. "Diachronic Corpus Pragmatics: Intersections and Interactions." In *Diachronic Corpus Pragmatics*, ed. by Irma Taavitsainen, Andreas H. Jucker, and Jukka Tuominen, 3-26. Amsterdam: John Benjamins.

Leech, Geoffrey, Marianne Hundt, Christian Mair, and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study.* Cambridge: Cambridge University Press.

Leech, Geoffrey. 2004. "Recent Grammatical Change in English: Data, Description, Theory." In Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23), Göteborg 22–26 May 2002, ed. by Karin Aijmer, and Bengt Altenberg, 61–81. Amsterdam: Rodopi.

Loureiro-Porto, Lucía. 2020. "Singular THEY in Asian Englishes: A Case of Linguistic Democratization?" In *Crossing Linguistic Boundaries: Systemic, Synchronic and Diachronic Variation in English*, ed. by Paloma Núñez-Pertejo, María José López-Couso, Belén Méndez- Naya, and Javier Pérez-Guerra, 187-209. London: Bloomsbury Publishing.

Loureiro-Porto, Lucía. 2021. "Linguistic Colloquialisation, Democratisation and Gender in Asian Englishes." In *Gender in World Englishes*, ed. by Tobias Bernaisch, 176-204. Cambridge: Cambridge University Press.

Loureiro-Porto, Lucía, and Turo Hiltunen. 2020a. "Democratization and Gender-Neutrality in English(es)." *Journal of English Linguistics* 48 (3): 215-232

Loureiro-Porto, Lucía, and Turo Hiltunen (eds). 2020b. Democratization and Gender-neutrality in English(es). Special issue of Journal of English Linguistics 48(3).

Mair, Christian, and Geoffrey Leech. 2006. "Current Change in English Syntax". In *The Handbook of English Linguistics*, ed. by Bas Aarts, and April McMahon, 318–342. Oxford: Blackwell.

Mair, Christian. 1997. "Parallel Corpora: A Real-Time Approach to the Study of Language Change in Progress." In *Corpus-Based Studies in English*, ed. by Magnus Ljung, 195-209. Amsterdam: Rodopi.

Mair, Christian. 2006. Twentieth-Century English: History, Variation, and Standardization. Cambridge: Cambridge University Press.

Seoane, Elena. 2006. "Changing Styles: On the Recent Evolution of Scientific British and American English." In *Syntax, Style and Grammatical Norms: English from 1500-2000,* ed. by Christiane Dalton-Puffer, Dieter Kastovsky, and Nikolaus Ritt, 191-211. Bern: Peter Lang.

"There is life beyond *however*": Adverbs of contrast in a learner-based corpus of L1 Spanish EFL writings

Carmen Maíz-Arévalo – *Complutense University of Madrid*

Keywords: *L2 writing, email, adverbs of contrast, English as a Foreign Language.*

The analysis of written compositions by learners of English as a foreign or second language (EFL/ESL) has been a subject of scholarly interest for decades, with a special focus on the use of adverbs (e.g. Chen, 2012; Gilquin and Paquot, 2008; Neff, 2008; Pérez-Paredes and Díez-Bedmar, 2019; Suzuki et al., 2012; Yilmaz and Dikilitas, 2017, among many others). In the case of L1 Spanish EFL learners, the adverbial expression of contrast has proven to be particularly difficult for EFL learners, who often perform a negative transfer from their Spanish L1 (Larsson et al., 2020; Mora Díaz and Gómez Orjuela, 2021; Pérez-Paredes and Sánchez-Tornel, 2014), adverbs of contrast like *however* have proven to be particularly difficult to master for Spanish EFL learners, who tend to overuse it to express contrast. Nonetheless, the study of this specific adverb (i.e. *however*) has been understudied in contrast with others, like *actually* (Pérez-Paredes and Bueno-Alastuey, 2019). The present study intends to redress this imbalance by focusing on the written production of 86 B1 students with Spanish as their L1, who were requested to write a pros and cons article. This convenience sample is part of the larger learner corpus FineDesc, which encompasses authentic writings by non-native learners of English at different levels and belonging to different text types (e.g. argumentative essays, emails, narratives, etc.). More specifically, the aim of this paper is to contrast these learners' use of adverbs and expressions of contrast such as 'however', 'nevertheless' or 'on the other hand' with native speaker use, using the subsection of LOCNESS corpus of British University students' essays. The analysis of the data will be aided by the use of Sketch Engine. Results show an overuse of initial adverbs of contrast, prepositional errors and lexical transfers which can be targeted explicitly in the classroom to improve the students' overall linguistic competence.

References

Chen, Z. 2012. Expression of Epistemic Stance in EFL Chinese University Students' Writing. *English Language Teaching*, *5*(10), 173-179.

Gilquin, G. & Paquot, M. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction*, *1*(1), 41-61.

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), 155-184.

Mora Díaz, L. M, & Gómez Orjuela, Y. 2021. Understanding the English language through a creative writing workshop: Adjectives and Adverbs essentials for EFL (English as a Foreign Language) learners. *Shimmering Words: Research and Pedagogy E-Journal*, *11*, 52-73.

Neff van Aertselaer, J. 2008. Contrasting English-Spanish interpersonal. *Phraseology in foreign language learning and teaching, 138*, 85.

Pérez-Paredes, P. & Bueno-Alastuey, M. C. 2019. A corpus-driven analysis of certainty stance adverbs: Obviously, really and actually in spoken native and learner English. *Journal of Pragmatics*, *140*, 22-32.

Pérez-Paredes, P. & Díez-Bedmar, M. B. 2019. Certainty adverbs in spoken learner language: The role of tasks and proficiency. *International Journal of Learner Corpus Research*, *5*(2), 253-279.

Pérez-Paredes, P. & Sánchez-Tornel, M. 2014. Adverb use and language proficiency in young learners' writing. *International Journal of Corpus Linguistics*, *19*(2), 178-200.

Suzuki, C., Fukushima, S., Kinjo, Y., Yoshihara, S., & Coxhead, A. 2012. Research results of corpus studies on language use in academic writing by EFL students compared with native speakers. *La investigación y la enseñanza aplicadas a las lenguas de especialidad y a la tecnología*, 139.

Yilmaz, E. & Dikilitas, K. 2017. EFL Learners' Uses of Adverbs in Argumentative Essays. *Novitas-ROYAL (Research on Youth and Language)*, *11*(1), 69-87.

Use of English loanwords containing V-ING type forms in Spanish, French and Italian: A study based on the Prague Aranea web corpora

François Maniez, María Belén Villar Díaz & Sandra Garbarino – *CeRLA, Lyon 2 University*

Keywords: *-ING morpheme, borrowing, comparable corpora, lexicology, loanwords, neology, Romance languages.*

Words starting with a verb root and ending with the *-ing* morpheme feature prominently among words borrowed from English in many Indo-European languages, and the rising popularity of the *-ing* morpheme has been attributed by some scholars (Picone 1996) to its nominalizing syntactic function. Such borrowings are frequently followed by the creation of equivalents coined by using native words in the receiving language, and occasionally by their inclusion in standard dictionaries (cf. the case of *brainstorming* and *remue-méninges* for French, Humbley 2008). Such neologisms occasionally present as hybrid borrowings (*relooking* or *surbooking* in French), or pseudo-Anglicisms such as *mailing*.

Using the Prague University Aranea web corpora, we studied the use of words beginning with a verb base and ending with the *-ing* morpheme (e.g. *shopping*) in Spanish, French and Italian. The extent of the borrowing phenomenon was found to be partially domain-dependent, as many loanwords seemed to relate to the domain of economy and finance (*banking*, *dumping, holding*, *rating, trading*), sports (*diving*, *jogging*, *racing*, *rafting*, *trekking*) as well as technology and communications (*e-learning*, *mailing, phishing, roaming, spamming, streaming*).

All borrowings from English ending with the *–ing* morpheme were extracted from the Prague Aranea web corpora in all three Romance languages. From a quantitative point of view, Italian seemed to be the language with most such borrowings from English, whereas French showed relatively more resistance to *-ing* loanwords, and Spanish even more. Italian seemed to exhibit a lesser degree of resistance to *-ing* forms than Spanish and French for many such forms (*doping*, *overbooking*, *roaming*, *trading*), some of them being used for concepts usually associated with other English words (e.g. *mobbing* for *harassment* in Italian, *zapping* for *channel surfing*).

Taking into account all of the *–ing* forms used in the Aranea web corpora for the three languages under study confirms the findings of previous research on the same topic (Maniez 2014) based on the Europarl parallel corpus (Tiedemann 2009) in terms of relative frequencies for the three languages concerned. However, by relying solely on comparable corpora for extraction of *–ing* forms, the present research eliminates translation biases and includes a much wider set of data, with an average of 3,16 per million words for the 100 most frequent *–ing* forms in the *Araneum Italicum Maius* corpus.

From a qualitative point of view, we focus more specifically on 35 high-frequency recent borrowings (such as *casting*, *coworking*, *crowdfunding*, *microblogging*, *networking*, *outsourcing*, *phishing*, *sponsoring* or *storytelling*) and examine their collocational patterns. For that particular set, similar quantitative results may be observed as in the whole corpus, with 41% more French *–ing* tokens and 63% more Italian *–ing* tokens than in Spanish.

We also focus on forms that are used much more widely in one language than in the other two, and on their translation equivalents in those languages, using the InterCorp v15 parallel corpus (Čermák & Rosen 2012). Such forms include words like *auditing*, *franchising*, *peacekeeping*, *porting* or *stalking* for Italian, *cocooning*, *relooking*, *snacking* or *teasing* for French and *bullying*, *greening*, *overclocking* or *ranking* for Spanish.

Bibliography

Alvar Ezquerra, M. (1995). *La formación de palabras en espagnol*, Madrid, Arco Libros.

Benardi, R. L. L'italien des institutions publiques, une langue bien perméable aux anglicismes. *N 13–décembre 2014*, 16.

Benko, V. (2014): Aranea: Yet Another Family of (Comparable) Web Corpora. In Sojka, P. – Horák, A. – Kopeček, I. – Pala, K. (eds), *TSD 2014, LNAI 8655*, 257–264. Springer International Publishing.

Bistarelli, A. (2008). L'interferenza dell'inglese sull'italiano. *inTRAlinea*, *10*, 1-11.

Cartier, E. & Lazar, J. (2021). Les anglicismes en français et en tchèque contemporains: le cas des formes en –ing. *AUC PHILOLOGICA*, *2020*(4), 117-132.

Čermak, F., Rosen, A. (2012): The case of InterCorp, a multilingual parallel corpus. In *International Journal of Corpus Linguistics*, 17(3), 411–427.

Furiassi, C., Pulcini, V., et González, F. R. (ed.) (2012). *The anglicization of European lexis*. John Benjamins Publishing.

Gómez Capuz, J. (1997). Towards a typological classification of linguistic borrowing (illustrated with anglicisms in romance languages). *Revista alicantina de estudios ingleses, No. 10 (Nov. 1997); pp. 81-94*.

Grossmann, M., Rainer F. (dirs.) (2004). *La formazione delle parole in italiano*, Tübingen, Niemeyer.

Humbley J. (2008). Emprunts, vrais et faux, dans le Petit Robert 2007, In Pruvost J. (dir.), Les journées des dictionnaires de Cergy: Dictionnaires et mots voyageurs. Les 40 ans du Petit Robert, de Paul Robert à Alain Rey, 221-238, Herblay, Éditions des Silves.

Maniez, F. (2014). "Implantation of English terms including the -ING morpheme in French, Spanish and Italian: A corpus-based study of the debates of the European Parliament" in Pascaline Dury, José Carlos de Hoyos, Julie Makri-Morel, François Maniez, Vincent Renner & María Belén Villar Díaz (dirs) *La néologie en langue de spécialité: détection, implantation et circulation des nouveaux termes*, pp. 189-201, Lyon, Travaux du CRTT.

Maniez, F. (2020). Use of English loanwords containing V-ING type forms in French and Italian. *Anglistica AION: An Intersciplinary Journal*, *24*(2), 83-98.

Nadvornikova, O. (2017). Le corpus multilingue InterCorp: nouveaux paradigmes de recherche en linguistique contrastive et en traductologie. *Studii de lingvistică*, *7*.

Picone M. D. (1996). *Anglicisms, Neologisms and Dynamic French*, Amsterdam, Benjamins.

Renner, V. & Fernández-Domínguez, J. (2015). 6. False Anglicization in the Romance languages: A contrastive analysis of French, Spanish and Italian. In *Pseudo-English* (pp. 147-158). De Gruyter Mouton.

Tiedemann J. (2009). News from OPUS – A collection of multilingual parallel corpora with tools and interfaces, in Nicolov N., Angelova G., Mitkov R. (eds.), *Recent Advances in Natural Language Processing V: Selected papers from RANLP 2007*, 237-248, Amsterdam, Benjamins.

Internet sources

Aranea, A Family of Comparable Gigaword Web Corpora, http://ucts.uniba.sk/aranea_about/index.html

InterCorp v15 parallel corpus, https://wiki.korpus.cz/doku.php/en:cnk:intercorp

OPUS, the open parallel corpus, http://opus.lingfil.uu.se/

Analyzing the diachronic variation of morphological productivity through textual genres and lexical domains with corpora

Valentina Maniglia – *Università degli Studi di Salerno*

Keywords: *diachrony, productivity, textual genres, lexical domains, corpora.*

The reliability of corpora for linguistic investigations is a largely debated question among scholars of different subfields. Although even in the field of morphological productivity there is active debate about that issue, different scholars concordantly argued that corpus linguistics allows to establish the variation of productivity through registers, genres, and domains (see, among others, Baayen 2009, Lefer 2012, etc.).

This general idea comes from considerations on the synchronic dimension of productivity. Our contribution aims to show that, from a diachronic perspective, corpora can be a precious source for detecting aspects of morphological productivity starting from the analysis of these three dimensions (register, genres, domains). Such analysis allows to enrich the diachronic method which generally consists in measuring productivity in terms of new words attested (see a.o. Neuhaus 1973; Schroder 2011; Berg 2020, 2021).

Bringing evidence from the case of prefixed words in the diachronic French corpus Frantext, in particular, we will see that, on the one hand, it is not always possible to make exhaustive conclusions about the historical quantitative variation of prefixal productivity in the French language only recurring to one corpus; on the other hand, a large corpus can offer preciuos inputs on the paths of diffusion of prefixes' productivity in terms of expansion of the variety of the textual genres and the number of lexical domains involved in their use (cf. also Peytard 1973).

We will provide data about different prefixes, such as *ex-, semi-, anti-*, etc. To give one example, if we want to look at the quantitative productivity of *ex-* over time (i.e. to count how many new words are coined with this prefix in different time spans), we should conclude that starting from a dozen of words coined in the 18th century, the prefix becomes largely productive in the 19th century (with about 50 new words coined) and loses productivity again in the 20th century. If we analyse the lexical domains the bases attached to the prefix belong to, we can observe a clear progression in the kind of domains: the first words present in the corpus seem to belong exclusively to the religious lexical domain (f.e. *ex-jésuite*, *ex-religieux, ex-franciscain, ex-capucin* etc). The more we progress in time, the more the lexical domains involved increase: from religion to politics (*ex-ministre, ex- président, ex-deputé*) and aristocracy (*ex-noble, ex-roi, ex-comte*), to more and more common professions (*ex- employé, ex-parfumeur- ex-negociante*) etc.

Not only do the lexical domains of the bases suggest a possible origin and a path of diffusion, but they also suggest what specific texts to look at for more exhaustive details about its changes in productivity (in terms of quantity). In fact, looking at religious texts we find that many other words are actually attested between the 17th and the 18th century (*ex-abbé, ex-pontife, ex-moine* etc.). As in these centuries religion influenced politics and administration, we find that in administrative texts and in newspapers articles of the following century, there are hundreds of words attested on the type *ex- ministre* etc.

In conclusion, we will show that the analysis of register, genres and domains is a fundamental step in looking at variation of productivity, especially for some morphological elements, which are not easily detectable with a simple search in a corpus.

Bibliography

ATILF. *Base textuelle Frantext* (En ligne). ATILF-CNRS & Université de Lorraine. 1998-2022. https://www.frantext.fr/ (consulté le 22 décembre 2022).

Baayen, Harald. 2009. Corpus linguistics in morphology: Morphological productivity. 10.1515/97831102-13881.2.899.

Berg, K. 2020. Changes in the productivity of word-formation patterns: Some methodological remarks. *Linguistics*, *58*(4), 1117-1150. https://doi.org/10.1515/ling-2020-0148.

Berg, K. 2021. Productivity, vocabulary size, and new words. A response to Säily (2016). *Corpus Linguistics and Linguistic Theory*, *17*(1), 177–187.

Lefer, Marie-Aude. 2012. La préfixation française à travers les genres et les domaines: étude de corpus. *SHS Web of Conferences.* 1. 10.1051/shsconf/20120100251.

Neuhaus, H. 1973. Zur Theorie der Produktivität von Wortbildungssystemen. In A. Cate & P. Jordens (Ed.), *Linguistische Perspektiven* (pp. 305-318). Berlin, New York: Max Niemeyer Verlag. https://doi.org/10.1515-/9783111712345.305.

Peytard Jean. 1973. De la diffusion d'un élément préfixal: « mini- ». In: *Langue française*, n°17. Les vocabulaires techniques et scientifiques. pp. 18-30; doi: https://doi.org/10.3406/lfr.1973.5618.

Schröder, A. 2011. On the Productivity of Verbal Prefixation in English (1st ed.). Narr Francke Attempto Verlag. Retrieved from https://www.perlego.com/book/1016797/on-the-productivity-of-verbal-prefixation-in-english-synchronic-and-diachronic-perspectives-pdf (Original work published).

The depiction of women in business texts: A corpus-driven study

María José Marín-Pérez,[1] Ángela Almela Sánchez-Lafuente[1] & Camino Rea-Rizzo[2]

University of Murcia[1] – Universidad Politécnica de Cartagena[2]

Keywords: *ESP, business English, specialised corpora, gender studies, discourse studies.*

Gender representations have been analysed in different fields implementing corpus-based/driven methodologies, using news texts as the major data sources (Baker, 2010; Baker & Levon, 2015; Zottola, 2018; Wilkinson, 2019; Karimullah, 2020). As regards business English, the number of analyses of this sort is much scarcer (Fuertes-Olivera, 2007; Power, Rak & Kim, 2020; Vázquez-Amador & Lario-de-Oñate, 2022), probably owing to the reduced amount of corpora available within this ESP (English for Specific Purposes) variety.

The *Wolverhampton Business English Corpus*[1] (*WBE*) is one of such datasets, it was collected from 23 different websites within the business area between the years 1999 and 2000. Its text typology varies considerably, reaching over 10 million words. This was the chief source used for the present research, still in progress, which aims at exploring the semantic preference of terms like *man* and *woman* as a point of departure towards the characterisation of gender roles within the business sphere. To that end, the collocate networks associated with both lexical items will be scrutinised and their main constituents classified into semantic categories. Their word sketches[2] (Kilgarriff *et al.,* 2014) will also be considered, as previous studies (Baker, 2010; Horton, 2018) show how syntactic roles such as object v. subject also contribute to the representation of one or the other gender in a very specific way, in conjunction with the semantic preferences deployed in each case.

Preliminary results complement the analysis of lexical gender as found in the *WBE* corpus by Fuertes-Olivera (2007), who concludes that the form *Ms.* seemed to have taken over *Mrs.* and *Miss* within the professional environment by the time the corpus started being available, whereas the business sphere remained dominated by males. In spite of that fact, having studied the collocates of *woman* and *man*, from a quantitative angle, the noun *woman* attracts a greater amount of collocates (709) than *man* (419) in the *WBE*. Nonetheless, their qualitative analysis goes in line with the findings shown above, as it is observed that motherhood is a relevant issue in the corpus. The term *woman* collocates with the modifier *pregnant* and is often followed by the prepositional phrase *with child,* whereas the term *man* is not associated with any of these ideas.

On the other hand, men's roles within companies, as defined by the verbs they collocate with, are depicted as active subjects who *buy, build, share, make* or *think*, whereas women, although there are some instances of collocates like *run*, *work* or *make*, tend to be associated with verbs that portray them as rather passive subjects, for instance, *need*, *learn* or *want,* according to their context of usage, also being linked to the lexical items such as *minority* or *discrimination.*

Finally, in order to compensate for the fact that the *WBE* is 23 years old, considering that there are no business corpora available that cover the second decade of this century, we decided to compare the usage of the compounds of the term *woman* (*spokeswoman, businesswoman, woman-friendly or woman-run*, amongst other), as found in the *WBE*, with their usage in the *SiBoL*[3] corpus of British newspapers, which includes a business section and covers from 1993 to 2021, so as to study their evolution throughout the 21st century.

References

Baker, P. 2010. 'Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English'. *Gender and Language,* 1(1): 125-149.

Baker, P., Levon, E. 2015. 'Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity'. *Discourse and Communication*, 9(2): 143-271.

Fuertes-Olivera, P.A. 2007. 'A corpus-based view of lexical gender in written Business English'. English for Specific Purposes, 26: 219–234.

Horton, R. H. 2018. 'A corpus analysis of girl and boy in spoken academic English and teaching activities to raise awareness about gendered discourse'. *TESOL Working Paper Series*, 16: 1-18.

Karimullah, K. 2020. 'Sketching women: a corpus-based approach to representations of women's agency in political Internet'. *Corpora*, 15(1): 21-53.

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: 7-36.

Power, K., Rak, L., Kim, M. 2020. 'Women in business media: A Critical Discourse Analysis of Representations of Women in Forbes, Fortune and Bloomberg BusinessWeek, 2015-2017'. *Critical Approaches to Discourse Analysis across Disciplines*, 11(2): 1-26.

Vázquez-Amador, M., Lario-de-Oñate, M.C. 2022. The Role of Women in Business English Textbooks (1970s-2010s). *ESP Today*, 10: 145–168.

Wilkinson, M. 2019. "Bisexual oysters': A diachronic corpus-based critical discourse analysis of bisexual representation in The Times between 1957 and 2017'. *Discourse and Communication* 13(2): 149-275.

Zottola, A. 2018. 'Transgender identity labels in the British press: A corpus-based discourse analysis'. *Language and Sexuality*, 7(2): 237-262.

[1] See http://www.elda.org/catalogue/en/text/W0028.html for more information on the *WBE* Corpus.

[2] Lists of collocates classified according to their grammatical categories and their syntactic bonds such as subject of, object of, modifier and the like.

[3] See https://www.sketchengine.eu/sibol-corpus/ for more information on the SiBoL corpus of British newspapers.

Analysis of subordination clauses in different newspaper domains: does the subject influence the syntactic structure?

Virginia Mattioli – *Independent researcher*

Keywords: *newspaper articles, newspaper domains, genre analysis, subordination, finite/non-finite clauses.*

According to register theory, linguistic features are correlated with specific communicative purposes and situational contexts of the texts (Biber & Conrad, 2019: 2). If many authors describe and compare different text types from a functional perspective, highlighting the importance of distinguishing and describing them (Rafajlovicová 2008; Biber & Conrad, 2019), few of them focus on the domains of a specific text type. Firstly, previous literature does not present any agreement about the definition of domain (van der Wees et al. 2015: 560) which each author describes from a different perspective, as the topic of a text or as a kind of text type, as Garzone (2015: 1) who considers academic or legal discourse examples of domain-specific communication; secondly, the definition of register analysis itself, as a linguistic examination focussed on the register, implies the analysis of "text varieties of a language associated with particular situations of use" (Biber, 2013: 191), hence, regardless the main subject of the text.

This paper aims to investigate whether the main assumption of the register theory could be applied also to the textual domains within the same text type, considering domain as a "broad 'subject field'", according to the definition proposed by Burnard (1995: 5) later accepted by Lee (2001: 51). Actually, even if the situational context could remain unvaried depending on the domain, the subject field could affect the communication purpose, hence, the linguistic features of the text. Starting from this research question, the objective of this paper is comparing subordination across five domains of newspaper articles, according to the hypothesis that the syntactic structure of the texts in analysis varies according to their domain. The study focusses on subordination for two reasons: firstly, sentences are the base of communication and ideas exchange as texts are created by their combination; secondly, subordination is an index of structural complexity in a language and, as such, it is frequently examined in contrastive studies involving different types of texts (Rafajlovičová, 2008:64).

The examined material includes a set of five corpora, balanced in terms of number of texts, representing tantamount domains of newspaper articles, namely, culture, economics, science, society and international news. The articles were selected from four English international newspapers: *The Guardian*, *CNN*, *BBC* and *The Express*. The corpus-based methodology adopted for the analysis involves four steps and allows for identifying and comparing the subordinate clauses of each corpus starting from the conjunctions they are introduced by. Firstly, an exhaustive list of subordinate introductive items is selected. The list, created at St Mt San Jacinto College for teaching purposes, incudes conjunctions and relative pronouns. Secondly, each item is searched for in each corpus by using the concordance list tool provided by AntConc (Anthony, 2022), which allows to observe the results in context. Thirdly, the results of the search for each item are reviewed in order to eliminate those cases in which the item searched is not used to introduce a subordinate clause and to distinguish finite and non-finite clauses. Finally, the frequency of all the resulting introductive items is added: as each item introduces a subordinative clause, their total frequency will correspond to the total number of subordinative clauses. The five corpora are analysed separately and, finally, the results are compared considering four variables: the type of subordinate clauses (nominal relative or adverbial), their function (conditional, causal, temporal, etc.), their finite or non-finite nature and, in case of non-finite structures, their verb (infinitive, gerund, participle).

The outcomes suggest the influence of the domain on the grammatical features of a text, calling for further studies examining greater corpora, different linguistic features and/or other text types. Examining the variation of the subordination in texts according to their domain will offer information about the use and the function of specific linguistic features within a certain topic, contributing to the description of the language from a functional

perspective with impacts on research and education. On the one hand, defining the specific grammatical features which characterize each domain will contribute to stylistics and discourse analysis from a theoretical perspective. On the other hand, as teaching a language implies teaching the use and the function of its grammatical forms (Rafajlovičová, n.d.:42), the results will improve L1 and L2 courses, by relating specific grammatical forms to specific topics. Similarly, offering information about the most frequent subordinate structures and forms used in newspaper articles according to their domain will be a useful resource also for journalism studies.

References

Anthony, L. 2022. AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software>.

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

Biber, D. 2013. "Register and discourse analysis." In *The Routledge Handbook of Discourse Analysis*, Routledge, 191-208.

Burnard, L. (Ed.) 1995. *The British national corpus users reference guide* (SGML version. Part of release 1.0 of the BNC). Oxford, UK: Oxford University Computing Services.

Garzone, G. E. 2015. "Genre analysis". *The international encyclopedia of language and social interaction*, John Wiley & Sons Inc., 1-17.

Lee, D. Y. W. 2001. "Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle", *Language Learning and Technology*, 5 (3), 37-72.

Rafajlovičová, R. 2008. The distribution of finite and non-finite subordinate clauses according to text type. *Linguistics Journal: Discourse and Interaction*, 1(2), 64-72.

Rafajlovičová, R. n.d. Subordinate clauses–their forms and functions in different text types. <https://www.academia.edu/download/78539018/Rafajlovicova.pdf>.

Van der Wees, M., Bisazza, A., Weerkamp, W. & Monz, C. 2015, July. "What's in a domain? Analyzing genre and topic differences in statistical machine translation." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol. 2, 560-566.

A study on the semantic preference of English near-synonymous suffixes through linguistic motion chart: Taking *-proof* vs. *-resistant* as an example

Qingnan Meng[1] & Martin Hilpert[2] – *Dalian Maritime University*[1] – *University of Neuchâtel*[2]

Keywords: *near-synonymous suffixes, semantic preference, diachronic study, linguistic motion chart.*

Previous studies on constructional alternations mainly focus on lexical, syntactic or discoursal levels. According to Goldberg's (1995) classical claim that "it's constructions all the way down", this phenomenon should apply to lower levels (e.g.: affixes) as well. On this account, this study explores the variation and changes in frequency distribution and the semantic preference of the nominal stems of two near-synonymous suffixes *-proof* and *-resistant* both diachronically and synchronically based on COHA (1820-2019) and GloWbE corpora. The results are visualized by static linguistic motion chart, through {ggplot2} package in R, following Primahadi-Wijaya-Rajeg and Rajeg's (2018) tradition.

It is shown that in all English varieties, these two suffixes display a distinct division of labor: *bullet-proof* is the most frequently used adjectival compound, followed by nouns related to costumes or means of transportation, whereas *-resistant* mainly collocates with medical and biochemical nominal stems such as *drug*, *antibiotic* and *methicillin*. In American English, the frequency of N-*resistant* increases sharply during the past 20 to 30 years, and it may well become a substitute for N-*proof* compounds in the near future. In other English varieties, due to geographical and sociocultural factors, the nominal stems collocating with these two suffixes boast some unique regional features. For instance, *rabbit-proof* is predominantly used in Australian English, so it is with *earthquake-proof*/*resistant* in New Zealand English, *fool-proof* in Indian, Kenyan and Singapore English, and *moisture-proof* in Hong Kong English. Apart from American English, in the other 19 English varieties, the type and token frequencies of N-*resistant* are still both relatively low, which displays their conservativeness in language use. Whether they may undergo "Americanization" in the near future remains to be further observed.

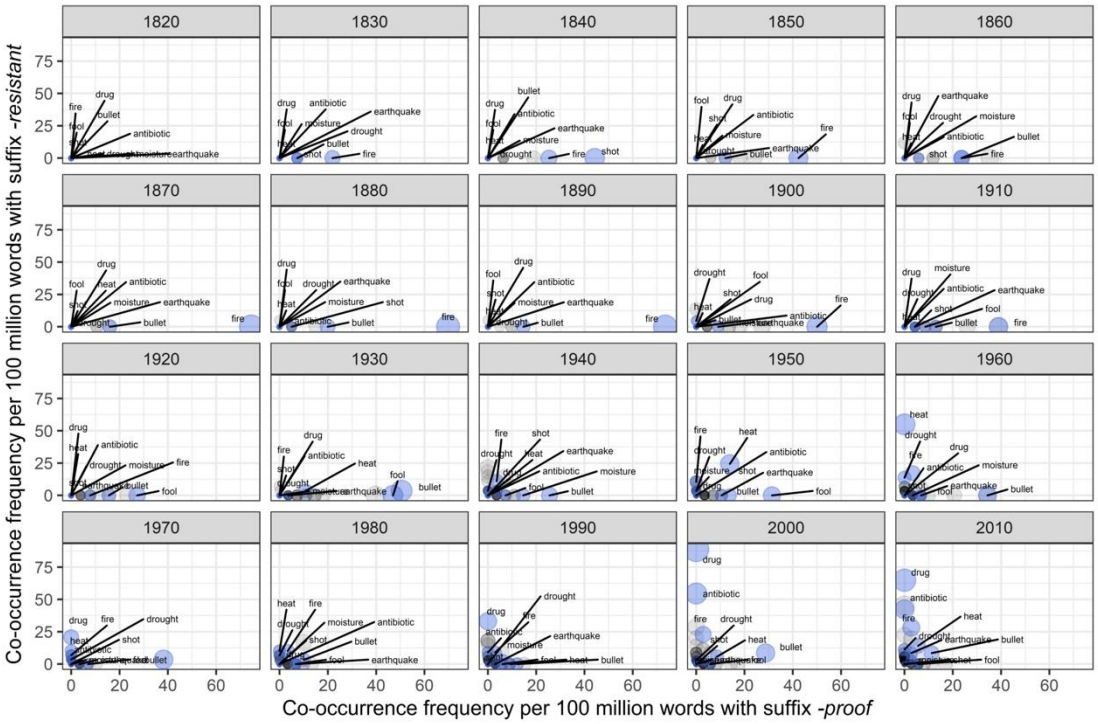

Figure 1 Frequency distribution of nominal stems cooccurring with "-proof" and "-resistant" in COHA (1820-2019) (excluding two outliers "water" and "methicillin")

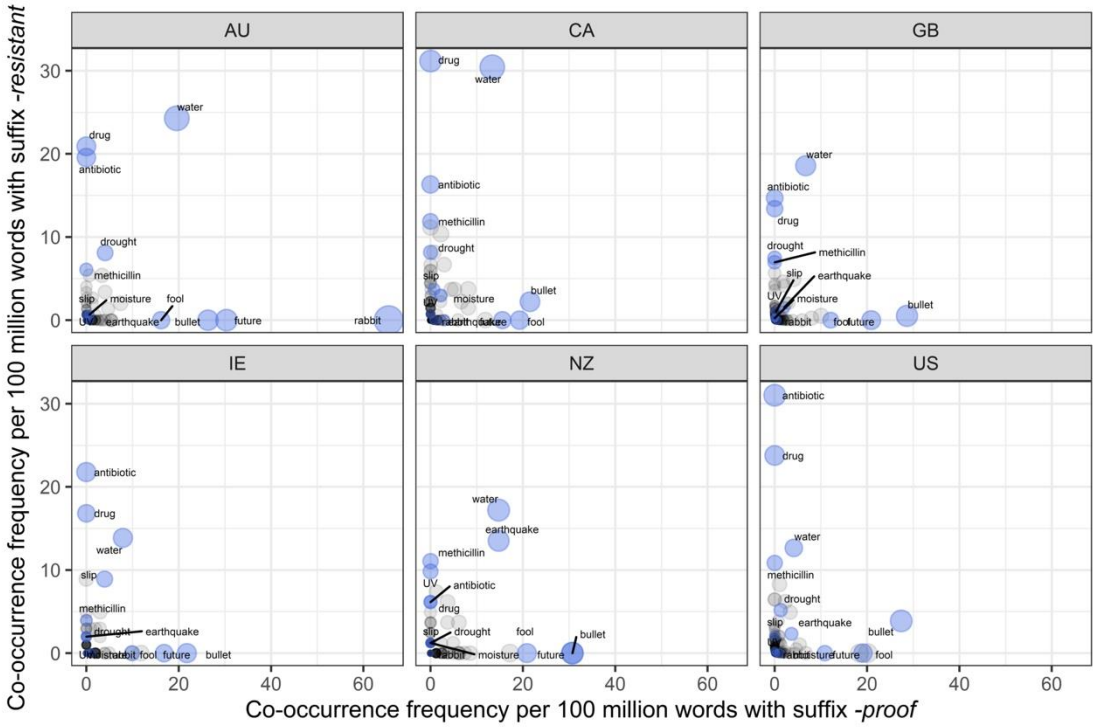

Figure 2 Frequency distribution of nominal stems cooccurring with "-proof" and "-resistant" in six inner circle countries based on GloWbE corpus

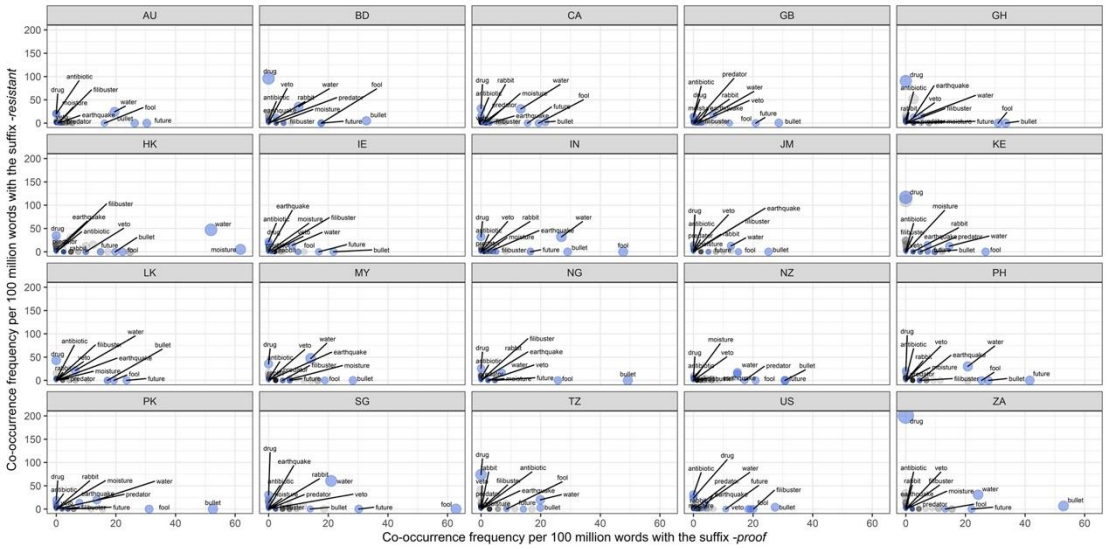

Figure 3 Frequency distribution of nominal stems cooccurring with "-proof" and "-resistant" in 20 English varieties based on GloWbE corpus

Bibliography

Davies, M. *Global Web-based English Corpus (GloWbE)*: 1.9 billion words, 2012-2013 [DB/OL]. [2021-09-25] https://www.english-corpora.org/glowbe/.

Davies, M. *The Corpus of Historical American English (COHA)*: 475+ million words, 1820-2019 [DB/OL]. [2021-09-25]. https://www.english-corpora.org/coha/

Davies, M. & Fuchs, R. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE) [J]. *English World-Wide*, (1): 1-28.

Goldberg, A. E. 1995. Constructions: A Construction Grammar Approach to Argument Structure [M]. Chicago: Chicago University Press.

Hilpert, M. 2011. Dynamic visualizations of language change: motion charts on the basis of bivariate and multivariate data from diachronic corpora [J]. *International Journal of Corpus Linguistics*, 435-461.

Kachru, B. B. 1985. Standards, Codification and sociolinguistic realism: the English language in the outer circle [A]. In R. Quirk & Widdowson, H. *English in the World: Teaching and Learning the Language and Literatures* [C]. Cambridge: Cambridge University Press, 11-30.

Pichler, K. 2016. *A Diachronic Perspective on Synonymy* [D]. Vienna, Austria: University of Vienna.

Primahadi-Wijaya-Rajeg, G. & I. M. Rajeg. 2018. Generating static linguistic motion charts [DB/OL]. *RPubs*. https://rpubs.com/primahadi/static_motion_charts (05 February 2018).

False friends or cognates? Using corpus data for checking out Spanish-Russian translation equivalents for obscene expressions

Mikhail Mikhailov – *Tampere University*

Keywords: *parallel corpora, comparable corpora, Russian-Spanish phraseology, cognates, obscene language.*

In any pair of languages one can find a lot of differences and some similarities. Spanish and Russian are very different lexically and morphologically, and language contacts were scarce due to geographical distances. Still, numerous similarities can be observed between the two languages. Some of them may be explained by borrowing via literary translation or via mutual sources like the Bible, some are plain coincidence (see e.g. Kreidlin 2022, Kutieva 2009). One of the most surprising similarities is the almost direct matching of obscene expressions containing the word *c . . . jo* / *h.j* 'male sexual organ'.

The Spanish expression *mandar al c . . . jo* and the Russian *poslat' na h.j* 'to send to the mail sexual organ' and their derivatives have very similar structure and are quite close in meaning: 'to refuse to cooperate by insulting the other party'. The main difference is that while in Spanish these expressions are just offensive language, the corresponding Russian expressions are believed to be extremely rude and are subject to taboo. For this reason some researchers call these matches false cognates and recommend to use other translation equivalents e.g. *poslat' k čërtu* 'to send to the devil' (e.g. Henao 2015). It is obvious, however, that comparing dictionary definitions and using own language competence do not provide a final answer and the data from parallel and comparable text corpora would be needed (McEnery & Xiao 2007).

For a better study of cross-lingual correspondences, a corpus-based study was performed. The search in Spanish-Russian parallel corpus from the Russian National Corpus (RNC) produced only 10 examples from literary texts with all Russian translations using non-obscene translation equivalents. However, the parallel Spanish-Russian concordance obtained from OpenSubtitles2018 (Sketch Engine) demonstrates a completely different picture: about one third of the total 1500 bitexts have obscene expressions as Russian translation equivalents. However, most of the data in OpenSubtitles has English as a source language and the concordance in question is in fact pseudo-parallel. Besides, most of the translations are performed by amateurs and are not intended for public watching.

To settle the matter, the data from monolingual corpora of Spanish and Russian was checked. The comparison of frequencies both from gigaword web corpora at Sketch Engine (esTenTen2018, ruTenTen2017) and from RNC and corpora of Spanish language from the Spanish Royal Academy (CREA, CORPES) shows that both in Spanish and in Russian exist similar trends in using the expressions in question with the only difference that the Russian *poslat' na h.j* is often replaced by euphemistic expressions *poslat' na her* 'letter X' / *hren* 'radish' / *fig* 'fig', etc. This means that these expressions indeed can be used (with certain precautions) as translation equivalents for the Spanish expression *mandar al c . . . jo.*

To sum everything up, the corpus data suggests that the Spanish-Russian expressions in question have become closer during the last decades and can be treated as cognates. Using parallel and comparable corpora provides important findings that make it possible to check interlingual lexical correspondences in much more effective manner than traditional methods permit.

References

Henao J. P. Duque 2015. One hundred years of solitude» in Russian translations [in Russian]. *Vestnik of Moscow State Linguistic University. Humanities.* 27 (738), 32-39.

Kreidlin, G. E., 2022. Semiotic conceptualization of the body and the comparative phraseology [in Russian]. *RSUH/RGGU Bulletin*. Series, 4:3, 338–347.

Kutieva, Marina 2009. Semantics of Phraseological Units with ornithonyms in Russian and Spanish languages. *Eslavística Complutense*, 2009, 9, 97-113.

McEnery, T. & Xiao, R., 2007. Parallel and comparable corpora: What is happening? In G. Anderman, & M. Rogers (Eds.). *Incorporating Corpora: the Linguist and the Translator*, 18-31. Clevedon.

Exploring indicators of L2 multi-word verb knowledge in the "English profile"

Elaine Millar – *University of Cantabria*

Keywords: *language learning, learner corpora, wordlists, CEFR, formulaic language.*

Multi-word verbs have long been perceived as a characteristic feature of English language usage (Bolinger, 1971), but it is really only fairly recently that corpus linguistic researchers have uncovered the empirical data to support this observation (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Claridge, 2000; Davies, 2009; Liu, 2011; Rodríguez-Puente, 2019). At the same time, corpus linguistic studies have also indicated that speakers of English as a second or foreign language (EFL) often struggle to achieve a 'native-like' degree of mastery of this phenomenon (Alejo-González, 2010a, 2010b). The present paper explores this issue, by examining indicators of learner English multi-word verb knowledge in the *English Profile* (https://www.englishprofile.org/). The *English Profile* is a largescale corpus-informed database which provides insight into the typical lexical-grammatical knowledge of EFL users at each level of the *Common European Framework of Reference for languages* (CEFR) A1-C2 proficiency scale (Council of Europe, 2001). The database draws from various sources, including the 50-million-word *Cambridge Learner Corpus* which is based on written examination scripts from candidates of *Cambridge English Language Assessment* examinations, and the *English Profile Corpus*, a purpose-built collection of language production collected from authentic EFL classroom contexts around the world.

To carry out the study, the database was consulted and information relating to multi-word verbs was transferred to *Microsoft Excel* software for analysis. A total of 526 multi-word verb lemmas and 728 senses were found to be listed in the database. The analysis suggests that the typical EFL learner will know and use four MWV lemmas and senses at A1 level, with a further 23 new lemmas and 27 senses added to their repertoire at A2 level. This number rises to 109 new MWV lemmas and 134 senses at B1, and when learners reach B2, the highest proportion of new MWV lemmas and senses come into use (196 and 278, respectively). Figures drop at C1 (62 new MWV lemmas, 99 new MWV senses) and then increase again at C2 level (132 new MWV lemmas, 186 new MWV senses). In terms of distribution by syntactic categories, most senses (511) listed in the database function as phrasal verbs, followed by prepositional verbs (170 senses) and phrasal-prepositional verbs (38 senses). There are also nine other uncategorisable MWV constructions listed in the database (e.g. *hold out hope*, *keep to yourself*). The *English Profile* is a unique reference source in that it can provide indicators of how a learners' knowledge of this phenomenon deepens semantically and syntactically as they progress through the CEFR proficiency scale. It is important to note, however, that the database is designed primarily for use in English language teaching practice (e.g. syllabus design, materials writing, assessment, etc.). As such, it does not provide raw data but rather curated information on learners' key knowledge. In light of this, the results presented here should be taken with caution, and a potential avenue for further investigation would be to conduct a direct analysis of multi-word verb usage in the corpus sources that inform the *English Profile* (i.e. the *Cambridge Learner Corpus* and *English Profile Corpus*).

References

Alejo-González, R. 2010a. L2 Spanish acquisition of English phrasal verbs: A cognitive linguistic analysis of L1 influence. In M. C. Campoy-Cubillo, B. Bellés-Fortuño, & M.-L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 149–166). London/New York: Continuum International Pub. Group.

Alejo-González, R. 2010b. Making sense of phrasal verbs: A cognitive linguistic account of L2 learning. *AILA Review*, *23*(1), 50–71. https://doi.org/https://doi.org/10.1075/aila.23.04ale.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Bolinger, D. 1971. *The phrasal verb in English*. Cambridge: Harvard University Press.

Claridge, C. 2000. Multi-word verbs in early modern English: a corpus-based study. Amsterdam-Atlanta: Rodopi B.V.

Council of Europe. 2001. Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge University Press.

Davies, M. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190. https://doi.org-/10.1075/ijcl.14.2.02dav.

Liu, D. 2011. The Most Frequently Used English Phrasal Verbs in American and British English: A Multicorpus Examination. *TESOL Quarterly*, *45*, 661–688. https://doi.org/10.2307/41307661.

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

Rodríguez-Puente, P. 2019. The English Phrasal Verb, 1650–Present: History, Stylistic Drifts, and Lexicalisation. Cambridge University Press.

Idiolectal stability across genres in Mexican Spanish

Andrea Mojedano Batel & Krzysztof Kredens – *Aston University*

Keywords: *idiolect, Spanish, language stability, cross-genre, authorship analysis, forensic linguistics.*

The notion of idiolect is well known in linguistics, but there is a vast lacuna between theoretical understandings of the concept and empirical studies of the phenomenon (Barlow, 2013). This study aims to identify and analyze linguistic features that show idiolectal cross-genre stability in Spanish, filling part of this lacuna. We follow the framework of forensic authorship analysis, which assumes that idiolectal features will recur with a relatively stable frequency (Coulthard et al., 2011). This assumption proves problematic because it is widely accepted that different levels of linguistic structure are susceptible to change over the speaker's lifespan (Labov, 1972; Bailey et al., 1991; Sankoff, 2005), under stress (Pennebaker & Lay, 2002), depending on the speaker's audience (Bell, 1984; Hay & Mendoza-Denton, 2010), across different genres (Taylor, 1994; D'Arcy et al., 2013), etc.

There are only a handful of authorship attribution studies using cross-genre and/or cross-domain data (e.g., Baayen et al., 2002; Goldstein-Stewart et al., 2009; Mollin, 2009; Litvinova et al., 2018a, 2018b), and the present corpus study is the first of its kind examining data in Spanish.

The participants in our study are six Mexican women and three Mexican men, between 30 and 60 years old at the time of their interview (2019-2020), all born and raised in central Mexico, and all currently working as professors or researchers at a public university in Mexico City; their native language is Spanish. We used convenience sampling to obtain the data.

Data collection for analysis involved obtaining the following linguistic data from each of the participants, subsequently anonymized:

1. 30 text messages (6,122 tokens in total), written in 2020.
2. 30 emails (33,203 tokens in total), written between 2019 and 2020.
3. Sociolinguistic interviews with each participant (18,839 tokens and 174 minutes of audio in total), conducted between 2019 and 2020.
4. Transcripts of monthly work meetings which took place between 2008 and 2019 (844,104 tokens in total).

Using word n-grams as our basic analytical category, we identified cross-genre idiolectal stability patterns. Lists of genre-recurring n-grams were generated using spaCy, an NLP library for Python, and a custom-made Python script. After retrieving the n-grams, we dispensed with: (a) all the highly frequent function words in Spanish, due to their low discriminatory potential, and (b) topic-dependent words. We then qualitatively examined contextual use of n-grams through the AntConc program (Anthony, 2017), in order to find idiolectally stable patterns.

Results from the analysis show that stable patterns in our corpus all belong to one of four areas of linguistic functions: (1) evaluative language, (2) deontic modality markers, (3) epistemic modality markers, and (4) expressions of quantity. Most interesting is that many of these stable features allow speakers to express their subjectivity by displaying how certain they are about the information they are providing or by regulating politeness (epistemic stance), or by indicating whether a proposition expressed by a command is obligatory, advisable or permissible (deontic stance): from these findings, we infer that examining expressions of epistemic and deontic modality can greatly aid forensic linguists in authorship analysis tasks, a finding that is in line with what has been reported for English (Kredens, 2002) and Russian (Litvinova et al., 2018a, 2018b), in a demonstration of cross-linguistic idiolectal stability.

References

Baayen, H., van Halteren, H., Neijt, A., & Tweedie, F. 2002. "An experiment in authorship attribution", en *JADT 2002: 6es Journées internationales d'Analyse statistique des Données Textuelles*, 69-75. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.679.2951&rep=rep1&type=pdf.

Bailey, G., Wikle, T., Tillery, J., & Sand, L. 1991. The apparent time construct. *Language variation and change*, *3*(3), 241-264.

Barlow, M. 2013. "Individual differences and usage-based grammar", *International Journal of Corpus Linguistics* 18/4, 443-478. https://dx.doi.org/10.1075/ijcl.18.4.01bar.

Bell, A. 1984. "Language style as audience design", *Language in society* 13/2, 145-204. http://www.jstor.org/stable/4167516.

Bloch, B. 1948. "A set of postulates for phonemic analysis", *Language* 24/1, 3-46.

Coulthard, M., Grant, T., & Kredens, K. 2011. "Forensic linguistics", en R. Wodak, B. Johnstone, & P.E. Kerswill (eds.) *The SAGE Handbook of Sociolinguistics*. London: SAGE publishing, 531-544. https://dx.doi.org/10.4135/9781446200957.n36.

D'Arcy, A., Haddican, B., Richards, H., Tagliamonte, S. A., & Taylor, A. 2013. Asymmetrical trajectories: The past and present of–body/–one. *Language Variation and Change*, *25*(3), 287-310.

Goldstein-Stewart, J., Winder, R., & Sabin, R. E. 2009. Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 336-344).

Grieve, J. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, *22*(3), 251-270.

Hay, J., Jannedy, S. & Mendoza-Denton, N. 1999. Oprah and/ay: Lexical frequency, referee design and style. In *Proceedings of the 14th international congress of phonetic sciences* (pp. 1389-1392). Berkeley, CA: University of California.

Koppel, M., Schler, J., & Argamon, S. 2013. Authorship Attribution: What's Easy and What's Hard? *Journal of Law and Policy, 21*(2), 317-331.

Kredens, K. 2002. Towards a corpus-based methodology of forensic authorship attribution: a comparative study of two idiolects. In B. Lewandowska-Tomaszczyk (Ed.), *PALC'01: Practical Applications in Language Corpora* (pp. 405-437). Peter Lang: Frankfurt am Mein.

Labov, W. 1972. *Language in the inner city: Studies in the Black English vernacular* (No. 3). Philadelphia: University of Pennsylvania Press. https://doi.org/10.1177/089124167600400410.

Litvinova, T., Litvinova, O., & Seredin, P. 2018. Assessing the level of stability of idiolectal features across modes, topics and time of text production. In *2018 23rd Conference of Open Innovations Association (FRUCT)* (pp. 223-230). IEEE.

Litvinova, T., Seredin, P., Litvinova, O., Dankova, T., & Zagorovskaya, O. 2018. On the stability of some idiolectal features. In *International Conference on Speech and Computer* (pp. 331-336). Springer.

Mollin, S. 2009. "I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics, 14*(3), 367–392.

Mufwene, S. 2010. "SLA and the emergence of creoles", *Studies in Second Language Acquisition* 32, 359–400. https://doi.org/10.1017/S027226311000001X.

Pennebaker, J. W. & Lay, T. C. 2002. Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality*, *36*(3), 271-282.

Sankoff, G. 2008. Cross-Sectional and Longitudinal Studies. In U. Ammon, N. Dittmar, K. Mattheier, & P. Trudgill (Eds.), *Volume 2: An International Handbook of the Science of Language and Society* (pp. 1003-1013). Berlin & New York: De Gruyter Mouton. https://doi.org/10.1515/9783110171488.2.7.1003.

Taylor, A. 1994. Variation in past tense formation in the history of English. *University of Pennsylvania working papers in linguistics*, *1*(1), 10.

Wright, D. 2017. "Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem", *International Journal of Corpus Linguistics* 22/2, 212-241. https://doi.org/10.1075/ijcl.22.2.03wri.

Mapping of political events related to the COVID-19 pandemic on Twitter using topic modelling and keywords over time

Antonio Moreno-Ortiz & Carla Fernández-Melendres – *University of Málaga*

Keywords: *topic modelling*, *keywords*, *political events*, *COVID-19*, *Twitter*.

The COVID-19 pandemic outbreak paused societies worldwide. As cities were forced to go on lockdown, people turned to social media platforms like Twitter to discuss ongoing events. This research aims to study the relationship between actual, real-world events related to the COVID-19 pandemic and the impact these events produced on social media. To achieve this objective, we employ topic modelling and keyword extraction techniques. Topic modelling is a Natural Language Processing technique that attempts to identify topics automatically from a collection of documents (Vayansky and Kumar, 2020). This is similar to keyword extraction but, unlike this, topic modelling algorithms return clusters of words that make up the topic. Thus, a second objective is to compare the results of these two methods when it comes to identifying the salient topics in a corpus.

Several studies have looked into the social media dynamics in the context of COVID-19. For instance, Xue et al. (2020) examined COVID-19-related discussions, concerns, and sentiments on Twitter using the machine learning approach known as LDA. Jiang et al. (2020) linked Twitter users to locations within the United States to see local discussions about COVID-19. Boon-Itt and Skunkan (2020) studied the public's perception of COVID-19 by analysing keyword frequency, sentiment analysis, and topic modelling. In Spanish, Argüero-Torales, Villares, and López-Herrera (2021) applied topic modelling to study Twitter discussions at the beginning of the COVID-19 pandemic.

Methodologically, we have used the publicly available and multilingual COVID-19 Twitter dataset collected from January 21, 2020 (and still ongoing) available via the COVID-19-TweetsIDs GitHub repository (Chen, Lerman & Ferrara, 2020). The data is collected using Twitter's streaming application programming interface (API) and the Tweepy library to follow specific keywords and trending accounts. Each tweet is categorised as an original tweet, a retweet (with or without a comment), or a reply. For this study, we will focus on tweets written in English from 2020 and 2021. We limited our study to the years 2020 to 2021, which contains 1 billion tweets (31 billion tokens), and extracted a random, time-stratified sample of 0,1%, which resulted in a total of approximately 1 million tweets (31 million tokens). To our knowledge, there has not been a study which identifies events over a long period of time (2 years), by using topic modelling and keywords.

In terms of methods, we employed unsupervised machine learning methods for both tasks. For topic modelling we used BERT embeddings and the BERTopic library (Grootendorst, 2022). Our script generates a full list of topics and assigned terms, a coherence score, and several data visualisations, such as topics-over-time graphs, heatmaps, and topic hierarchies. For keyword extraction, we used *TextRank* (Mihalcea & Tarau, 2004), a language-independent, graph-based ranking model. We then compare results returned by both methods in terms of usefulness and, finally, provide an interpretation of results by relating the extracted topics to the situation of the global pandemic at different stages of the crisis.

Bibliography

Agüero-Torales, M. M., Vilares, D., & López-Herrera, A. G. 2021. Discovering topics in Twitter about the COVID-19 outbreak in Spain. *Procesamiento Del Lenguaje Natural*, *66*, 177–190.

Boon-Itt, S. & Skunkan, Y. 2020. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *JMIR Public Health and Surveillance*, *6*(4), e21978. https://doi.org/10.2196/21978.

Chen, E., Lerman, K., & Ferrara, E. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance*, *6*(2), e19273. https://doi.org/10.2196/19273

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Mihalcea, R. & Tarau, P. 2004. TextRank: Bringing Order into Texts. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 404– 411.

Vayansky, I. & Kumar, S. A. P. 2020. A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. 2020. Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of Medical Internet Research*, *22*(11), e20550. https://doi.org/10.2196/20550.

Strategies for large social media corpora analysis: Sampling and keyword extraction methods

Antonio Moreno-Ortiz, Chantal Pérez-Hernández & María García-Gámez – *University of Málaga*

Keywords: *Covid-19 language, large-scale social media corpus, sampling methods, sampling sizes, keyword extraction.*

In the context of the Covid-19 pandemic, social media platforms such as Twitter have become of great importance for users to exchange news, ideas, and perceptions. Researchers from fields such as discourse analysis and the social sciences have resorted to this type of content to explore public opinion and stance on this topic, and they have tried to gather information through the compilation of large-scale corpora that have been made available to the academic community (Banda et al., 2020; Dimitrov et al., 2020; Lamsal, 2021). Of these, the corpus created by Chen et al. (2020) stands out as the largest, both in terms of size (with over 31 billion words) and time span, as the data were collected from January 21, 2020, and the process is still ongoing.

However, the size of such corpora is both an advantage and a drawback, as it requires users to implement their own Natural Language Processing (NLP) techniques, as manual, qualitative analysis is unfeasible. The problem is that such techniques are often computationally intensive and difficult to learn, which usually becomes a limitation for researchers. Moreover, desktop corpus tools such as *Wordsmith* (Scott, 1996) and *AntConc* (Anthony, 2022), or web-based tools that allow uploading user corpora, such as *SketchEngine* (Kilgarriff et al., 2014), simply cannot handle such massive amounts of text as they do not have text-indexing capabilities (in the case of desktop applications), or do not allow uploading such large amounts of text. Therefore, it is necessary to come up with suitable methodological underpinnings, as well as specific strategies, that facilitate managing and exploring such large-scale corpora.

This study aims to provide methodological and practical cues on how to manage the contents of Chen et al. (2020)'s Covid-19 corpus. The main objective is to compare and evaluate, in terms of efficiency and efficacy, available methods to handle large-scale social media corpora. In this way, this study leverages and compares the use of different methods and approaches. First, we aim to compare the use of differing sample sizes to assess whether it is possible to achieve similar results despite the size difference, and to evaluate sampling methods such as proportional-to-size sampling (PPS) following a specific data management approach to storing the original corpus. Second, this work will examine two keyword extraction tools that have different methodological approaches to the process: the traditional method used in corpus linguistics, which employs a reference corpus to compare word frequencies using a range of different statistical measures, and graph-based techniques as developed in NLP applications. These objectives are tackled using an experimental methodology, and evaluation of results will be performed employing specific formal metrics where possible, as assessing keyword extraction performance or quality is prey to subjective interpretation (Gabrielatos, 2018). The methods and strategies discussed in this study enable valuable quantitative and qualitative analyses of an otherwise intractable mass of social media data.

Bibliography

Anthony, L. 2022. *AntConc (Version 4.0.10)*. Waseda University. https://www.laurenceanthony.net/software.

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, K., Tutubalina, E., & Chowell, G. 2020. *A large-scale COVID-19 Twitter chatter dataset for openscientific research—An international collaboration* (Version 30) [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.4065674.

Chen, E., Lerman, K. & Ferrara, E. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance*, *6*(2), e19273. https://doi.org/10.2196/19273.

Dimitrov, D., Baran, E., Fafalios, P., Yu, R., Zhu, X., Zloch, M., & Dietze, S. 2020. TweetsCOV19—A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2991–2998. https://doi.org/10.1145-/3340531.3412765.

Gabrielatos, C. 2018. Keyness analysis: Nature, metrics and techniques. In C. Taylor & A. Marchi (Eds.), *Corpus Approaches to Discourse: A Critical Review* (pp. 225–258). Routledge.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. 2014. The Sketch Engine: Ten years on. *Lexicography*, 7–36.

Lamsal, R. 2021. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*, *51*(5), 2790–2804. https://doi.org/10.1007/s10489-020-02029-z.

Scott, M. 1996. *WordSmith Tools manual*. Oxford University Press.

Management of phraseology for the construction of a corpus-based controlled natural language

Leticia Moreno-Pérez – *University of Valladolid*

Keywords: *phraseology, corpus linguistics, controlled natural language, food and drink.*

Phraseology has attracted much attention from researchers to date, mainly due to its importance in language and the difficulties it poses for linguists in terms of conceptualization, classification, and scope (Granger and Paquot, 2008; Gray and Biber, 2015, among others). The evolution of linguistic research into technological fields, such as Natural Language Processing (NLP), has led to new challenges that are still being addressed, along with others that will emerge as more sophisticated tools are developed.

The aim of this paper is to describe how the process followed within the framework of the xxxx Research Group to handle multiword expressions (MWEs) could be applied to the construction of a Controlled Natural Language (CNL) for the food and drink industry. Using the exhaustive collection of data mined from corpora in the Group's previous project – a collection of writing generators for wine, cheese, tea, dried meats, biscuits and recipes – we will explain how data could be processed in order to include it in the architecture of a CNL. For this purpose, this paper will focus on the description of the different stages suggested to systematize and process MWE- related data retrieved from a multi-layer, domain-restricted English-Spanish comparable corpus comprised of more than 1.5 million words (Rabadán, Ramón and Sanjurjo-González, 2021).

The starting point is to obtain the most prototypical MWEs for each of the fields named above, in order to find both distinctive and recurring patterns, followed by a second stage which involves tagging MWEs at 4 levels: part-of-speech, rhetorical, semantic and pragmatic. The premise behind this methodology is that, as all these layers are interwoven in discourse, the more information the system is provided with, the more accurate the resulting CNL will be: it will offer the user specific options depending on the intention or the part of the genre being drafted, all of which have been checked by expert linguists in the field, and that are the most common ones in that specialized context.

The main turning point, the one that switches from previous tools to a CNL, is to process the existing list of tagged MWEs. The challenge posed by this phase is to adapt the data retrieved from corpora to the specific needs and aim of such a tool. More specifically, a process of standardization has to be carried out, in order to determine what a MWE is in the field of food and drink promotional texts. Also, MWEs need to be classified depending on their characteristics and tags, in order to find common patterns for this field; to do so, a specific tagging is required, not only labeling MWEs as single units, but also tagging their individual components. This, again, follows the idea of including as much lexical information as possible in the CNL.

The final step, the development of the architecture of a CNL, will need to overcome the common technical issues in the field (Villavicencio, et al., 2005; Baldwin and Kim, 2010; Parmentier and Waszczuk, 2019, among others) while allowing the tool to show all these data to the user.

Bibliography

Baldwin, T. and Kim, S. N. 2010. Multiword Expressions. In N, Indurkhya and F. J, Damerau (Eds.). *Handbook of Natural Language Processing* (2nd ed., pp. 267- 292). Boca Raton, USA: CRC Press.

Granger, S. and Paquot, M. 2008. Disentangling the phraseological web. In Granger, S. and Meunier, F. (Eds.). *Phraseology. An interdisciplinary perspective.* 27-49. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Gray, B. and Biber, D. 2015. Phraseology. In D. Biber and R. Reppen (Eds.). *The Cambridge Handbook of English Corpus Linguistics* (Cambridge Handbooks in Language and Linguistics, 125-145). Cambridge: Cambridge University Press.

Parmentier, Y. and Waszczuk, J. (Eds.). 2019. Representation and parsing of multiword expressions: Current trends. (Phraseology and Multiword Expressions 3). Berlin: Language Science Press.

Rabadán, R., Ramón, N. and Sanjurjo-González, H. 2021. Pragmatic Annotation of a Domain-Restricted English-Spanish Comparable Corpus. *Bergen Language and Linguistics Studies 11(1)*, 209-23.

Villavicencio, A., Bond, F., Korhonen, A., and McCarthy, D. 2005. Editorial: Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language, 19 (4)*, 365–377.

Multiword locative adverbial constructions in Portuguese

Izabela Müller[1], Nuno Mamede[2] & Jorge Baptista[1] – *University of Algarve[1] & University of Lisboa[2]*

Keywords: *multiword expressions, idiom, adverb, lexicon-grammar, Portuguese*

This paper proposes a new syntactic-semantic class of multiword (= compound) *locative* adverbs in Portuguese, e.g., ***aqui e acolá*** lit.: 'here and there' 'somewhere/everywhere, spottly': *O Pedro mora* aqui e acolá lit.: 'Pedro lives here and there' 'in several, undefined places'; ***por montes e vales*** lit.: 'through hills and valleys' 'everywhere, cross-country': *O príncipe cavalgava* por montes e vales lit.: 'The prince rode through hills and valleys'. These compound adverbs (a.k.a. *locuções adverbiais* 'adverbial locutions' in Portuguese grammatical terminology) present several internal combinatorial constraints (M. Gross 1982, 1986b; Guimier 1996; self-reference-1). For example, (i) permutation of coordinated elements is blocked (self-reference-1): **acolá e aqui*, ?**por vales e montes*; (ii) the coordination is obligatory: º*O Pedro mora* aqui 'Pedro lives here' (unlike Athe compound, the single-word adverbs have a deictic value); º/**O príncipe cavalgava* por montes/vales; (iii) gender/number variation is blocked: **O príncipe cavalgava* por monte e vale 'through hill and valley'; (iv) insertions of determiners and/or modifiers are blocked: **pelos montes e pelos vales* 'through_the hills and the valleys', **por montes altos e vales profundos* 'through high hills and the deep valleys'. As the examples' translation shows, these adverbs' meaning is often non-compositional (i.e., idiomatic), though they still present some generic locative meaning. This class of *locative adverbs* can be formally (syntactically) defined by:

(a) constituting an adequate answer to an interrogative adverb such as *onde* 'where': *Plantámos árvores* aqui e acolá*!* '[We] planted trees here and there' Q: Onde *plantaram essas árvores?* 'Where did you plant those trees? A: Aqui e acolá 'Everywhere';
(b) and (b) being an adequate complement of locative verbal constructions, that is, verbs requiring a locative complement (Baptista & Mamede 2020), e.g. *morar* 'live/reside': *O Pedro morava* aqui 'Pedro lived here', *O Pedro atalhou* por aqui '*Pedro shortened-the-way through* here'.

This study is part of an ongoing investigation that aims to identify, collect, and provide a syntactic-semantic description of multiword adverbs in (Brazilian) Portuguese (self-reference-2) based on similar studies for European Portuguese, French, and Spanish (Palma, 2009; Laporte, 2008a, 2008b, 2018; Català *et al.,* 2020). We adopt the theoretical-methodological framework of Lexicon-Grammar, proposed by Maurice Gross (1975, 1981, 1986a, 1996) and based on the Operator Transformational Grammar of Zellig S. Harris (1976, 1982, 1991), in conjunction with the semantic criteria for classifying the adverbs proposed by Molinier & Levrier (2000). Amongst the classes of adverbials they proposed for the French adverbs ending in *-ment* '-ly', and though locative adverbs have been often mentioned in the literature (Grevisse 1993:1442-45; Bechara, 1999; Costa 2008: 44), it was found that this specific set of adverbs, the locatives, were not included, perhaps because those adverbs did not present a locative value. It is this gap that the paper aims to fulfill, concerning locative compound adverbs. Naturally, this class is not homogenous, and the paper also proposes some subclasses to distinguish several aspects, such as location precision/fuzziness, extension, and idiomaticity. A tentative indication of the semantic roles is also presented.

So far, our study has collected approximately 150 compound adverbs in Portuguese that can be considered locative (from a lexicon of +3,300 compound entries). Many of these expressions are common to European and Brazilian varieties of the language, while others, especially the most idiomatic ones, pertaining to a single variety.

References

Baptista, J., & Mamede, N. 2020. *Dicionário gramatical de verbos do português*. Faro: Editora da Universidade do Algarve.

Bechara, E. (1999). Moderna gramática da língua portuguesa. *Rio de Janeiro: Lucerna.*

Català, D., Baptista, J., Palma, C. 2020: Problèmes formels concernant la traduction des adverbes composés (espagnol/portugais). *Langue(s) & Parole* 5, pp. 67–82. https:// ddd.uab.cat/pub/languesparole/languesparole_a2020n5/languesparole_a2020n5p67.pdf

Costa, J. 2008. *O advérbio em português europeu*. Lisboa: Edições Colibri.

Grevisse, M. 1993. *Le Bon Usage* (edited by André Goose, 13th ed.). Louvaine-la-Neuve: Ducolot.

Gross, M. 1975. *Méthodes en syntaxe*. Paris: Hermann.

Gross, M. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages* 63: pp. 7-52. Paris: Larousse.

Gross, M. 1982. Une classification des phrases figées du français. *Revue québécoise de linguistique* 11:2, pp. 151-185. Montreal: Presses de l'Université du Québec à Montréal.

Gross, M. 1986a. *Grammaire transformationnelle du français: 3 - Syntaxe de l'adverbe*. Paris: ASSTRIL.

Gross, M. 1986b. Lexicon-Grammar. The representation of compound words. In Proceedings of the *11th International Conference on Computational Linguistics, COLING'86*, Bonn, West Germany, pp. 1-6. https://aclanthology.org/C86-1001.pdf

Gross, M. 1996. Lexicon-grammar. In: Brown, K., & Miller, J. (eds.) *Concise Encyclopedia of Syntactic Theories*, pp. 244–259. Cambridge: Pergamon.

Guimier, C. 1996. *Les adverbes du français: le cas des adverbes en* -ment. Paris: Editions Ophrys.

Harris, Z. S. 1976. *Notes du cours de syntaxe*. Transl. and presented by Maurice Gross. Paris: Éditions du Seuil.

Harris, Z. S. 1982. *A Grammar of English on Mathematical Principles*. Wiley-Interscience. New York; John Wiley & Sons.

Harris, Z. S. 1991. *A Theory of Language and Information - A Mathematical Approach*. Oxford: Clarendon Press.

Laporte, E., & Voyatzi, S. 2008. An electronic dictionary of French multiword adverbs. In *Language Resources and Evaluation Conference. Workshop towards a shared task for multiword expressions*, pp. 31-34. https://shs.hal.science/halshs-00286546/file/MWE AdverbLexicon.pdf

Laporte, E., Nakamura, T., & Voyatzi, S. 2008. A French corpus annotated for multiword expressions with adverbial function. In *Linguistic Annotation Workshop. Language Resources and Evaluation Conference* (LREC), pp. 48-51. https://halshs.archives-ouvertes.fr/halshs-00286541/file/CorpusAdverbialsCorrectResults.pdf

Laporte, É. 2018. Choosing features for classifying multiword expressions. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, pp. 143–186. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.1182597

Molinier, Ch., & Levrier, F. 2000. *Grammaire des adverbes: description des formes en*-ment. Genève: Librairie Droz.

Palma, C. 2009. *Estudo Contrastivo Português-Espanhol de Expressões Fixas Adverbiais*. (MA Thesis), Universidade do Algarve, Faculdade de Ciências Humanas e Sociais, Faro, Portugal, https://sapientia.ualg.pt/bitstream/10400.1/428/1/CristinaPalma2009 (TM).pdf.

Los corpus orales del español centroamericano

Danny Fernando Murillo Lanza – *University of Valencia*

Palabras clave: *corpus orales, español de Centroamérica, corpus Ameresco-Tegucigalpa, conversación coloquial.*

Los estudios que parten del análisis de corpus lingüísticos, ya sea de materiales orales, escritos o mixtos, son cada vez más frecuentes. Esto se debe al esfuerzo que han dedicado numerosos investigadores al diseño, construcción y publicación de una diversidad de corpus. No obstante, su desarrollo no ha sido igual en todos los géneros discursivos, ni en todos los canales de producción (escritos u orales), como tampoco han sido igual de representativos en todas las variedades diatópicas del español.

En el mundo hispánico, hay trabajos, como el de Briz y Albelda (2009), Briz (2012, 2018), Solís García (2018) y Briz y Carcelén (2019); que, además, de presentar y recopilar aquellos corpus –sobre todo, los de lengua hablada u oral– del español, ofrecen una mirada crítica sobre el estado, el diseño, la construcción, el acceso y el uso de estos corpus. En términos generales, señalan que, a pesar de que el avance en la construcción de corpus del español ha sido notable, en la mayoría de los corpus más grandes lo oral ocupa poco espacio, por un lado, y los corpus situacionales y de conversaciones coloquiales son todavía pocos, por otro lado.

A esto se puede añadir que no todas las variedades o normas regionales del español se encuentran representadas de igual forma, ni en cantidad ni en calidad. Para el caso, en el contexto regional, la construcción y el acceso a corpus que recojan muestras escritas y, sobre todo, muestras orales que den cuenta de las diversas variedades y normas lingüísticas del español de Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica y Panamá (el español centroamericano, el cual cuenta con más de 44 millones de hablantes) ha sido mínimo y discreto en comparación con otras regiones y países hispanohablantes.

Dado que reconocemos que la elaboración de cualquier corpus –sea escrito u oral– implica la inversión de mucho tiempo y esfuerzo por parte de los investigadores, hemos pensado que sería justo (para estos investigadores) y beneficioso para la lingüística del español centroamericano poder dar cuenta de los principales corpus orales que se han construido. Nos centramos en los corpus orales, pues son los que menos atención reciben (Briz y Albelda, 2009) y los que, a su vez, suponen una mayor dificultad en su diseño y elaboración (Briz, 2012: 124).

En consecuencia, esta comunicación tiene como objetivo, por un lado, recopilar y presentar cuáles son los principales corpus orales del español centroamericano; y, por otro, reflexionar sobre sus fortalezas y debilidades. Esta recopilación ha seguido los criterios de recolección ya empleados por Briz y Albelda (2009) y los ha adaptado a sus objetivos. Cabe destacar que la recopilación de los corpus ha sido localizada gracias a trabajos como el de Briz y Albelda (2009) y el de Briz y Carcelén (2019). No obstante, esta comunicación aspira a completar la información ofrecida en estos trabajos y pretende dar algunos datos importantes sobre los corpus orales centroamericanos, tales como la persona y la institución que se encargó de la recolección del microcorpus, la cantidad de material recogido, el período en el que fueron grabadas, la cantidad y las características de los hablantes que participan, entre otros.

Los resultados preliminares indican que la mayoría de los corpus orales del español centroamericano que se han diseñado y construido forman parte de macroproyectos como el PRESEEA, Ameresco, el Proyecto de la Norma Culta «Juan M. Lope Blanch», entre otros.

Referencias bibliográficas

Briz Gómez, A. 2012. «Los déficits de los corpus orales del español (y de algunos análisis)», en T. Jiménez Juliá, B. López Meirama, V. Vázquez Rozas y A. Veiga Rodríguez (coords.), *Cum corde et in nova grammatica. Estudios ofrecidos a Guillermo Rojo*. Santiago de Compostela: Servizo de Publicacións da Universidade de Santiago de Compostela, pp. 115-137.

Briz Gómez, A. 2018. «Los corpus de conversaciones coloquiales. La elaboración del corpus *AMERESCO* (Español coloquial de América y España)». Ponencia de clausura en el I Col·loqui Internacional de Lingüística de Corpus (LingCor 2018), Valencia, del 13 al 14 de diciembre de 2018.

Briz Gómez, A. y Albelda Marco, M. 2009. «Estado actual de los corpus de lengua hablada y escrita: I+D», en *El español en el mundo. Anuario del Instituto Cervantes 2009*. Madrid: Instituto Cervantes y AEBOE, pp. 165-225.

Briz, A. y Carcelén, A. 2019. «El futuro iberoamericano del español: la investigación del español oral y en español». En *Anuario del Instituto Cervantes*. Instituto Cervantes.

Solís García, I. 2018. «Corpus españoles dialógicos para el análisis de la conversación», CHIMERA. *Romance Corpora and Linguistic Studies*, 5 (1), pp. 117-129.

The use of fillers among instructors of Spanish as a second language

Oihane Muxika-Loitzate[1], Nausica Marcos-Miguel[2] & Silvia Aguinaga-Echeverría[1]

University of Navarra[1] – Denison University[2]

Keywords: *fillers, teacher-talk, Second Language Acquisition, a ver*, *vale.*

Fillers play an important role in communication by expressing pragmatic nuances, but second language instructors do not usually teach them in class (Basurto et al., 2016; Erten, 2022). Although learners may be exposed to them through teacher-talk, as there is variation on lexical richness depending on the teaching context (see Horst, 2009), it is not clear whether learners will receive systematic exposure to fillers.

This study analyzes a bilingual corpus of 12 online Spanish lessons taught at a US university to determine which fillers instructors use and in which classroom modes, i.e., focusing on classroom management, materials, skills and systems, or learners' oral fluency (see Walsh, 2006). Specifically, the fillers *a ver* and *vale* are examined.

The research questions are: (a) do instructors use *a ver* and *vale* in their Spanish classes?, (b) if they do, do they show variability in their uses across and within instructors? and (c) in which classroom modes do these fillers appear and which pedagogical goals do they pursue?

We analyzed data from three instructors from Argentina (T1), Spain (T2), and Romania (T3). They recorded themselves in March-April 2020. The transcriptions included 375 tokens of *vale* and 69 of *a ver*. Results show variability in their use as only T1 and T2 used *vale*, whereas the three instructors used *a ver* (see Table 1). Instructors T1 and T2 did not only use these fillers when speaking Spanish, but also when speaking English, as in (1):

(1) *If you see somebody in the waiting room, you let me know,* ***¿vale?***

	Vale		***A ver***	
	with sentences in Spanish	with sentences in English	with sentences in Spanish	with sentences in English
T1	**69%** (n=100)	**31%** (n=44)	**97.30%** (n=36)	**2.7%** (n=1)
T2	**81%** (n=188)	**19%** (n=43)	**96.15%** (n=25)	**3.85%** (n=1)
T3	0	0	**100%** (n=6)	**0%** (n=0)

Table 1. Occurrences of *vale* and *a ver* in Spanish and English sentences by instructor.

T2's use of *vale* was expected, given that this filler has been reported in varieties of peninsular Spanish (Basurto et al. 2016), but it was unexpected for T1, who spoke a variety of Spanish from Argentina. This could be due to T1's contact with speakers from Spain. With regards to *a ver,* all instructors used it. However, T3 used it with the literal meaning "to see" and never as a filler, i.e., she never used any of the fillers under analysis. Finally, the results showed that T1 and T2 used their fillers mostly in managerial mode, as they used *vale* and *a ver* after or while giving instructions related to the course or its assignments.

In brief, there was variation among instructors: whereas T1 and T2 showed similarities in their use of the two fillers, T3 deviated from the other two. Thus, this data suggests some effects on production of fillers by advanced users of Spanish compared to native speakers. Additionally, given that the classes were online, our analysis shows use of fillers for pedagogical goals, such as *looking for something on the screen* or *explaining information related to the videoconferencing software,* which could help future analyses of teacher-talk in digital platforms, i.e., in a

digital managerial mode. Teaching of fillers might still be necessary as exposure is not systematic.

Bibliographic references

Basurto Santos, N. M., Hernández Alarcón, M. M., & Mora Pablo, I. 2016. Fillers and the development of oral strategic competence in foreign language learning. *Porta Linguarum*, 25: 191-201. [http://hdl.handle.net/10481/53916]

Erten, S. 2014. Teaching fillers and students' filler usage: A study conducted at ESOGU preparation school. *International Journal of Teaching and Education,* 2(3), 67.

Horst, M. 2009. 5. Revisiting Classrooms as Lexical Environments. In T. Fitzpatrick, & A. Barfield (Eds.). *Lexical processing in second language learners* (pp. 53-66). Multilingual Matters.

Walsh, S. 2006. Investigating classroom discourse. Routledge.

Do learners acquire the function of the English passive along with its form? Case study of Armenian learners

Emma Nemishalyan – *University of Santiago de Compostela*

Keywords: *corpus linguistics, pragmatics, passive voice, second language acquisition.*

It has been long argued that learners' mother tongue (L1) influences the second language (L2) acquisition and production; some have shown that L1 has a huge impact on their L2, some have debunked the "myth" that L1 influences the acquisition and the production of L2. If there are discrepancies in terms of the impact of L1 on L2, there is none in regard to the acquisition of form and function: it is believed that even advanced learners of a language acquire the form more easily than its function (Carroll et al.,2000); Granger et al., 2002; Stutterheim and Lambert, 2005). However, so far, no research has been carried out on the Armenian learners of English with the aim to discover whether the aforementioned theories are applicable for them too. The aim of this research is to fill this gap and understand whether Armenian learners' pattern of the use of the passive voice is similar to the Native ones' pragmatically and whether their L1 has a role in their pattern. To this end, about 27.000 tokens were taken from both native speakers' (LOCNESS) and Armenian learners' corpora (compiled by me in compliance with ICLE guidelines, including 100 essays by upper-intermediate to advanced level Armenian learners) and were compared in the light of the use of the passive. Both quantitative and qualitative methods of analysis have been applied.

Firstly, with the help of Contrastive Interlanguage Analysis (hereafter CIA) (Granger, 2007) the most common to the least common purposes of the use of the passive voice have been classified in both native and Armenian learners' corpora. Secondly, based on the quantitative data, the reasons behind the use patterns of the passive voice in both corpora have been analyzed. Special emphasis has been given to the Armenian learners' choice of the passive voice, looking into the pragmatic aspect of the passive voice in Armenian to track the connection.

Taking into account the fact that the pragmatic aspect of the passive voice in Armenian is not comprehensively studied and the main function of it is to thematize the logical object of the sentence, it was hypothesized that Armenian learners would mostly use the English passive with that intention because of the influence of their L1. The results yielded from the research show that Armenian learners did use the form of the passive voice mostly without errors. Moreover, when it came to the pragmatics, the pattern was not drastically different from the native speakers' one, which comes to disprove the hypothesis that learners struggle with the acquisition of the function of grammar issues.

Bibliography

Carroll, M., Murcia-Serra, J., Matorek, M. and Bendiscioli, A. 2000. The relevance of information organization to second language acquisition studies. The descriptive discourse of advanced adult learners of German. *Studies in Second Language Acquisition* 22 (3): 441-466.

Granger, S., Dagneaux, E. L., & Meunier, F. 2002. *The International Corpus of Learner English.* Handbook and CD-ROM.

Granger, S. 2007. A Bird's-eye View of Computer Learner Corpus Research.

Stutterheim, C. V. & Lambert, M. 2005. Cross-linguistic analysis of temporal perspectives in text production.

Stutterheim, C. von. 2003. Linguistic structure and information organisation: The case of very advanced learners. In S. H. Foster-Cohen and S. Pekarek-Doehler (eds) *EUROSLA Yearbook,* 183-206. Amsterdam: John Benjamins.

Journalistic texts across languages: Specificities of rhetorical devices. What parallel corpora tell us about genre in French and in English

Raluca Nita – *University of Poitiers*

Keywords: *journalistic translation, coherence, genre, parallel corpora.*

Journalistic translation has received great interest from translation studies and communication studies scholars (Valdeón 2015a, Zanettin 2021) who have pointed out the relation between translation and news production and have described journalistic translation mainly as target text and culture oriented, making strong claims for "an extension of the term translation" (Van Doorslaer 2010). They have indeed underlined frequent cultural and linguistic transformations of the source text relative to the cultural and linguistic specificities of the target text and described the process of journalistic translation as localisation, domestication, contextualization or reframing (Bielsa et al. 2009, Bani 2006). At the same time, journalistic translation has also been an important support for contrastive linguistic studies based on parallel corpora (Chuquet & Paillard 2006, Gilquin 2006, Lansari 2006) which have taken into account newspapers as a genre but not necessarily the specificities of journalistic translations which are revealed by corpus compilation and alignment.

Adopting a linguistic view on parallel journalistic texts in French and English and taking into account the target language and culture constraints involved in journalistic texts as a genre, we will deal with frequent cases of modifications in news translations (Bani 2006, Valdeón2006, van Doorslaer 2010) like omissions and syntactic reorganization of sentences. We will point out recurrent, corpus-based linguistic patterns that accompany this type of modifications and that could suggest a balanced relation between source text and target text, contrary to the model of translation provided by translation studies scholars. Our focus will be on the way in which coherence with respect to topic and to genre (Carter Thomas 2021) is achieved both in the source and target texts within the structural changes that affect the latter as a consequence of medium constraints.

Our corpus is made up of French original texts published in *Le Monde* and translated into *The Guardian Weekly* (from 1991 to 2021, about 740.000 words) and of English originals from English newspapers around the world translated into French in *Le Courrier International* (from 1991 to 2021, about 730.000 words). The newspaper articles in the corpus are mainly news articles and features which can influence the type of translation. We are in the presence of relative "stable" sources (Valdeón 2015b) whose authors and newspapers are clearly identified in the target text which is, in its turn, presented as a translation in the target newspaper. By describing the process of corpus building and corpus alignment, we advocate the idea that journalistic texts as a genre (Biber and Conrad 2009, Bell 1991) can make translations a valid linguistic support for empirical and theoretical studies on textual coherence in contrastive linguistics.

While omissions in our corpus generally concern non essential information taking the form of parenthetical comments or quotations as illustrations of main arguments, they do modify cohesion devices and the coherence of the text from the point of view of topic presentation. We will point out the relation between cohesion markers (conjunctions, nouns) and textual coherence in the two texts.

Sentence structure changes are an effect of different textual rhetoric in the two languages and while certain changes occur under similar conditions both in literary and journalistic translations, others can be genre specific. Such is the case of non-verbal sentences in French (for instance, indefinite article + noun with anaphoric function + relative clause) which are typical of journalistic texts in French (Combettes 2007, Emmott & al. 2006, Grinshpun 2011) and provide insight into different ways of achieving information focus in French and English. The recurrent use of canonical verbal sentences as English equivalents will be regarded from a linguistic and

textual point of view as specific genre realizations. Such equivalents involve, indeed, not only syntactic transformations but also modal changes to keep up with the topic provided by the source text while adopting target text rhetoric in terms of topic organization.

References

Bani, Sara, 2006, "An Analysis of Press Translation Process", *in* K. Conway and S. Bassnett (ed.) *Translation in Global News, Proceedings of the conference held at the University of Warwick*, 23 June 2006, 35-45.

Bell, Allan, 1991, *The Language of News Media*, Blackwell.

Bielsa, Esperança and Bassnett, Susan, 2009, *Translation in Global News*, Routledge.

Biber, Douglas and Conrad, Susan, 2009, *Register, genre and style*, Cambridge University Press.

Carter-Thomas, Shirley, 2021, « Cohérence, cohésion et structures textuelles: liens et interactions avec la notion de genre », *in Cohérence et cohésion textuelles*, Lambert Lucas, 31-47.

Combettes, Bernard, 2007, « Discontinuité et cohérence discursive: le cas des ajouts après le point », *Cahiers de praxématique*, 48.

Chuquet, Hélène et Paillard, Michel, 2006, « Les adjectifs composés en X + V + *-ing*: prédication, collocations, traductions », *Palimpsestes* 19, 13-34.

Conway, Kyle and Bassnett, Susan (ed.), 2006, *Translation in Global News*, *Proceedings of the conference held at the University of Warwick, 23 June 2006*, The Centre for Translation and Comparative Cultural Studies, University of Warwick, Coventry CV4 7AL, United Kingdom.

Emmott, C., Sanford, A. J. & L. Morrow, 2006, "Sentence fragmentation. Stylistic aspects", in *Encyclopedia of Language and Linguistics*, vol. 11, K. Brown (ed.) Elsevier, Oxford, p. 241-251

Gambier, Yves, 2006, "Transformations in International News", in K. Conway and S. Bassnett (ed.), 2006, *Translation in Global News,* 9-22.

Gilquin, Gaëtanelle, 2006, « Constructions causatives en *faire faire* et *make*: qui se ressemble ne d'assemble pas toujours », *in* H. Chuquet et M. Paillard (ed.) *Causalité et contrastivité. Études de corpus*, Rennes, Presses universitaires de Rennes, 93-112.

Grinshpun, Yana, 2011, « Phrase averbale et presse écrite: le cas des constructions en [UN + N + expansion] », *in* F. Lefeuvre et I. Behr (ed.) *Les énoncés averbaux autonomes entre grammaire et discours*, Ophrys, 187-203.

Gutiérrez, Miren, 2006, "Journalism and the Language Divide", in K. Conway and S. Bassnett (dir.) *Translation in Global News*, *Proceedings of the conference held at the University of Warwick, 23 June 2006*, The Centre for Translation and Comparative Cultural Studies, University of Warwick, Coventry CV4 7AL, United Kingdom, 29-34.

Lansari, Laure, 2006, Les périphrases « aller » + inf. et « be going to » en français et en anglais contemporains, Thèse soutenue à l'Université de Poitiers.

Van Doorslaer, Luc, 2010, "The double extension of translation in the journalistic field", Across Languages and Cultures 11 (2), 175-188.

Valdeón, Roberto A., 2015a, "Fifteen years of journalistic translation research and more", *Perspectives*, 23:4, 634-662,

Valdeón, Roberto A., 2015b, "(Un)stable sources, translation and news production", *Target*, 27:3, 440-453.

Valdeón, Roberto A., 2014, From adaptation to appropriation: framing the world through news translation", *Linguaculture*, 1, 51-62453.

Zanettin, Frederico, 2021, *News Media Translation*, Cambridge University Press.

An analysis of the syntactic development of Czech texts written by non-native speakers

Michaela Nogolová, Michaela Hanušková, Radek Čech & Miroslav Kubát – *University of Ostrava*

Keywords: *syntactic development, Czech language, Slavic language.*

Syntactic analysis is an important part of second language acquisition (SLA) studies. Over the past 30 years, syntactic complexity has been regarded as a theoretical framework for analysing the syntactic development of L2 learners. However, scholars have recently focused on finding other ways to measure SLA syntactic development, particularly those that consider the dependency structure of clauses or sentences. Following this trend, we apply the Linear Dependency Segment (LDS; Mačutek et al. 2021) to SLA syntactic development. The LDS is a recently proposed unit between word and clause that reflects both the dependency structure in a clause and the linear order of the sentence. Specifically, it is defined as "the longest possible sequence of words belonging to the same clause in which all linear neighbours (i.e., words adjacent in a sentence) are also syntactic neighbours (i.e., they are connected by an edge in the syntactic dependency tree which represents the sentence)" (Mačutek et al. 2021). In this analysis, the relationships between a) the average length of clauses measured in LDS and language proficiency levels and b) the length of LDS and language proficiency levels are investigated.

The language material consists of the CzeSL-SGT learner corpus (Šebesta et al. 2014), which contains Czech texts written by non-native speakers collected between 2009 and 2013. In our study, we use 5,721 texts covering levels A1–C1 (according to the Common European Framework of Reference for Languages). For a comparison with L2 results, we also utilize the reference corpus (REF-CZ) consisting of texts written by Czech pupils and students at elementary and secondary schools (SKRIPT2012; Šebesta et al. 2013). Specifically, we use 87 texts from high school students in fourth grade. All texts are processed with UDPipe 2.0 (Straka 2018). This tool is used for parsing and morphological tagging, which is necessary for the determination of clauses, LDS, and their lengths.

The analysis is comprised of the following steps. First, we calculate the average length of a clause (ACL) in terms of the number of LDS and the average length of LDS in words. Next, we statistically test the differences between particular proficiency levels. Finally, we focus on the cross-linguistic influence of the learners' first language by comparing the results of texts written by Slavic and non-Slavic speakers at the same proficiency levels. These results are also statistically tested.

The results indicate that the higher the proficiency level, the longer the ACL and the shorter the average LDS length. As for ACL, there are statistically significant differences between all levels, except C1 and REF-CZ. In the case of average LDS length, the results reveal slower development. Particularly, statistically significant differences are not found between consecutive levels, but they occur between more distant levels (e.g., between A1 and B2, A2 and B2). Finally, non-Slavic speakers have a higher average LDS length in most cases and a shorter ACL at all levels compared to their non-Slavic counterparts. Thus, the linear dependency segment is a useful syntactic unit to study second language acquisition.

References

Mačutek, J., Čech, R., Courtin, M. 2021. The Menzerath-Altmann law in syntactic structure revisited. In: *Proceedings of the Second Workshop on Quantitative Syntax*, Association for Computational Linguistics, pp. 65–73.

Straka, M. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In: *Proceedings of CoNLL 2018*. The SIGNLL Conference on Computational Natural Language Learning, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 197–207.

Šebesta, K., Bedřichová, Z., Šormová, K., Štindlová, B., Hrdlička, M., Hrdličková, T., Hana, J., Petkevič, V., Jelínek, T., Škodová, S., Poláčková, M., Janeš, P., Lundáková, K., Skoumalová, H., Sládek, Š., Pierscieniak, P., Toufarová, D., Richter, M., Straka, M. & Rosen, A. 2014. *CzeSL-SGT: CzeSL-SGT – a corpus of non-native speakers' Czech with automatic annotation*, version 2 from 28 Sep 2014. Ústav Českého národního korpusu FF UK, Praha.

Šebesta, K., Goláňová, H., Jelínek, T., Jelínková, B., Křen, M., Letafková, J., Procházka, P., Skoumalová, H. 2013. *SKRIPT2012: akviziční korpus psané češtiny, přepisy písemných prací žáků základních a středních škol v ČR*. Ústav Českého národního korpusu FF UK, Praha.

The role of age in the use of emoji and emoticons in Twitter discourse

Paloma Núñez-Pertejo & Ignacio Palacios-Martínez – *University of Santiago de Compostela*

Keywords: *digital discourse, Twitter, emoji, emoticons, age.*

Digital communication is a rapidly evolving domain and has dramatically altered the way we communicate in our everyday interactions. The emergence of *Internet Slang* or *Netspeak* (cf. Crystal 2006) has led to what has been termed *Computer-Mediated Communication* (CMC), which is conducted through a variety of social media, including applications such as *Facebook*, *WhatsApp*, *Twitter*, *Instagram* and *TikTok*, among others (Herring 1996, 2001; Zappavigna 2012; Squires 2016).

This paper will examine the use of two particular graphic devices or *graphicons* (Herring & Dainas 2017, 2018, 2020): *emoji* and *emoticons* or *smileys* (cf. Dresner & Herring 2010; Hsiao & Hsieh 2014; An et al. 2018; Jaeger et al. 2018; Wiese & Labrenz 2021; Koch et al. 2022) in the language of Twitter users, looking specifically at the possible impact of the age factor on the selection of such devices. This area has received little scholarly attention (Nguyen et al. 2013; Flekova et al. 2016), in contrast to the study of other factors or variables here, including users' gender, race, political orientation, cultural background, and regional origins (cf. Nishimura 2015; Herring & Dainas 2018; Albawardi 2018; Miltner 2021; López-Rúa 2021, among others).

To this end, we will consider two samples of data: a set of posts extracted from the accounts of three British rappers in their mid-twenties (BackRoad Gee, Berwyn and Enny), which will be compared to a collection of tweets drawn from the accounts of three well-known London singers born in the 1950s and 1960s (Elton John, Phil Collins and Samantha Fox). The study will include an analysis of the frequencies of both emoji and emoticons, their positions in posts, the main collocates, and their interrelation with other graphic features, etc. Special attention will be paid to the pragmatic functions conveyed by these communicative devices (McCulloch 2019; Dainas & Herring 2021; Zappavigna & Logi 2021), since they can be used in very much the same way as more traditional pragmatic markers. Indeed, emoji and emoticons can be said to fulfil a variety of expressive functions, including humour and irony (Yus 2021) while also contributing to the establishment and maintenance of social relationships.

Our preliminary findings indicate that the use of emoticons (or smileys) over the last few years has greatly declined in both young and mature adults, but especially among younger users. By contrast, emoji are very frequent in the exchanges of both groups of posters, and the use of these appears to vary according to the topic in question. Final position is by far the most common in all cases, although they also occur frequently on their own, thus becoming equivalent to a complete turn. Differences between the two groups of users are also found in terms of the selection and frequency of particular emoji, as well as in their meanings and certain pragmatic functions identified.

Overall, our findings seem to suggest that age, which has consistently been found to play a significant role in off-line communication (Holmes 1992; Eckert 1998; Stenström et al. 2002; Cheshire 2006; Tagliamonte 2016), also has a role in digital platforms such as Twitter. This is observed not only at the lexical, grammatical and discourse levels, but also in the use of emoji and emoticons, as well as in other pervasive features in Twitter discourse which are directly related to the oralization of written texts (Yus 2011), such as letter/word lengthening and letter spacing, abbreviations typical of internet language, the use of capitals and asterisks, repeated punctuation marks, the replacement of words by numbers, etc.

Selected References

Albawardi, Areej. 2018. The translingual digital practices of Saudi females on WhatsApp. *Discourse, Context & Media* 25: 68-77.

An, Jiaxin, Tian Li, Yifei Teng & Pengyi Zhang. 2018. Factors influencing emoji usage in smartphone mediated communications. In Gobinda Chowdhury, Julie McLeod, Val Gillet & Peter Willett (eds.). *Transforming Digital Words*. Springer. 423-428.

Dainas, Ashley R. & Susan C. Herring. 2021. Interpreting emoji pragmatics. In Chaoqun Xie, Francisco Yus & Hartmut Haberland (eds.). *Approaches to Internet Pragmatics: Theory and Practice*. Amsterdam/Philadelphia: John Benjamins. 107-144.

Herring, Susan C. & Ashley R. Dainas. 2017. "Nice picture comment!" Graphicons in Facebook comment threads. *Proceedings of the Hawaii International Conference on System Sciences* 50: 2185-2194.

Herring, Susan C. & Ashley R. Dainas. 2018. Receiver interpretations of emoji functions: a gender perspective. In Sanjaya Wijeratne, Emre Kiciman, Horacio Saggion & Amit.

P. Sheth (eds.). Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media (Emoji2018). Stanford, CA, USA.

Herring, Susan C. & Ashley R. Dainas. 2020. Gender and age influences on interpretation of emoji functions. *ACM Transactions on Social Computing* 3(2), article 10.

Hsiao, Kun-An & Pei-Ling Hsieh. 2014. Age difference in recognition of emoticons. In Sakae Yamamoto (ed.). *Human Interface and the Management of Information. Information and Knowledge in Applications and Services*. Part II. Springer. 394-403. Koch, Timo K., Peter Romero & Clemens Stachl. 2022. Age and gender in language, emoji, and emoticon usage in instant messages. *Computers in Human Behavior* 126.106990.

Jaeger, Sara R., Yisun Xia, Pui-Yee Lee, Denise C. Hunter, Michelle K. Beresford & Gastón Ares. 2018. Emoji questionnaires can be used with a range of population segments: Findings relating to age, gender and frequency of emoji/emoticon use. *Food Quality and Preference* 68: 397-410.

López Rúa, Paula. 2021. Men 😊 and women ♥ on Twitter: A preliminary account of British emoji usage in terms of preferred topics and gender-related habits. *Language@internet* 19, article 3.

Wiese, Heike & Annika Labrenz. 2021. Emoji as graphic discourse markers. In Daniël Van Olmen & Jolanta Šinkūnienė (eds.). *Pragmatic Markers and Peripheries*. Amsterdam/Philadelphia: John Benjamins. 277-300.

Corpus paralelos español-asturiano para el entrenamiento de sistemas de traducción automática neuronal

Antoni Oliver[1], Cristina Valdés[2] & Víctor Suárez[1] – *Universitat Oberta de Catalunya[1] – University of Oviedo[2]*

En este trabajo se presenta la tarea de compilación de corpus paralelos español-asturiano en el marco del proyecto TAN-IBE: Traducción automática neuronal para las lenguas románicas de la península Ibérica. Este proyecto tiene como objetivo el entrenamiento de sistemas de traducción automática neuronal entre las siguientes lenguas románicas: español, portugués, catalán, gallego, asturiano, aragonés y aranés. El proyecto pone una especial atención a las tres últimas lenguas de la lista anterior, ya que son las que disponen menos corpus paralelos.

En primer lugar, ofrecemos un análisis de los corpus disponibles en la colección Opus Corpus para el par español-asturiano. En el momento de escribir este artículo había un total de 10 corpus paralelos disponibles, sumando un total de 7.2 millones de segmentos. La mayoría de ellos de un tamaño muy reducido, a excepción del corpus CCMatrix, que cuenta con un total de 6.3 millones de segmentos. Una primera inspección visual de este corpus ya nos confirmó que, de entre estos segmentos, muchos presentaban problemas: los segmentos de la lengua de llegada están en lenguas diferentes del asturiano, y/o los segmentos de partida y llegada no son realmente equivalentes de traducción. Para detectar los segmentos problemáticos se ha desarrollado un algoritmo que verifica las lenguas de los segmentos y ofrece un índice de confianza respecto a que el segmento de la lengua de llegada sea el equivalente de traducción del segmento en la lengua de partida. El algoritmo utiliza un detector de lenguas y la representación de los segmentos de partida y de llegada mediante embeddings a partir de un modelo multilingüe. Esto nos permite calcular la distancia entre estos embeddings para determinar si son equivalentes de traducción. Este algoritmo se ha implementado en una herramienta, *MTUOC-PCorpus-rescorer,* que se puede obtener y utilizar libremente.

Posteriormente, describimos el trabajo de obtención de textos paralelos y su posterior alineación automática para crear los corpus paralelos. Para esta tarea se evalúan dos estrategias de alineación automática: una más clásica basada en Hunalign (Varga et al., 2005) y una nueva implementación basada de nuevo en *embeddings* de oraciones utilizando un modelo multilingüe. Estas dos técnicas se han implementado en la herramienta *MTUOC-aligner* y en este artículo se presenta la evaluación comparativa de oraciones permite, además de la búsqueda de segmentos traducidos en corpus comparables. En este artículo se presenta el uso y evaluación de esta técnica.

Reactions in the UK to Johnson's and Truss's resignations

Aroa Orrequia-Barea – *University of Cádiz*

Keywords: *Johnson, Truss, Twitter, newspapers, headlines, corpus linguistics.*

The United Kingdom has gone through a period of political chaos during 2022. It all started in July when Boris Johnson, Prime Minister at the moment, decided to resign after several scandals, such as breaking lockdown rules and attending parties; and the economic crisis the country was facing. Afterwards, Liz Truss won the race in the leadership contest to replace Johnson as Prime Minister and was appointed on the 6th of September. However, her economic reforms led to cuts, inflation and the cost-of-living crisis which made her resign just forty-five days later, on the 20th of October.

This study aims to analyse the British's reactions to these events in microtexts belonging to two different genres: news headlines and Twitter. On the one hand, two corpora of headlines have been compiled, one including 1,649 news headlines related to Johnson and the other one containing 1,339 headlines referring to Truss. They belong to the news published the day they resigned and the day after. On the other hand, two corpora of tweets have been compiled over the same period of time. The criterion was that they include hashtags mentioning both politicians' names and/or surnames (i.e., #truss, #LizTruss, #liztruss; #johnson, #BorisJohnson, #borisjohnson). Truss's corpus consists of 245 tweets, whereas Johnson's corpus amounts to 174 tweets.

The texts have been analysed using the methodology of Corpus-assisted Discourse Studies (Partington & Haarman 2004; Baker 2006) which combines both Corpus Linguistics and Discourse Studies methods and helps the researcher to analyse the data both quantitatively and qualitatively. Sketch Engine has been used to analyse the text using corpus linguistics techniques, such as keywords, wordlists and collocations (Kilgariff et al. 2014). Additionally, the concordance function has been used for close reading and to interpret and make sense of the patterns found from a (critical) discourse studies perspective to uncover "non-obvious meanings" (Partintong, Duguid & Taylor 2013).

Despite Johnson having done something unethical, Truss's resignation has been more high-profile than Johnson's. This can be seen in the 'Liz Truss Lettuce' campaign started by the *Daily Star* to see whether a lettuce would outlast Truss's premiership. The campaign was followed by most of the country (#LizVsLettuce on Twitter) as the tabloid streamed it online placing a photograph of Truss next to the lettuce, which was characterised with a blonde wig.

Considering these facts, this paper aims to answer the following two research questions: a) Has Truss been differently treated from Johnson? If so, how and why? and b) Is there any difference in terms of genre?

Preliminary results illustrate that Truss's mandate, though short, is mainly defined not by what she actually did, as happens in Johnson's case ('beergate', 'partygate', 'brexiteer' are some of Johnson's corpus keywords), but for the lettuce episode as the keywords of the corpus ('lettuce', 'outlast') show. Additionally, although both leaders performed the same action, that is to resign, the word 'chaos' is more frequent in Truss's corpora than in Johnson's. Other interesting words referring to Truss are 'shortest-serving' and 'quitter'. The former is actually true but the latter does not retrieve any hits in Johnson's corpora, though it refers to something that he also did. The results are quite similar in both genres.

References

Baker, Paul. 2006. *Using Corpora in Discourse Analysis. Using Corpora in Discourse Analysis*. London: Continuum. https://doi.org/10.1007/9781139764377.013

Kilgarriff, Adam; Baisa, Vít; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michelfeit, Jan; Rychlý, Pavel; Suchomel, Vít. 2014. The Sketch Engine: Ten Years On. *Lexicography* 1/1, 7–36. https://doi.org/10.1007-/s40607-014-0009-9.

Partington, Alan; Duguid, Alison and Taylor, Charlotte. 2013. *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)* (Vol. 55). John Benjamins Publishing.

Partington, Alan & Haarman, Louann. 2004. *Corpora and Discourse*. Peter Lang.

Particularidades de la prosa modernista de Valle-Inclán: análisis desde la lingüística de corpus

Andrés Ortega Garrido – *Università degli Studi di Bergamo*

Palabras clave: *Valle-Inclán, modernismo, lingüística de corpus, estilística de corpus.*

La compleja trayectoria literaria de Ramón del Valle-Inclán (1866-1936) se caracteriza por la transformación sucesiva de su estética, que parte del Modernismo, pasa por una fase de transición donde se agudizan ciertas particularidades de su estilo y concluye con la deriva carnavalesca que supone el esperpento, una ética y una estética que constituye la personal visión del mundo por parte del autor gallego en sus últimos años. Esta evolución se desarrolla de manera especialmente relevante en el plano estético, con una renovación progresiva del estilo y de la expresión meramente lingüística, como ha estudiado tradicionalmente la crítica (Montolío Durán 1992a, 1992b; Abad Nebot, Peces Gómez 1995).

Para el presente trabajo hemos tomado en consideración los textos en prosa pertenecientes a esa primera etapa creativa de Valle-Inclán, en concreto los libros puramente anclados en el modernismo literario (*Femeninas*, *Epitalamio*, *Corte de amor*, *Jardín umbrío*, *Flor de santidad* y las cuatro *Sonatas*), publicados entre 1895 y 1908, y las tres novelas pertenecientes al ciclo *La guerra carlista* (*Los cruzados de la causa*, *El resplandor de la hoguera* y *Gerifaltes de antaño*), que vieron la luz entre 1908 y 1909 y oscilan entre el Modernismo y un estilo más depurado que anuncia un cambio de estética. El objetivo de nuestra investigación consiste en un análisis del léxico presente en estas obras para establecer la existencia de formas de expresión recurrentes que puedan ayudar a distinguir y clasificar de manera concreta y exacta las características propias del estilo valleinclaniano en esta primera etapa de su producción en prosa. En concreto, analizamos frecuencias generales de uso de las principales categorías gramaticales dotadas de contenido semántico, así como n-gramas y palabras clave. Además, profundizamos en determinadas agrupaciones léxicas en torno a verbos, sustantivos y adjetivos, tanto aquellos que presentan una alta frecuencia como los que han sido identificados como palabras clave; de este modo, intentaremos discernir en estos textos preferencias de uso con una cierta carga semántica.

Como marco teórico y metodológico nos movemos en el ámbito de la lingüística de corpus, más en concreto de la estilítica de corpus (Semino, Short 2004; Stubbs 2005; Mahlberg 2013, 2014, 2016. Para el caso del español, véase Piccioni 2015; Nieto Caballero 2018; Nieto Caballero, Ruano San Segundo 2020, Chierichetti 2022). Para la realización del estudio hemos compilado un corpus de textos a través de ediciones digitalizadas de las obras citadas, generalmente correspondientes a las ediciones últimas publicadas en vida del autor. Como herramienta informática, nos valemos de Sketch Engine, que nos permite realizar un óptimo análisis de frecuencias coherente con los objetivos de investigación prefijados.

El análisis del corpus mediante herramientas informáticas hace posible establecer una preponderancia de elementos léxicos referidos al campo semántico de los sentidos y de las sensaciones, lo cual corresponde a lo que tradicionalmente se ha señalado respecto al Modernismo literario. Con todo, no solamente es reseñable la abundancia de voces y agrupaciones léxicas orientadas hacia la expresión de la experiencia sensorial, sino también en referencia a campos semánticos relativos a la localización física (perceptible sobre todo mediante la observación de los n-gramas) y a la descripción de ciertos espacios y momentos del día, todo ello asociado a determinados ámbitos semánticos que refuerzan la propia experiencia sensorial.

Bibliografía

Abad Nebot, F.; 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), Barcelona, 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 & M. Mahlberg (eds.), *The Corpus Linguistics Discourse: In honour of Wolfgang Teubert* (Studies in Corpus Linguistics; vol. 87), John Benjamins Publishing Company, 77-94.

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. Granada, Comares.

Mahlberg, M. 2013. *Corpus stylistics and Dicken's Fiction*. Nueva York, Routledge, 5-25. Mahlberg, M. 2014. "Corpus stylistics", en M. Burke (ed.) 2014, *The Routledge Handbook of Stylistics*, Londres/Nueva York, Routledge, 387-392.

Mahlberg, M. 2016. "Corpus stylistics", en V. Sotirova (ed.) 2016, *The Bloomsbury Companion to Stylistics*, Londres/Nueva York, Bloomsbury, 139-156.

Montolío Durán, E. 1992a. Gramática en la caracterización de Valle-Inclán. Análisis sintáctico, pragmático y textual de algunos mecanismos de caracterización. Barcelona, Promociones y Publicaciones Universitarias.

Montolío Durán, E. 1992b. "La conciencia lingüística de Valle Inclán: la voluntad de renovar la lengua literaria", en *Actas del II Congreso Internacional de Historia de la Lengua Española. Tomo II*, Madrid, 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., Ruano San Segundo P. 2020. Estilística de corpus: nuevos enfoques en el análisis de textos literarios. Berlín, Peter Lang.

Piccioni, S. 2015. Lingüística de corpus y literatura. Aproximaciones cuantitativas al análisis del estilo. Saarbrücken, Editorial Académica Española.

Scott, M. 1997. "PC Analysis of Key Words – And Key Key Words". *System*, 25 (2), 233- 245.

Semino, E., Short, M. 2004. Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Narratives. London, Routledge.

Stubbs, M. 2005. "Conrad in the computer: examples of quantitative stylistics methods". *Language and Literature*, 14 (1), 5-24.

Social actors in Venezuelan presidential tweets: A corpus-assisted critical discourse study

Silvia Peterssen – *Autonomous University of Madrid*

Keywords: *polarisation; social actor analysis; corpus-assisted critical discourse studies; Venezuela.*

Social media networks, such as Twitter, Facebook, and Instagram, have become productive discourse spaces for politicians around the globe, who use them to disseminate their polarising views, ideologies, and political agendas (Baider & Constantinou, 2014; Masroor et al., 2019). Considering this issue, and taking into account the socio-economic, political and institutional crisis that Venezuela has been facing, this study looks at the discursive construction of polarisation in a corpus of tweets of two Venezuelan leaders, namely, Nicolás Maduro, president of Venezuela, and Juan Guaidó, self-proclaimed interim president of Venezuela. When Juan Guaidó declared himself interim president of the country in January 2019, despite the victory obtained by Nicolás Maduro in the presidential elections held in May 2018, a conflict known as the Venezuelan Presidential Crisis began. Despite its impact on the Venezuelan national politics and international relations, there are few studies that have analysed this crisis from a critical corpus-assisted or corpus-based discursive approach (Baker & McEnery, 2015; Charteris-Black, 2004; Taylor & Marchi, 2018). Hence, the goal of this paper is to explore their Twitter polarising narratives through this perspective, particularly focusing on the first year of the crisis (2019-2020).

Polarisation is understood as a socio-cognitive discursive phenomenon grounded on the division between 'Us' (i.e., the ingroup) and 'Them' (i.e., the outgroup) that lies at the core of ideologies and social identities (Hogg, 2016; Oktar, 2001; Van Dijk, 1998). This study investigates polarisation in Maduro and Guaidó's tweets through the analysis of their social actor representations (Darics & Koller, 2019; Krendel et al., 2022). The tweets were collected using the web-scraping software Octoparse (https://www.octoparse.es), and then compiled and processed with Sketch Engine (Kilgarriff et al., 2014). Social actors were extracted from the most frequent nouns and keywords of the corpus and analysed through collocation and concordance analyses. More specifically, the modifiers, nouns, and pronominal possessors that collocated with these actors were examined. In addition, following Van Leeuwen's social actor analysis (2008), the roles of the actors were annotated, and processes of personalisation and impersonalisation were looked at. Preliminary results suggest (i) the significative role of the Venezuelan people in Maduro's tweets vs. that of Venezuela and the opposition in Guaidó's; (ii) the use of the pronominal possessor "our" in both narratives to positively characterise ingroups; (iii) the activeness of the outgroup vs. the passiveness of the ingroup, hence reinforcing the dynamics of ingroup victimisation and outgroup blaming; and (iv) the collectivisation and objectivation of the ingroup social actors vs. the individualisation and negative labelling of the outgroup. Overall, this study points to the strategic and polarising role that positive ingroup and negative outgroup presentations of social actors have in the Twitter narratives of Nicolás Maduro and Juan Guaidó and moves the field forward by applying corpus methods to the investigation of polarisation and social actor representations.

References

Baider, F. H. & Constantinou, M. 2014. Language of Cyber-Politics: "Imaging/Imagining" Communities. *Lodz Papers in Pragmatics*, *10*(2). https://doi.org/10.1515/lpp-2014-0012

Baker, P. & McEnery, T. (Eds.). 2015. *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Palgrave Macmillan. https://doi.org/10.1057/9781137431738

Charteris-Black, J. 2004. *Corpus Approaches to Critical Metaphor Analysis*. Palgrave Macmillan. https://doi.org-/10.1057/9780230000612

Darics, E. & Koller, V. 2019. Social Actors "to Go": An Analytical Toolkit to Explore Agency in Business Discourse and Communication. *Business and Professional Communication Quarterly*, *82*(2), 214–238. https://doi.org/10.1177/2329490619828367

Hogg, M. A. 2016. Social Identity Theory. In S. McKeown, R. Haji, & N. Ferguson (Eds.), *Understanding Peace and Conflict Through Social Identity Theory: Contemporary Global Perspectives* (pp. 3–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-29869-6_1

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

Krendel, A., McGlashan, M., & Koller, V. 2022. The Representation of Gendered Social Actors Across Five Manosphere Communities on Reddit. *Corpora*, *17*(2), (1-25). https://eprints.lancs.ac.uk/id/eprint/155332-/5/The_representation_of_gendered_social_actors_across_five_manosphere_communities_on_Reddit_clean.pdf

Masroor, F., Khan, Q. N., Aib, I., & Ali, Z. 2019. Polarization and Ideological Weaving in Twitter Discourse of Politicians. *Social Media and Society*, *5*(4). https://doi.org/10.1177/2056305119891220

Oktar, L. 2001. The Ideological Organization of Representational Processes in the Presentation of us and them. *Discourse \& Society*, *12*(3), 313–346. https://doi.org/10.1177/0957926501012003003

Taylor, C. & Marchi, A. (Eds.). 2018. *Corpus Approaches to Discourse. A Critical Review*. Routledge. https://www.routledge.com/Corpus-Approaches-to-Discourse-A-Critical-Review/Taylor-Marchi/p/book/9781138895805

Van Dijk, T. A. 1998. *Ideology: A multidisciplinary approach*. SAGE. https://doi.org/10.4135/9781446217856

Van Leeuwen, T. 2008. Discourse and Practice: New Tools for Critical Discourse Analysis. Oxford University Press.

"This is an extortion note" – A corpus-driven genre analysis of commercial extortion letters

Marton Petyko, Lucia Busso, Sarah Atkins, Emily Chiang, Nabanita Basu & Tim Grant – *Aston University*

Keywords: *extortion letter, move analysis, genre, forensic linguistics.*

Malicious communications, of which extortion letters are a type, are a key area of study for forensic linguistics with many investigative applications (Nini, 2017). However, the question of whether extortion letters exhibit sufficient regularity to constitute a 'genre' has remained vastly underexplored. Some studies have identified that the form or 'text-type' of extortion letters is only minimally shaped by norms, with writers borrowing from other genres such as business letters but with the texts open to a large degree of individual variation (Fobbe, 2020). Nevertheless, such studies are often necessarily limited to small datasets, often single case studies. We present here a corpus-driven analysis of a comparatively large series of 39 commercial extortion letters and emails from historic cases in the UK (2008-19), called the *Excrow* corpus (Extortion CoRpus Of Writings) (ca 9200 words).

Using Swales' (1990) *move analysis,* we explore whether conventional discourse structures of a genre (i.e., moves) can be identified in extortion letters. We further combine findings from this qualitative genre analysis with quantitative evidence from corpus linguistics and machine learning.

The paper is innovative in two main respects. Firstly, we develop a reliable and replicable bottom-up method for corpus-based move analysis, taking clauses as basic units of analysis. Using this methodology, the research team developed a code set of eleven key moves, some of which we define as core functions for the genre (>90% of texts), while others may be typical (>50% of texts) or atypical (< 50% of texts). These moves are opening, sign-off, demand, instructions, justification, threat, demonstrating credibility, consequences, additional persuasion, statement of purpose, and pre-announcement. The letters were annotated using this code set and inter-rater reliability analysis among coders was performed (Chiang, 2018; Rau and Shih 2021). This is – to the best of our knowledge – the first work to employ such a methodology. Secondly, we use our coded set of moves to address two further research questions: (1) Are there consistent sequential patterns of moves in Excrow? (2) Do letters reliably cluster based on moves frequency?

We use sequence analysis (Gabadinho et al., 2011) and n-gram analysis to address the first question, and clustering algorithms for the second. Results indicate a high degree of variability in move sequences, and no obvious recurring move patterns. However, we were able to cluster the letters in coherent and stable groups based on move prevalence.

References

Chiang, E. 2018. Rhetorical moves and identity performance in online child sexual abuse interactions. PhD Thesis, Aston University.

Fobbe, E. 2020. Text-linguistic analysis in forensic authorship attribution. Journal of Language and Law, 9, 93-114.

Gabadinho, A., Ritschard, G., Mueller, N. S., & Studer, M. 2011. Analyzing and visualizing state sequences in R with TraMineR. Journal of statistical software, 40(4), 1-37.

Nini, A. 2017. Register variation in malicious forensic texts. International Journal of Speech, Language and the Law, 24(1), 1-35.

Rau, G. & Shih, Y. S. 2021. Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. Journal of English for academic purposes, 53, 101026.

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

Avoiding biases and assuring representativeness during the compilation of the social media section of a corpus on pseudoscientific discourse

Luis Puente-Castelo – *University of A Coruña*

Keywords: *corpus compilation, pseudoscientific discourse, social media corpora.*

From resistance to vaccination, to climate change denialism and alternative accounts of historical events, the spread of pseudoscientific ideas, "which purport to offer alternative accounts to those of science or claim to explain what science cannot explain" (Grove 1985: 219), has become a major problem in our society.

As one of the main characteristics of pseudoscientific discourse is its attempt to supersede or pass as genuine science discourse, it is of particular interest to describe this discourse from a linguistic point of view, and thus, possibly, identify particularly characteristic linguistic uses.

In order to do so, a new corpus of pseudoscientific discourse is currently being compiled. This corpus is a *c.* one million-word corpus trying to represent pseudoscience as a whole, and consequently presents texts belonging to a number of different "pseudo-disciplines" (homeopathy, antivaccination, climate change denialism, Flat Earth movement, creationism…) as well as different genres, going beyond articles, books and position papers to also include web publications, comments on web chats and forums, and contributions found in social media, trying to present as wide of a snapshot of the phenomenon as possible.

Besides presenting the current state and latest developments in the design and compilation of this corpus, this paper addresses one particular methodological problem which appeared during the process of compilation of the social media section of this corpus.

Initially, the search for examples of pseudoscientific discourse in social media was being done by means of a series of pre-established keywords, chosen for their high use in the corresponding pseudo-discipline, whose results were then manually checked. This assured all of the examples of text in the corpus were real instances of pseudoscientific writing, but, while analysing a first batch of these results, a glaring problem became obvious: the results were biased, as the very keywords which were being used to locate the examples were being over-represented, as they were present in all of the samples.

In order to avoid this problem, a new method has been devised. It involves using those keywords to locate individual user profiles rather than instances of texts, and then searching within those individual user profiles for further examples of pseudoscientific discourse, not necessarily containing these keywords. By using this method, we can obtain a collection of texts which can still be representative while also avoiding these biases.

The paper will also address some of the other measures taken to ensure representativeness and balance, such as limiting the number of samples per author to just one so as to avoid idiosyncrasies, discarding translations, and compiling samples taking into account parameters such as the sex of the author or geographical variety, so that they don't influence the variety of the results (Moskowich, 2021).

Bibliography

Grove, J. W. 1985. Rationality at Risk: Science against Pseudoscience. *Minerva, 23*: 216–240.

Moskowich, Isabel. 2021. The making of the Corpus of English Life Sciences Texts (CELiST), a bunch of disciplines. In Moskowich, Isabel; Lareo, Inés and Camiña Rioboó, Gonzalo (eds.), *"All families and genera": Exploring the Corpus of English Life Sciences Texts.* Amsterdam: John Benjamins. 2–19.

English-Spanish fictive dialogue vs. prefabricated orality: A study on addressee-oriented conversational markers

Rosa Rabadán-Álvarez & Camino Gutiérrez-Lanza – *University of León*

Keywords: *English-Spanish conversational markers, fictive dialogue, prefabricated orality, standardization, interference.*

Narrative texts recreate non-spontaneous conversation, i.e., "fictive dialogue" (Brumme & Espunya 2012), intending to reflect real-life orality in written form. Audiovisual scripts also present fake speech, i.e., "prefabricated orality" (Baños-Piñero & Chaume 2009), written to be delivered orally, as produced by the actors on screen. An essential feature of oral discourse is how the relationships and expectations of the participants are marked linguistically, for example, by means of conversational markers (CMs). Addressee-oriented CMs, also known as alterity markers, regulate the exchange between the speaker and the hearer and help to define their relationship, frequently asking the latter for cooperation (Borreguero Zuloaga 2015). These CMs tend to be polyfunctional items and have different procedural meanings depending on context; e. g., *bueno*, the most frequent translation of English *well* (48%), may indicate resumption, a remark, surprise or (dis)agreement.

This paper explores how addressee-oriented CMs behave in Spanish translated from English in the two registers, fictive dialogue and prefabricated orality, and how this compares to non-translated Spanish. In this paper register is understood in a Hallidayan sense meaning variation in language use depending on the function and, here, mode (Halliday & Hasan 1989). To do so, we have started from a widely accepted list of Spanish addressee-oriented CMs: *oye/oiga*, *mira/mire*, *venga (ya)/vamos*, *hombre/mujer*, and *bueno*, when it conveys "a certain kind of (dis)agreement" (Martín Zorraquino & Portolés Lázaro 1999: 4171-4190). First, we searched these CMs in two parallel corpora, P-ACTRES 2.0 for narrative and TRACEci for scripts, applying back translation, i.e., searching in the translated Spanish subcorpora. From this query, we obtain their translation frequency and English triggers, i.e., *look*, *well*, *OK*, *hey*, *come on*, etc. Next, the CMs were looked up in the corresponding subcorpora of CORPES XXI, the RAE reference corpus of non-translated Spanish, which features both narrative and scripts in its Fiction subcorpus.

Results show that (i) in non-translated Spanish, all CMs are more widely used in the scripts than in narrative, except for *venga ya,* for which the difference is not statistically significant; (ii) in the case of translated Spanish, corpus data indicate that the use of CMs is, in all cases, also markedly higher in the scripts than in narrative fiction, which indicates that translations are normalized in this respect since they follow the trend in original Spanish; (iii) when contrasting translated and non-translated data, CMs are underused in translated narrative, with zero translation taking an average of 23% of all cases. However, in translated scripts, CMs are generally overused, except for *hombre/mujer*, for which the difference is not significant.

This overuse of CMs in translated as compared to non-translated scripts suggests a hypercharacterization of orality in this genre, as opposed to translated narrative, where the CMs are underused, pointing to a flattening of the oral character of the dialogue. These quantitative results are interpreted in terms of third-code features (Frawley 1984), such as standardization and interference (Toury 2012). Most translation solutions choose the most frequent -and polyfunctional- options in Spanish, e. g. *bueno,* whereas other possibilities are systematically overlooked. The consequence is the growing standardization of some items, which become central in exchange marking, while alternative options are ignored. A second reason underlying these choices seems to be interference: the dictionary equivalent of the English item is preferred, prioritizing lexical correspondence over procedural meaning.

This paper is part of a more ambitious project that explores discourse markers in translated Spanish in different genres and domains (Author & Author 2023).

References

Baños-Piñero, R. & Chaume, F. 2009. Prefabricated Orality. A Challenge in Audiovisual Translation. *inTRAlinea*. Special Issue: *The Translation of Dialects in Multimedia*. http://www.intralinea.org/archive/article/1714

Borreguero Zuloaga, M. 2015. A vueltas con los marcadores del discurso: de nuevo sobre su delimitación y sus funciones, in Lala, Letizia; Ferrari, Angela and Stojmenova, Roska (eds.). *Testualità, Fondamenti, unità, relazioni / Textualité. Fondements, unité, relations / Textualidad. Fundamentos, unidad, relaciones*, Firenze: Franco Cesati, 151-170.

Brumme, J. & Espunya, A. (eds.). 2012. *The Translation of Fictive Dialogue*. Amsterdam/New York: Brill.

Frawley, W. 1984. Prolegomenon to a theory of translation. In W. Frawley (Ed.), *Translation: Literary, Linguistic and Philosophical Perspectives* (pp. 159-175). Associated University Press.

Halliday, M. A. K. & Hasan, R. 1989. Language, context and text: Aspects of language in a social-semiotic perspective (2nd ed.). Oxford: Oxford University Press.

Martín Zorraquino, M.A. & Portolés Lázaro, J. 1999. Los marcadores del discurso. In I. Bosque & V. Demonte (Dirs.), *Gramática Descriptiva de la Lengua Española* (pp. 4051-4213). Espasa Calpe.

Author & Author. In press, 2023. Interference, explicitation, implicitation and normalization in third code Spanish: Evidence from discourse markers. *Across Languages and Cultures* 24.

Toury, G. 2012. *Descriptive Translation Studies and beyond*. John Benjamins. (Original work published 1995). https://doi.org/10.1075/btl.100.

Algunas consideraciones sobre la modificación en español: un análisis discursivo-funcional en datos del CORPES XXI

Bárbara Ribeiro-Fante – *University of Oviedo*

Palabras clave*: sintagma nominal, modificación, adjetivo, Gramática Discursivo- Funcional.*

La Gramática Discursivo-Funcional, de Hengeveld y Mackenzie (2008) (GDF), teoría base de este estudio, describe el proceso de modificación de manera diferente a la tradición lingüística y gramatical (véase Lapesa, 1974; Lujan, 1980; Penadés Martínez, 1988; Bosque 1993; Demonte 1982,1999; Allarcos Llorach, 1999; RAE, 2009). Para estos autores, la modificación adquiere diferentes matices semánticos motivados por la posición del adjetivo respecto al sustantivo. Sin embargo, para la GDF, la modificación representa una estrategia cuyo papel es suministrar información léxica sobre una variable pragmática o semántica que todavía no está plenamente especificada por el núcleo de esa variable. Según Hengeveld (2008), García Velasco y Rijkhoff (2008), Keizer (2012, 2020), Giomi (2020) y García Velasco (2022), la relación núcleo-modificador se realiza por un proceso semántico de naturaleza restrictiva, como en *un coche azul,* en el que *coche* identifica el referente a que se alude en el SN y *azul* restringe todavía más el referente, lo que hace posible identificarlo frente a otras referencias (el coche azul y no el coche rojo); o no restrictiva, como en *la nieve blanca*, en que el adjetivo solo evoca una característica inherente del nombre (la nieve es blanca en cualquier situación).

A la vista de lo anterior, el objetivo general de la presente investigación es analizar los sintagmas nominales (SSNN) del español actual con datos extraídos del Corpus del Español del Siglo XXI (CORPES XXI). Específicamente, analizamos SSNN como (i) *un libro muy bonito*, (ii) *unos trajes regionales blancos* y (iii) *en mayor o menor éxito,* a fin de explicar los fenómenos que se manifiestan en cada uno de ellos, como la intensificación del adjetivo *bonito* en (i), la doble modificación en *regionales blancos* en (ii) y la coordinación adjetiva *mayor o menor* en (iii), y cómo esas cuestiones se relacionan con el tipo de modificación restrictiva o no restrictiva. Con respecto a la recopilación de datos, nos restringimos a la modalidad escrita del español actual. Por eso, seleccionamos textos comprendidos entre los años 2017 a 2020, en que existen tipos textuales de distintos registros (formales e informales), lo que es adecuado para abarcar la mayor representatividad posible del español contemporáneo, conforme propone Rojo (2019).

De esa manera, seleccionamos solamente la variedad del español peninsular dentro del soporte *web* de tipo *blog*, en todos los temas posibles, como actualidad, ocio y vida cotidiana; arte, cultura y espectáculos; ciencia y tecnología; ciencias sociales, creencias y pensamientos; política, economía y justicia; y salud. Mediante la lectura de todos los documentos obtenidos, excluimos del análisis los adjetivos que configuran predicativos o participios, pues este trabajo se centra únicamente en la modificación léxica del sustantivo, lo que significa que consideramos validos solamente los datos de SSNN con sustantivos modificados lexicalmente por uno o más adjetivos. Los resultados parciales de este estudio, obtenidos a partir de un análisis cualitativo, muestran que el ordenamiento y disposición de los modificadores del sustantivo en la lengua española actual puede seguir motivaciones y principios semánticos y pragmáticos como los identificados en la bibliografía funcional y tipológica mencionada sobre la restricción y no restricción adjetiva. Se espera, por medio de esa descripción, obtener una clasificación más adecuada de la variedad de adjetivos existentes en español y, a la vez, por medio del uso del CORPES XXI, explicar por qué determinados tipos de modificación son prototípicos de ciertos contextos y no de otros.

Bibliografía

Alarcos Llorach, Emilio. 1999. *Gramática de la lengua española*, Madrid, Espasa Calpe.

Bosque, I. 1993 Sobre las diferencias entre los adjetivos relacionales y los calificativos. *Revista Argentina de Lingüística.* vol. 9, pp. 9-48.

Demonte, V. 1982 El falso problema de la posición del adjetivo: dos análisis semánticos. *Boletín de la Real Academia Española (BRAE).* vol. 62. Madrid: Imprenta Aguirre.

Demonte, V. 1999 El adjetivo: clases y usos. La posición del adjetivo en el sintagma nominal. In: Bosque, I.; Demonte, V. (Org.). *Gramática descriptiva de la lengua española.* Madrid: Espasa-Calpe, v. 1: Entre la oración y el discurso, p. 129-215.

García Velasco, D. 2022. Modification and context. *Open Linguistics,* vol. 8, no. 1, pp. 524-544. https://doi.org/10.1515/opli-2022-0206.

Giomi, R. 2020. Headedness and modification in Functional Discourse Grammar. *Glossa*: *a journal of general linguistics*, p.1–32. DOI: https://doi.org/10.5334/gjgl.1290.

Hengeveld, K; Mackenzie, L. 2008. Functional Discourse Grammar: a typologically- based theory of language structure. Oxford: University Press.

Hengeveld, K. 2008. Prototypical and non-prototypical noun phrases in Functional Discourse Grammar. In Rijkhoff, J.; García Velasco, D. (Ed.*). The noun phrase in functional discourse grammar.* Berlin: Mouton de Gruyter, p. 43-62.

Keizer, E. 2012. Proforms in Functional Discourse Grammar. In: *Language Sciences* 34(4), p. 400–420.

Keizer, E. 2020. The problem of non-truth-conditional, lower-level modifiers: A Functional Discourse Grammar solution. In: *English Language and Linguistics:* Cambridge University Press, pp.1–28. doi:10.1017/S1360674319-00011X.

Lapesa Melgar, R. 1975. La colocación del calificativo atributivo en español. In: *Homenaje a la memoria de Don Antonio Rodríguez-Moñino:* 1910-1970, p.329-346.

Luján, M. 1980. Sintaxis y semántica del adjetivo. Madrid: Ediciones Cátedra. Penádez Martínez, I. *Perspectivas de análisis para el estudio del adjetivo calificativo en español.* Servicio de Publicaciones de la Universidad de Cádiz, 1988.

Real Academia Española. 2022. *Corpus del español del siglo XXI* (CORPES XXI).

Real Academia Española; Asociación de Academias de la Lenguas Española. *Nueva gramática de la lengua española*. Madrid: Espasa, 2009-2011.

Rijkhoff, J. 2008. Layers, levels and contexts in Functional Discourse Grammar. In: Rijkhoff, J.; García Velasco, D. (eds.). *The noun phrase in functional discourse grammar.* Berlin: Mouton de Gruyter, pp. 1-42.

Rijkhoff, J.; García Velasco, D. 2008. Introduction. In Rijkhoff, J.; García Velasco, D. (eds.). *The noun phrase in functional discourse grammar.* Berlin: Mouton de Gruyter, p. 1-42.

Rojo, G. 2021. Introducción a la Lingüística de Corpus en español. New York: Routledge.

Introducing a language-universal operationalization of syntactic interference/normalization in translation using comparable corpora

Matt Riemland – *Dublin City University*

Keywords: *syntax*, *interference*, *normalization*, *corpus-based translation studies.*

This presentation introduces a language-universal methodology for empirically measuring syntactic interference/normalization in translation using comparable corpora. Interference is the influence of source-language (SL) features in translation. Conversely, normalization is an exaggeration of target-language (TL) features in translation. By operationalizing and quantifying these concepts, linguistic features of translated texts may therefore be located on a spectrum spanning between the polarities of interference and normalization.

This novel methodology measures interference/normalization on the syntactic level. Syntax is fundamentally composed of parts of speech (POS). Sequences of POS may be referred to as POS n-grams, where three consecutive POS constitute a 3-gram, four consecutive POS constitute a 4-gram, and so on. For multilingual corpora using different language-specific POS tagsets, tagsets may be standardized to a universal POS tagset. Relative frequencies (RFs) may then be calculated for each POS n-gram observed. For any given POS n-gram (e.g. adjective- adjective-noun), its respective RF in all three relevant corpora – the comparable SL copus, the comparable TL corpus, and the translated text(s) – may be calculated.

An observed POS n-gram's RF in the comparable TL corpus may serve as the expected value of the same POS n-gram's corresponding RF in the translated text(s), since translated texts must be grammatical in the TL. The observed value's (meaning the POS n-gram's RF in the translation corpus) deviation from this expected value in the direction of the POS n-gram's corresponding RF in the comparable SL corpus constitutes interference. Deviation from this expected value in the opposite direction constitutes normalization (i.e. an exaggeration of the TL syntax). For each POS n-gram, syntactic interference/normalization may therefore be expressed as a percentage, where the distance between the observed value and the expected value is divided by the distance between the n-gram's corresponding values in the comparable SL corpus and the comparable TL corpus. A weighted average of the syntactic interference/normalization scores of all POS n-grams in the translated text(s) may then express the overall measure of syntactic interference/normalization in translation. This presentation offers visual representation of this calculation to clarify its underlying concept. It further illustrates three hypothetical scenarios of syntactic interference/normalization in translation relative to comparable SL and TL corpora.

Previous studies on syntactic interference/normalization in translated texts operationalize these features in a manner that is tailored to specific language pairs and requires parallel corpora (see Teich 2003; De Sutter and Van de Velde 2008; Hansen-Schirra 2011). Because the empirical methodology presented here is language-universal, it will thus allow for a direct comparison of the degree of syntactic interference/normalization across translations in various language pairs. This universal applicability will assist researchers in categorizing the ways in which explanatory variables pertaining to discrepancies between SLs and TLs – such as in language prestige or linguistic similarity – influence the degree to which translated texts exhibit syntactic interference/normalization. The innovative and finetuned methodology introduced in this presentation fulfills De Sutter and Lefer's (2020) calls for corpus-based translation studies to surpass the simplistic or "teddy-bear" operationalizations of translation features that have thus far been used in the discipline.

Bibliography

De Sutter, Gert, and Marie-Aude Lefer. 2020. "On the Need for a New Research Agenda for Corpus-Based Translation Studies: A Multi-Methodological, Multifactorial and Interdisciplinary Approach." *Perspectives* 28 (1): 1–23. https://doi.org/10.1080/0907676X.2019.1611891.

De Sutter, Gert, and Marc Van de Velde. 2008. "Do the Mechanisms That Govern Syntactic Choices Differ between Original and Translated Language? A Corpus-Based Translation Study of PP Extraposition in Dutch and German." In *Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS)*, 1–38. Hangzhou.

Hansen-Schirra, Silvia. 2011. "Between Normalization and Shining-through: Specific Properties of English-German Translations and Their Influence on the Target Language." In *Multilingual Discourse Production: Diachronic and Synchronic Perspectives*, edited by Svenja Kranich, 135–62. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Teich, Elke. 2003. Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts. Vol. 5. Berlin and New York: Mouton de Gruyter.

"And there they are er I don't know er *for example*": How do EFL students use example markers?

Paula Rodríguez-Abruñeiras – *University of Santiago de Compostela*

Keywords: *example markers, for example, for instance, exemplification; argumentation, selection, EFL.*

Speakers of any language resort to examples to facilitate communication by illustrating their words with cases in point (Triki 2021: 1). In English, the markers used to introduce examples (known as *example markers*, EMs) are *including*, *included*, *such as*, *as*, *say*, *like*, *e.g.*, *for example* and *for instance* (Quirk et al. 1985 or Meyer 1992). Although most studies refer to sequences providing examples as cases of so-called exemplification, Eggs and McElholm (2013) actually distinguish three types of sequences: exemplification, argumentation and selection, although this classification only applies to the EMs *for example* and *for instance*.

Common as the use of examples is, this discourse strategy may pose important difficulties for both L1 and L2 learners (Siepmann 2005: 257), which is why more studies on this discourse function are much needed. The aim of this paper is precisely to investigate the use of EMs in a learner corpus, more specifically *The Santiago University Learner of English Corpus* (SULEC, Palacios Martínez 2002-2022). The paper will try to answer the following questions:

RQ1. What is the frequency of EMs across the different proficiency levels (preintermediate, intermediate and advanced)?

RQ2. How does text type (spoken vs. written) influence the use of EMs?

RQ3. Are there gender differences in the use of EMs?

RQ4. What is the distribution of EMs in exemplification, argumentation and selection?

Overall, 1,051 examples with EMs were retrieved from SULEC, *for example* (446 tokens) and *like* (305) being by far the most common options, whereas *included* (2), *say* (1) and *e.g.* (1) are barely attested. All the markers under analysis prevail in written texts, which seems to point out that the use of examples is commonplace in the written form. If we take into account informants' level of proficiency, the data indicate that speakers show a clear tendency to use *for example* regardless of their level of English. However, its use drops as informants become more proficient in English, thus showing an opposite trend to *for instance* and *such as*, which tend to gain ground in more advanced levels. When it comes to gender, all EMs are more recurring in female productions, except for *like* and *such as*, which predominate in the discourse of males. Finally, the data show that examples with *for example* and *for instance* are mostly given in the form of argumentation, whereas the frequency of exemplification and selection fluctuates depending on the EM used and informants' level of English.

References

Eggs, Ekkehard and Dermot McElholm. 2013. Exemplifications, Selections and Argumentations: The Use of Example Markers in English and German. Bern: Peter Lang.

Meyer, Charles F. 1992. *Apposition in Contemporary English*. Cambridge: CUP.

Palacios Martínez, Ignacio (dir., SPERTUS Research Group). 2002-2022. *The Santiago University Corpus of Learner English (SULEC)*. Santiago: University of Santiago de Compostela. Available at https://sulec.cesga.es/

Quirk, Randolph, Sidney Greenbaum, Greoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Siepmann, Dirk. 2005. Discourse Markers across Languages: A Contrastive Study of Second-level Discourse Markers in Native and Non-native Text with Implications for General and Pedagogic Lexicography. London & New York: Routledge.

Triki, Nesrine. 2021. Exemplification in research articles: Structural, semantic and metadiscursive properties across disciplines. *Journal of English for Academic Purposes* 54:1-13. https://doi.org/10.1016/j.jeap.2021.101039.

El *storytelling* como recurso periodístico para narrar movimientos sociales

Estéfano Rodríguez-Peláez[1] & Beatriz Sánchez-Cárdenas[2]– *University of Niza[1]* – *University of Granada[2]*

Palabras clave: *chalecos amarillos, narratologías, storytelling, estructuras argumentales.*

Los textos periodísticos sobre movimientos sociales extranjeros se caracterizan por adaptar la información a un lector foráneo que no está, a priori, familiarizado con esa realidad. El objeto de estudio de esta investigación es el movimiento francés conocido como los "chalecos amarillos". Surgió en Francia en noviembre de 2018. Empezó siendo un movimiento en contra del aumento del precio del carburante propuesto por el gobierno de Emmanuel Macron. Poco a poco, la revuelta se fue convirtiendo en una crítica general en contra de la política del presidente francés. Analizamos cómo influye la construcción del relato en la recepción de la información del lector de *La Razón* y *El País* y, más concretamente, qué marco semántico se forja en el lector español. Comparamos los resultados de ambos periódicos, pues estos siguen ideologías políticas diferentes. Así nuestra hipótesis es una divergencia de la presentación y caracterización del movimiento entre los dos periódicos. Para poder comparar la información, hemos compilado un corpus *ad hoc* tras una búsqueda exhaustiva en la hemeroteca del periódico que se amplió en la base de datos FACTIVA.

Basamos nuestro análisis de corpus en los principios de la técnica del *storytelling* (Salmon, 2007; Marti y Pélissier, 2012; Pélissier y Eyrès, 2014). En un mundo polarizado, donde quien domina el relato de los hechos es dueño de la "verdad", el *storytelling* se ha vuelto una táctica relevante para las construcciones narrativas. El enfoque metodológico adoptado en este estudio aporta una novedad con respecto a las investigaciones previas, ya que estudiamos este movimiento social bajo el prisma de esta técnica. Esto nos permite acceder a los marcos cognitivos que se activan en el lector. Para ello, analizamos las "proposiciones narrativas" de la narración (Adam, 1997; Heidmann y Adam, 2010; Adam, 2018). Se trata de los segmentos lingüísticos que constituyen el armazón del relato. Etiquetamos las proposiciones narrativas con una tipología semántica que permite inferir una generalización de las estructuras semánticas subyacentes en la narración. Las técnicas narrativas basadas en el *storytelling* se apoyan, como es natural, en recursos lingüísticos.

Con esta premisa, llevamos a cabo un análisis de corpus. Una vez identificadas las estructuras a analizar, etiquetamos los segmentos mediante el programa ATLAS.ti (Friese, 2017). Este etiquetado se basa en cuatro tipologías semánticas: las categorías conceptuales de los sustantivos (Flaux & Van De Velde, 2000; Sánchez-Cárdenas, 2010; Sánchez-Cárdenas y Ramisch, 2019), así como sus roles semánticos (Fillmore, 1972; Van Valin, 1983, 2005; Faber, 2015), los dominios léxicos de los verbos (Faber y Mairal- Usón, 1999) y, por último, los circunstantes. Al agrupar los elementos etiquetados según su similitud, emergen las estructuras recurrentes, lo que muestra los esquemas mentales subyacentes sobre los que los periodistas construyen su relato.

Los resultados preliminares de nuestro estudio muestran qué esquemas mentales se activan en cada periódico. El análisis indica que ambos medios presentan una realidad del movimiento modificada si bien existen diferencias en la representación de ambos, en concreto en lo que respecta a la caracterización y a la cuantificación de los manifestantes.

Bibliografía

Adam, J. M. 1997. "Unités rédactionnelles et genres discursifs: cadre général pour une approche de la presse écrite", *Pratiques 94*/1, 3-18.

Adam, J. M. 2018. Souvent textes varient. Génétique, intertextualité, édition et traduction. París: Classiques Garnier.

Faber, P. 2015. "Frames as a framework for terminology". *Handbook of terminology 1*/14, 14-33.

Faber, P. & Mairal-Usón, R. 1999. "Constructing a Lexicon of English Verbs", *Language Design: Journal of Theoretical and Experimental Linguistics* 2, 150-152.

Fillmore, C. 1972. "Subjects, speakers, and roles". *Semantics of natural language* 1-24. Holland: Riedel.

Flaux, N. & Van de Velde, D. 2000. *Les noms en français: esquisse de classement*. París: Editions Ophrys.

Friese, S. 2017. *ATLAS.ti. 8 Windows – Full Manual*. Berlin: Atlas.ti Scientific Software Development GmbH

Heidmann, U. & Adam, J. M. 2010. Textualité et intertextualité des contes. Perrault, Apulée, La Fontaine, Lhéritier… París: Classiques Garnier.

Marti, M. & Pélissier. 2012 Le storytelling. Succès des histoires, histoire d'un succès. París: L'Harmattan.

Pélissier, N & Eyries, A. 2014. Fictions du réel: le journalisme narratif. *Les Cahiers de Narratologie* 26, 1-10.

Salmon, C. 2007. Storytelling. La Machine à fabriquer des histoires et à formater les esprits, París: La Découverte.

Sánchez-Cárdenas, B. 2010. Paramètres linguistiques pour la conception d'un dictionnaire électronique bilingue (français-espagnol) destiné à la traduction: le cas des verbes de comptage [Tesis doctoral] Universidad de Granada / Universidad de Estrasburgo.

Sánchez-Cárdenas, B. & Ramisch, C. 2019. "Eliciting specialized frames from corpora using argument-structure extraction techniques", *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 25*/1, 1-31.

Van Valin, R. D. 1983. "Pragmatics, ergativity and grammatical relations". *Journal of Pragmatics 7*/1, 63-88.

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

Multidimensional analysis of subgenres in Spanish literary texts

Ignacio Rodríguez-Sánchez – *Universidad Autónoma de Querétaro*

Palabras clave: *análisis multidimensional, literatura en español, subgéneros.*

Dentro de los estudios literarios se suele establecer la adscripción de una obra literaria a un género mediante un vínculo a una tradición histórica y al concepto académico de canon; sin embargo, a veces, la clasificación de subgéneros resulta problemática, y más aún en el caso de determinadas obras de carácter híbrido (por ejemplo, qué similitues y diferencias habría entre una biografía, una novela biográfica, biografía novelada). En el ámbito de la vanguardia y de la cultura popular, se hacen propuestas y se toman decisiones de clasificación basadas en muchos y muy variados criterios: estilísticos, temáticos, sobre la aparición de ciertos personajes, eventos, el tono que imprime el autor, la época, sin excluir también los criterios editoriales, como por ejemplo el público al que va dirigida la obra (novela juvenil).

El Análisis Multidimensional es una técnica estadística que permite la clasificación de textos mediante la la computación de la presencia y ausencia de decenas de rasgos. Estos rasgos se agrupan en dimensiones como narratividad, objetividad (Biber,1988, 2003). El objetivo de esta investigación es averiguar si el Análisis Multidimensional, aplicado a la clasificación de textos literarios permite aportar objetividad a los criterios que se usan para determinar la inclusión de las obras dentro de un subgénero así como para estimar las distancias entre esas mismas obras.

La estadística multivariada o multivariante establece cómo dos o más variables se correlacionan entre sí en grados diversos. Como señalan Almela, Berber Sardinha y Cantos-Gomez (2022), "estos procedimientos estadísticos ayudan al investigador a resumir los datos y reducir la cantidad de variables necesarias para describirlos, explorando la contribución de varios factores a un mismo evento" (p. 546). A diferencia de técnicas informáticas basadas en la inteligencia artificial, pueden considerarse también variables explicativas de la variable dependiente, aunque hay autores que recomiendan mucha cautela a la hora de realizar interpretaciones (Gould, 2006).

Se realizó con éxito un análisis de agrupación jerárquica basado en las palabras referidas a 85 partes del cuerpo para agrupar (por subgénero y por autor) más de 40 obras de narrativa de una veintena de autores. Este trabajo fue una adaptación de algunas investigaciones del grupo de Moretti en Stanford (Alison et al. 2011, entre otros). El proyecto que presentamos aquí amplia el corpus de esa investigación para cubrir subgéneros como la novela negra, policial, narcoliteratura, histórica, ciencia ficción, romántica y juvenil, así como varias obras de carácter híbrido y autores canónicos.

En la presente investigación nuestro preprocesamiento incluye el etiquetado de categoría gramatical y morfológico con Freeling 4.1. Los rasgos que se están analizando parten del análisis clásico de Biber (1988, 2003). Sin embargo, se añadieron rasgos que no eran relevantes para el análisis de textos escritos en inglés (por ejemplo, el uso de subjuntivo o morfemas verbales de persona o número). Asímismo, se está experimentando en la inclusión de palabras de ciertos campos semánticos (sentimientos, presencia de elementos de la naturaleza). Así que el conteo inicial se realiza con cerca de 100 rasgos.

El análisis factorial, a través del proceso de rotación de factores, permite elegir interpretaciones entre una amplia variedad de posibles opciones. La reducción de los rasgos a solo aquellos que presentan un con peso estadísticamente significativo simplifica la labor. Sin embargo, aún no se cuenta con datos que permitan identificar adecuadamente las dimensiones que permitan agrupar con precisión los textos literarios que forman parte del corpus experimental.

Referencias

Allison, S., Heuser, R., Jockers, M., Moretti, F., & Witmore, M. 2011. *Qualitative formalism: An experiment*. Stanford Literary Lab. https://litlab.stanford.edu/pamphlets/

Almela, Á., Berber Sardinha, T., & Cantos-Gómez, P. 2022. Métodos multidimensionales basados en corpus del español. En G. Parodi, P. Cantos-Gómez, & C. Howe (Eds.), *Lingüística de corpus en español*. Routledge.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. 1995. Dimensions of Register Variation: A Cross-Linguistic Comparison. Cambridge University Press.

Gould, S. J. 2006. *The Mismeasure of Man*. W. W. Norton & Company.

The use of manner adverbials as disjuncts in *The Mary Hamilton Papers*

Jesús Romero Barranco – *Universidad de Granada*

In English, adverbs may modify not only adjectives and other adverbs, but also noun phrases, prepositional phrases, particles, numerals and sentences or clauses. Adverbs that modify sentences or clauses are labelled adverbials, and can have different functions in these environments: 1) circumstance adverbials, adding information about the action described in the clause (time, manner and place), i.e. *He took it in slowly but uncomprehendingly*; 2) stance adverbials, conveying the speaker/writer's assessment of the proposition in the clause, i.e. *His book undoubtedly fills a need*; and 3) linking adverbials, connecting stretches of text (phrases, sentences and paragraphs), i.e. *Nevertheless, the review represents substantial progress* (Biber et al. 1999: 548-549). From a historical perspective, the nineteenth century was a transitional period in some changes that took place in the English adverb phrase, and a case in point could be the use of adverbs of manner as disjuncts, which modify a whole sentence or clause (i.e. *Surprisingly, we got there in time*) and usually express the speaker's attitude (Denison 1998: 234). It has been argued that many of these adverbials were first used in these environments in the seventeenth century, and that they were widespread by the eighteenth century (Swan 1988, 1990).

As far as I have been able to investigate, the phenomenon has not been studied so far in historical private correspondence and, therefore, the present paper sets out with the following objectives: 1) to assess the use and distribution of manner adverbials as disjuncts in Late Modern English (1740-1850); 2) to classify the adverbials according to their function in the sentence, that is, circumstance adverbials or stance adverbials; and 3) to observe whether the socio-biographical information of informants constitutes a decisive factor in the use of manner adverbials as disjuncts in Late Modern English. The source of evidence comes from *The Mary Hamilton Papers*, a collection of letters belonging to Mary Hamilton, one of the most well-connected and highly cultured women in eighteenth-century British polite society (Denison et al. 2019).

References

Biber, Douglas, Stig Johansson, Geoffrey N. Leech, Susan Conrad, Edward Finegan. 1999. *Grammar of Spoken and Written English*. London: Longman.

Denison, David. 1998. "Syntax". *The Cambridge History of the English Language, Vol. IV, 1776-1997*, edited by Suzanne Romaine. Cambridge: Cambridge University Press. 92-329.

The Mary Hamilton Papers (c.1740-c.1850). Compiled by David Denison, Nuria Yáñez- Bouza, Tino Oudesluijs, Cassandra Ulph, Christine Wallis, Hannah Barker and Sophie Coulombeau, University of Manchester. In progress, 2019-.

Swan, Toril. 1988. Sentence Adverbials in English: a Synchronic and Diachronic Investigation. Oslo: Novus.

Swan, Toril. 1990. "The development of sentence adverbs in English". *Tromso Linguistics in the Eighties*, edited by E. H. Jahr & O. Lorentz. (Tromso Studies in Linguistics 11.) Oslo: Novus Press, 369-88.

Análisis de la dimensión cultural en equivalentes terminológicos (español-francés) del sector forestal

Sara Rupérez-León – *University of Valladolid*

Palabras clave: *equivalencia terminológica, traducción, cultura, semiótica cognitiva, lingüística de corpus, semántica distribucional, sector forestal.*

Uno de los principales retos a los que se enfrenta todo traductor, y que lo diferencia de cualquier recurso informático, es la capacidad de identificar las variables que influyen en la elección de equivalentes terminológicos. En este proceso no solo intervienen parámetros lingüísticos. La cultura adquiere un papel primordial, pues vertebra las decisiones del traductor al manifestar el grado de (re)conocimiento que posee el público meta con respecto a un concepto lexicalizado del texto origen. Desde esta perspectiva, el referente de toda expresión terminológica queda supeditado a su interpretación tanto en su contexto de producción como de recepción.

Este estudio pretende averiguar cómo influye la dimensión cultural en la lexicalización y búsqueda de equivalentes terminológicos en lenguas española y francesa. Para ello, presentamos una propuesta teórico-práctica basada en la semiótica cognitiva (Gottlieb, 2018) y en el modelo de paracultura, diacultura e idiocultura iniciado por Ammann (1989) y continuado por Reiss y Vermeer (1996), Nord (1997), Witte (2000) y Martín (2003).

Como ámbito de aplicación, nos centramos en términos compuestos propios del sector forestal. La importancia atribuida al bosque deriva de su concepción como patrimonio colectivo. No obstante, las investigaciones interlingüísticas (español-francés) que existen sobre este campo de especialidad son insuficientes, lo que ha entorpecido el desarrollo de una terminología forestal contrastiva.

Este ámbito plantea varias dificultades que podrían explicar parcialmente la falta de recursos terminológicos disponibles. Además de la tecnicidad de su terminología, el sector forestal se encuentra íntimamente ligado a su correspondiente marco jurídico, donde la dimensión cultural adquiere especial relevancia. Esto conlleva la identificación de culturemas, entendidos como una asimetría lingüística y/o conceptual entre las lenguas involucradas, y la dificultad manifiesta de trasladarlos a la lengua de destino. Ejemplos de ello en este campo son los términos *suerte de pinos* en español y *affouage* en francés. Ambos se refieren a un privilegio que poseen los vecinos de un municipio con respecto a los rendimientos obtenidos en los aprovechamientos madereros del monte. Sin embargo, estos dos conceptos presentan diferencias sustanciales debido a su anclaje en su correspondiente marco jurídico, lo que, unido a la falta de equivalentes interlingüísticos acuñados y a la opacidad de su significado, convierte a estos términos en culturemas. Esto hace que sea necesario abordar su equivalencia desde una estrategia concreta.

En este estudio, presentamos una propuesta para resolver cómo se lexicalizan y establecen equivalencias interlingüísticas dentro de la terminología forestal con base en la semiótica cognitiva y el modelo cultural.

Como metodología adoptamos la Lingüística de corpus con la compilación de un corpus comparable en lenguas española y francesa denominado FORESCOR, así como la semántica distribucional para determinar el grado de equivalencia a través de la comparación de cotextos (Cabezas-García, 2021). Como herramienta se utiliza Sketch Engine (SkE), referente por sus amplias funcionalidades en este campo (Sánchez-Cárdenas y López Rodríguez, 2020; Cabezas-García, 2021).

Tras la extracción y selección terminológica, nuestra propuesta se divide en las siguientes fases: (i) análisis de términos compuestos con base en la semiótica cognitiva y los parámetros de paracultura, diacultura e idiocultura; (ii) aplicación de la semántica distribucional en la búsqueda de equivalentes español-francés y evaluación de este

método en SkE; (iii) análisis de posibles estrategias traductológicas; y (iv) propuesta de equivalentes terminológicos interlingüísticos. Nuestra comunicación muestra qué protocolo puede adoptarse para el traslado de términos compuestos forestales (entre ellos, culturemas) según este nuevo paradigma cultural.

Referencias bibliográficas

Ädel, A. 2020. Chapter 1. Corpus compilation. En M. Paquot, & S. Gries (Edits.), *A Practical Handbook of Corpus Linguistics* (1ª ed., págs. 3–24). Nueva York: Springer Nature Switzerland AG.

Ammann, M. 1989. Grundlagen der modernen Translationstheorie: ein Leitfaden für Studierende. Heidelberg: Instituts für Übersetzen und Dolmetschen.

Cabezas García, M. 2021. Metodología para la traducción de términos compuestos mediante corpus. *Mutatis Mutandis, 14*(2), 451-468. https://doi.org/10.17533/udea.mut.v14n2a08

Castillo Rodríguez, C., Díaz Lage, J., & Rubio Martínez, B. 2020. Compiling and analyzing a tagged learner corpus: a corpus-based study of adjective uses. *Círculo de Lingüística Aplicada a la Comunicación*(81), 115-136.

Cepeda Ortega, J. 2018. Una aproximación al concepto de identidad cultural a partir de experiencias: el patrimonio y la educación. *Tabanque. Revista Pedagógica* (31), 244-262.

García-Miguel, J. M. 2022. Lingüística de corpus: de los datos textuales a la teoría lingüística. *Estudios de Lingüística del Español, 45*, 11-42. https://raco.cat/index.php/Elies/article/view/403735

Gómez González-Jover, A. 2002. La equivalencia como cuestión central de la traducción en las instituciones de la Unión Europea. *Actas del I Congreso Internacional «El Español, Lengua de Traducción»*. Almagro. 438-457.

Gottlieb, H. 2018. Semiotics and translation. En K. Malmkjaer (Ed.), *The Routledge Handbook of Translation Studies and Linguistics* (págs. 45-63). Londres: Routledge.

Martín De León, C. 2003. *Metáforas en la traductología funcionalista.* Las Palmas de Gran Canaria: Universidad de las Palmas.

Nord, C. 1997. Translating as a Purposeful Activity: Functionalist Approaches Explained. Mánchester: St. Jerome Publishing.

Pérez Hernández, M. C. 2002. Terminografía basada en corpus: principios teóricos y metodológicos. En P. Faber, & C. Jiménez (Edits.), *Investigar en terminología* (págs. 127-166). Granada: Comares.

Reiss, K. & Vermeer, H. 1996. *Fundamentos para una teoría funcional de la traducción*. Madrid: Ediciones Akal.

Sánchez Cárdenas, B. & López Rodríguez, C. I. 2020. *Retos de la traducción científico-técnica profesional: Teoría, metodología, recursos* (Vol. Interlingua). Granada: Comares.

Valdenebro Sánchez, J. 2019. Estudio contrastivo del anisomorfismo cultural (Francia y España) de la terminología penal. *Hikma: Estudios de Traducción*, *18(1)*, 231-260. https://helvia.uco.es/bitstream/handle/-10396/19478/hikma_18_01_08.pdf?sequence=1&isAllowed=y

Witte, H. 2000. Die Kulturkompetenz des Translators: Begriffliche Grundlegung und Didaktisierung. Tubinga: Stauffenburg.

Witte, H. 2005. El traductor como mediador cultural. Fundamentos teóricos para la enseñanza de la lengua y cultura en los estudios de traducción. *El Guiniguada*, *1(3)*, 407-414.

Equivalencia interlingüística en estructuras fraseológicas verbo- nominales de conceptos especializados

Beatriz Sánchez-Cárdenas – *University of Granada*

Palabras clave: *fraseología, colocaciones verbales, terminología, equivalencia interlingüística.*

Desde una perspectiva traductológica, es fundamental considerar los segmentos del texto que van más allá de unidades léxicas individuales para establecer correspondencias interlingüísticas. Por lo tanto, el dominio de la fraseología es una necesidad para cualquier traductor que desee producir un texto lo más natural y correcto posible (Corpas Pastor 2008, Tutin 2014). Las unidades fraseológicas se dan a todos los niveles del discurso científico, tales como expresiones metatextuales (*formular una hipótesis*), marcadores intrapersonales (*como bien es sabido*), conectores lógicos (*por lo tanto*), expresiones de actitud (*defender una postura*), marcadores modales (*hasta cierto punto*) (Tutin & Falaise 2013), y combinaciones verbo-nominales (*expulsar lava*) (Buendía Castro y Sánchez Cárdenas 2012). Este estudio se centra en el estudio de colocaciones de verbo-nominales en discursos especializados. Entendemos por "colocación verbo-nominal" combinaciones de dos o más palabras muy frecuentes formadas por estructuras en la que hay un sustantivo y un verbo o un verbo más un sustantivo, donde el sustantivo es la base con la que coloca el verbo (Buendía 2013; Buendía, Montero y Faber 2014). Describimos un protocolo de extracción y análisis de corpus de fraseología, diseñado para representar las equivalencias interlingüísticas verbo-nominales en recursos terminológicos destinados a la traducción. En esta comunicación presentamos los resultados del análisis de un subcorpus en inglés y otro español de un millón de palabras cada uno formados por textos científicos que tratan los procesos asociados con la "deforestación". Mediante búsquedas complejas en el corpus, se extrajeron patrones léxicos en forma de triples (sustantivo- verbo-sustantivo), donde el término es el núcleo de un argumento situado antes o después del verbo: "La deforestación provoca un aumento global de la temperatura"; "La emisión de gases de efecto invernadero contribuye a la deforestación".

Nuestras técnicas de extracción utilizan dos herramientas de PNL: Sketch Engine (Kilgarriff et al 2014) y MWEtool kit (Ramisch 2015). Extraemos de manera automática combinaciones con la estructura "nombre-verbo-nombre". Una vez extraídas, se seleccionan manualmente de acuerdo con su pertinencia. Se evalúa la exhaustividad y precisión de ambas herramientas para esta tarea y se analizan los errores de extracción. El mismo protocolo se llevó a cabo en dos corpus comparables, en español y francés, con resultados similares. Por último, los triples en ambos idiomas se agrupan automáticamente mediante una medida de similitud semántica basada en los "word embeddings" (Camacho-Collados & Taher 2020) utilizando el software Gensim. Se generan todas las combinaciones posibles de sustantivos y verbos que pueden aparecer en cada posición del triple. Se obtienen así las estructuras verbo-nominales fraseológicas recurrentes del término analizado tanto en español (*la demanda creciente de alimentos impulsa la deforestación*; *la ganadería comercial contribuye a la deforestación*), como en inglés (*soybean expansion leads to deforestation*; *technological change increases deforestation*).

En la segunda parte del estudio, se reflexiona sobre el concepto de equivalencia interlingüística. En concreto, se propone situar la equivalencia verbal de los discursos especializados a nivel pragmático-textual, alejándonos de una perspectiva exclusivamente lingüística, con el objetivo de mejorar la idiomaticidad de las traducciones. Ilustramos esta idea mediante ejemplos concretos utilizando los resultados de nuestro estudio. Para ello se compara la calidad de segmentos traducidos a la luz de los resultados de nuestro estudio con traducciones basadas en diccionarios bilingües tradicionales o en herramientas de traducción automática.

Referencias bibliográficas

Buendía Castro, Miriam 2013. Phraseology in Specialized Language and its Representation in Environmental Knowledge Resources. PhD Thesis. Universidad de Granada, Granada, Spain.

Buendía Castro, Miriam and Sánchez Cárdenas, Beatriz 2012. Linguistic knowledge for specialized text production. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12),* Calzolari, N., Choukr, K., Declerc, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J. & Piperidis, S. (eds), 622-626. Istanbul.

Buendía Castro, Miriam, Montero Martínez, Silvia and Faber, Pamela 2014. Verb collocations and phraseology in EcoLexicon. *Yearbook of Phraseology,* 5(1):57-94. De Gruyter.

Camacho-Collados, José y Mohammad Taher Pilehvar 2020. Embeddings in natural language processing. *Proceedings of the 28th international conference on computational linguistics: tutorial abstracts,* 10-15.

Claveau, Vincent, & L'Homme, Marie-Claude. 2006. Discovering and organizing noun-verb collocations in specialized corpora using inductive logic programming. *International Journal of Corpus Linguistics,* 11(2), 209-243.

Corpas Pastor, Gloria 2008. Investigar con corpus en traducción: los retos de un nuevo paradigma (Vol. 49). Peter Lang.

Faber Pamela, and Ricardo Mairal. 1999. Constructing a Lexicon of English Verbs. Mouton de Gruyter. Granger, Sylviane, and Meunier, Fanny (eds) 2008. *Phraseology: An interdisciplinary perspective.* John Benjamins Publishing.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovvář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel 2014. The Sketch Engine: ten years on. *Lexicography,* 1: 7-36.

L'Homme, Marie-Claude 2015. Predicative lexical units in terminology. Language Production, Cognition, and the Lexicon, 75-93. Springer International Publishing.

León-Araúz, Pilar 2022. Terminology and equivalence. In L'Homme, Marie-Claude, and Pamela Faber. *Theoretical Perspectives on Terminology,* John Benjamins Publishing, 477–502.

Ramisch, Carlos 2015. *Multiword Expressions Acquisition: A Generic and Open Framework.* Theory and Applications of Natural Language Processing series, XIV. Springer.

Tutin, Agnès 2014. *L'écrit scientifique: du lexique au discours.* F. Grossmann (Ed.). Presses universitaires de Rennes.

Williams, Geoffrey 2005. English Collocation Studies: The OSTI report. *International Journal of Lexicography,* 18(3): 391-393.

Does *violencia de género* translate *domestic violence*, and vice-versa? Parallel contrastive naming practices (English/Spanish) in media language about violence against women

José Santaemilia & Sergio Maruenda – *University of València*

Keywords: *violence against women, naming practices, corpus-based approach, translation, contrastive pragmatics.*

Naming, defining and translating sensitive socio-ideological realities (abortion, same- sex marriage, or violence against women, among others) is increasingly becoming more difficult and problematic day by day, as well as constituting a highly ideological task. In this paper we resort to the combined potentials of corpus linguistics, contrastive pragmatics and translation studies in order to unveil the ways we conceptualise, understand or translate key terms in newspaper stories about violence against women, in English and in Spanish. To this end, we will focus on the main labelling (or naming) practices surrounding the VAW phenomenon as depicted (and translated) in English (*The Guardian* and *The Times*) and Spanish (*El País* and *El Mundo*) quality papers, in a macrocorpus containing ca. 20 million words (2000-2020). In short, we study "how sexual and domestic violence against women is framed through language" (Klein 2013: 1), both in English and in Spanish.

This cross-cultural perspective, which benefits from the triangulation of data offered by corpus linguistics, cross-cultural pragmatics and translation studies aims at explaining cultural "imprints" (Stefan Hauser & Martin Luginbühl 2012) in media communication about VAW –i.e. how media texts from different socio-cultural communities differ (or concur) in transmitting socio-cultural norms and values, connected to the media framing of the phenomenon of violence against women, and to the associations activated by the linguistic representation of this phenomenon.

A corpus-based contrastive pragmatics perspective allows us to shed new light on the naming practices that are used, in English and Spanish, to define and characterise the same socio-ideological reality, giving us deeper insights into the individual languages involved (English and Spanish) and also on the common ground and ideological connotations carried by the terminology used. Although the reality of VAW may be the same everywhere, the differing language or terminology used, the (dis-)preferred expressions and ideological associations employed, may end up producing (slightly or significantly) different versions of reality and, in some sense, we are witnessing a parallel process of translating –i.e. of (re)inscribing a highly sensitive social issue into the very fabric of its specific society.

The fact that there are significant differences (between English and Spanish) in the most widely used terms to refer to VAW is a powerful indication of the diverging ideological routines, prejudices or stereotypes ingrained in a specific linguaculture (Friedrich 1989) around this topic, and is a worthwhile object of a corpus-based, contrastive and translation-related investigation.

Bibliography

Baker, Paul et al. 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.

Baker, Paul & Erez Levon 2015. "Picking the right cherries? A comparison of corpus- based and qualitative analyses of news articles about masculinity." *Discourse and Communication* 9(2): 221–236.

Caldas-Coulthard, Carmen Rosa & Rosamund Moon 2010. "'Curvy, hunky, kinky': Using corpora as tools for critical analysis". *Discourse & Society* 21(2): 99-133.

Carranza Márquez, Aurelia 2010. "Testimonies in the British and Spanish Parliaments: A contrastive study on domestic/gender violence". *Journal of Pragmatics* 42: 2172- 2180.

Conway, Kyle 2015. "News translation and culture." *Perspectives* 23(4): 517-520. Ebeling, Jarle (1998) "Contrastive Linguistics, Translation, and Parallel Corpora". *Meta* 43(4): 602-615.

Flotow, Luise von & Joan W. Scott 2016. "Gender studies and translation studies: "Entre braguette" – connecting the transdisciplines." In Yves Gambier & Luc van Doorslaer (eds.) *Border Crossings. Translation studies and other disciplines.* Amsterdam/Philadelphia: John Benjamins. 349-374.

Friedrich, P. 1989. "Language, ideology, and political economy." *American Antropologist* 91(2): 295-312.

Hauser, Stefan & Martin Luginbühl (eds.) 2012. *Contrastive media analysis: approaches to linguistic and cultural aspects of mass media.* Amsterdam/ Philadelphia: John Benjamins.

Klein, Renate (ed.) 2013. *Framing sexual and domestic violence through language.* Basingstoke/London: Palgrave.

Kranich, Svenja 2016. Contrastive pragmatics and translation: Evaluation, epistemic modality and communicative styles in English and German. Amsterdam/Philadelphia: John Benjamins.

Kranich, Svenja 2019. "Contrastive approaches to pragmatics and translation." In Tipton, Rebecca & Louisa Desilla (ed.) *The Routledge Handbook of Translation and Pragmatics.* London/New York: Routledge. 115-128.

Santaemilia, José 2013. "Translating international gender-equality institutional/legal texts: The example of 'gender' in Spanish." *Gender and Language* 7(1): 71–92.

Schmied, Josef 2009. "Contrastive corpus studies". In Lüdeling, Anke & Merja Kyt. (eds.) *Corpus Linguistics: An International Handbook. Volume 2*. Berlin/New York: Walter De Gruyter. 1140-1158.

Vandepitte, Sonia & Gert De Sutter 2013. "Creativity Contrastive Linguistics and Translation Studies." In Gambier, Yves & Luc van Doorslaer (eds.) *Handbook of Translation Studies. Vol. 4*. Amsterdam/Philadelphia: John Benjamins. 36-41.

Exploring the sociolinguistic development of the FACE diphthong in late modern Derbyshire dialect: A corpus-based diachronic study

Paula Schintu-Martínez – *University of Salamanca*

Keywords: *Derbyshire dialect, dialect writing, enregisterment, FACE diphthong.*

Historical sociolinguistic research has often been questioned in terms of the representativeness of its data sources and scientific reliability. Unlike contemporary sociolinguistic variation, which can easily be explored via the retrieval of first-hand information from participants, the indirect reconstruction of historical dialects is much more challenging, especially given the scarcity of direct information about the linguistic practices of bygone speech communities. Nevertheless, as Conde-Silvestre and Hernández-Campoy (2012) claim, third-wave developments in historical sociolinguistics have "conferred upon the discipline both 'empirical ease' and 'historical confidence'" via the corpus-based analysis of written sources, which "partly solves the fragmentary nature of historical material [and] ensures that variability in past stages can reliably be reconstructed" (3). Third-wave sociolinguistic approaches to historical varieties of English have thus renewed the interest in written records of language, particularly, as Hodson (2016: 28) points out, in the form of literature. Literary representations of dialect have proven to be useful tools not only to understand linguistic variation in the past, but also in order to unearth processes of indexicality and enregisterment (Silverstein 1976; Agha 2003), "both of which play a remarkable role in linguistic change" (Sánchez-García and Ruano-García 2020: 54-55).

This paper explores the sociolinguistic development of 19th and 20th-century Derbyshire dialect by focusing on the non-standard respellings used in literary representations of the variety to portray the local pronunciation of the FACE lexical set (Wells 1982). I draw on indexicality and enregisterment and focus on a corpus of thirteen texts extracted from *The Salamanca Corpus* with a threefold purpose. First, I aim at (1) identifying the most common dialectal realisations of the FACE diphthong in the county, while (2) ascertaining if time-dependent indexical changes may have affected the way in which this feature was understood and thus represented over the course of the Late Modern period. Besides, I also seek to (3) determine whether such recreation varied depending on the type of dialect representation, i.e. dialect literature or literary dialect (Shorrocks 1996), and, therefore, on the degree of familiarity with the variety of the author and the audience addressed. The corpus-based quantitative analysis of the texts has revealed discernible patterns in the depiction of Derbyshire FACE. These include a set of respellings which suggest non-standard pronunciations like [a] (e.g. *mak* 'make'), [e] (e.g. *mebbe* 'maybe'), and [i:]/[ɪə] (e.g. *neame* 'name'), which largely correlate with contemporary non-literary metalinguistic evidence on the dialect, thus pointing at third-order indexicality and enregisterment. Nevertheless, the data has likewise unveiled variation in the representation of FACE throughout the period analysed and across text types, revealing meaningful indexical shifts operating not only over time, but also over changing populations of language users.

Bibliography

Agha, Asif. 2003. The Social Life of Cultural Value. *Language & Communication* 23: 231–273.

Conde-Silvestre, Juan. C. and Juan M. Hernández-Campoy. 2012. Introduction. In Juan C. Conde-Silvestre, and Juan M. Campoy eds. *The Handbook of Historical Sociolinguistics*. Blackwell Publishing Ltd., 1-8.

Hodson, Jane. 2016. Talking like a servant: What nineteenth-century novels can tell us about the social history of the language. *Journal of Historical Sociolinguistics* 2/1, 27–46.

Sánchez-García, Pilar and Javier Ruano-García. 2020. Some Challenges behind the Compilation of the Salamanca Corpus: The Wiltshire Dialect as a Case Study. *Nexus* 2, 53-66.

Shorrocks, Graham. 1996. Non-Standard Dialect Literature and Popular Culture. In Juhani Klemola et al. eds. *Speech Past and Present. Studies in English Dialectology in Memory of Ossi Ihalainen*. Frankfurt am Main: Peter Lang, 385- 411.

Silverstein, Michael. 1976. Shifters, Linguistic Categories, and Cultural Description. In Ben G. Blount ed. *Language, Culture, and Society. A Book of Readings*. Long Grove: Waveland, 187-221.

Wells, John C. 1982. *Accents of English*. Cambridge University Press.

***The Organic Food Corpus*: A multilingual resource for the understanding of consumer attitudes towards organic food products**

Vasiliki Simaki – *Lund University*

Keywords: *corpus compilation, social media, user generated content, organic food consumption.*

In this paper, the design, compilation and analysis of the Organic Food Corpus (OFC) is presented. The corpus was created within the LangTool project[2] that studied consumer attitudes towards sustainable food consumption. The consumption of sustainable products[3] is seen as one of the ways that sustainable development, which has a great potential to deliver social well-being and contribute to environmental protection[4], can be achieved.

Consumers express their opinions and attitudes about products/services in various online spaces, and the number of online reviews has exploded over the past two decades (Vásquez 2014). Consumer feedback is important not only for other consumers, but also for the service/product provider. In many cases, a review may be more powerful than traditional forms of advertising (Cochrum 2011), and it can even have an economic impact on businesses (Ghose and Ipeirotis 2011). For the compilation of the OFC, user generated content from different social media sources (Twitter, Facebook, YouTube, Reddit, Instagram) was collected based on a set of thematic keywords terms in English, Swedish and Greek such as *organic, GMO, ekologiskt, hållbart, βιολογικά, οργανικά*, etc. The search focused on keywords related to organic food products, as they are strongly related to sustainable food products and development (Abeliotis et al. 2010), and they are strictly defined (e.g., Jones et al. 2001, Schmid and Sinabell 2006) and regulated[5].

For the data extraction, APIs and web scraping tools were used. This process was performed from February to May 2019, and texts including any of the given keywords from 2006 until 2019 were extracted. The size of the OFC is about 50 million words, and includes content in 19 languages. However, the largest part of the corpus consists of texts in English. The analysis of subsets of the corpus, focusing mainly on the English, Swedish and Greek content, showed that there are two main categories of texts related to organic food: *consumer* and *commercial* content. While most of the commercial content is promotional and aims to trigger the consumer sensitivity towards health and environmental concerns, the consumer content is more diverse in terms of functions and communicational goals. Consumers' language and their stances (my approach is based in Simaki et al. 2020 and Simaki et al. 2022) towards organic food were then explored. Both positive and negative stances that were strongly supported used in many cases markers such as hashtags (e.g., #*Weloveorganic*, #*healthy*) and hyperlinks. Consumers who have a negative stance often distrust that organic products are actually organic, and they consider them as a costly choice that is not really efficient to health and environment. Based on a thematic analysis of the data, organic food is strongly related to concepts such as sustainability, healthy and environment-friendly lifestyles, vegetarian and vegan choices. During the OFC creation, several issues related to ethical, formatting and language problems, which have been discussed in previous studies (e.g., Hernández 2014), were faced. Noise was also a challenge due to the amount of the data, and various tasks were performed to reduce this phenomenon.

[2] https://projekt.ht.lu.se/langtool/

[3] https://www.un.org/sustainabledevelopment/sustainable-development-goals/

[4] https://sdgs.un.org/2030agenda

[5] https://agriculture.ec.europa.eu/farming/organic-farming_en

Bibliography

Abeliotis, K., Koniari, C., & Sardianou, E. 2010. The profile of the green consumer in Greece. *International Journal of Consumer Studies*, *34*(2), 153-160.

Cockrum, J. 2011. Free marketing: 101 low and no-cost ways to grow your business, online and off. John Wiley & Sons.

Ghose, A. & Ipeirotis, P. G. 2010. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, *23*(10), 1498-1512.

Hernández, Nuria. 2014. New media, new challenges: exploring the frontiers of corpus linguistics in the linguistics curriculum. *Research in Corpus Linguistics* 1: 17-31.

Jones, P., Clarke-Hill, C., Shears, P. and Hillier, D. 2001. Retailing organic foods. *British Food Journal*, Vol. 103 No. 5, pp. 358-365.

Schmid, E. and Sinabell, F. 2007. Modelling Organic Farming at Sector Level. An Application to the Reformed CAP in Austria. *WIFO Working Papers No. 288*.

Simaki, V., Paradis, C., Skeppstedt, M., Sahlgren, M., Kucher, K., & Kerren, A. 2020. Annotating speaker stance in discourse: the Brexit Blog Corpus. *Corpus Linguistics and Linguistic Theory*, *16*(2), 215-248.

Simaki, V., Seitanidi, E., & Paradis, C. 2022. Evaluation of stance annotation in Twitter data. *Research in Corpus Linguistics*.

Vásquez, C. 2014. *The discourse of online consumer reviews*. Bloomsbury Publishing.

Disciplinary variation in academic discourse:
A multi-dimensional analysis of research papers in 'hard' and 'soft' sciences

Elizaveta Smirnova[1] & Javier Pérez-Guerra[2] – *HSE University*[1] – *University of Vigo*[2]

Keywords: *professional academic writing, corpus, multi-dimensional analysis, disciplinary variation, 'hard' sciences, 'soft' sciences.*

In the relevant literature it has been shown that academic discourse varies considerably not only in terms of word frequencies, rhetorical moves and collocational patterns (see, for example, Hyland, 2008), the use of particular grammatical structures (see, for example, Hiltunen, 2016), but also as regards functional dimensions (among others, Staples et al., 2016; Crossley et al., 2019). From this perspective, this study aims at determining the degree of similarity and of differences between so-called 'hard' and 'soft' sciences through the application of Douglas

Biber's Multi-Dimensional (MD) analysis – the labels 'hard' and 'soft', attributed to Storer (1967), are used to compare scientific fields on the basis of perceived methodological rigour, exactitude and objectivity. The MD technique consists of the implementation of "factor analysis for the extraction of latent dimensions of variation from patterns of co-occurrence of linguistic features" (Nini, 2019: 67). Deviations in the way in which dimensions are materialised in different disciplines can contribute to research in disciplinary variation and on the peculiarities of professional academic writing, with practical implications for EAP (English for Academic Purposes) instruction.

Specifically, this study undertakes a quantitative and qualitative analysis of register variation (*à la* Biber, 1988), conducted on a 1,597,000-word corpus of research articles in four ('soft') arts and social sciences (business studies, linguistics, history and political science), and four ('hard') life and physical sciences (mathematics, engineering, chemistry and physics), published in leading journals. The goal is twofold: first, to describe the linguistic factors that exert an influence on register variation in academic discourse and, second, to test the hypothesis that there are significant differences in the realisation of such linguistic features across disciplines.

The MD analysis was performed using the Multidimensional Analysis Tagger (Nini, 2015). Subsequently, factor analysis allowed to detect the most significant features contributing to register variation in professional academic writing. After factor analysis, a single factor proved to explain almost 67 percent of variance. In light of the grouping of the features and their contribution (factor loadings) into such a factor, the latter's interpretation or, in functional terms, of the 'dimension', has been 'Involved versus argumentative style'. On the one hand, this dimension is justified by the salient frequencies of features with positive loadings such as amplifiers, present-tense verbs and prepositional phrases (commonly functioning as situational adverbials), which increase discourse vividness. On the other hand, the argumentative pole is explained via the contribution of features with negative loads such as those purported to bleach agent arguments (passives and first-person subjects), those instantiating links in discourse (conjunctions, predicting modals), narrative features (time adverbials), and other strategies taken as proxies for increased linguistic complexity (*wh*-clauses, gerunds, participial clauses). Once the relevant features and their contribution to the resulting dimension were determined, the disciplines were plotted along the dimension. Such a scaling revealed a distinct trend of soft sciences clustering toward the top of the scale, that is, closer to the Involved style pole, whilst the hard sciences were located in the lower part, interpreted as evincing Argumentative style.

References

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University.

Crossley, S. A., Kyle, K., & Römer, U. 2019. Examining lexical and cohesion differences in discipline-specific writing using multi-dimensional analysis. In Berber Sardinha, T. & Veirano Pinto M.(eds.) *Multi-Dimensional Analysis: Research Methods and Current Issues*. London, New York: Bloomsbury Academic, 189-216.

Hiltunen, T. 2016. Passives in academic writing: Comparing research articles and student essays across four disciplines. In López-Couso, M. J., Méndez-Naya, B., Núñez-Pertejo, P. & Palacios-Martínez, I. M. (eds.) *Corpus Linguistics on the Move: Exploring and Understanding English through Corpora.* Amsterdam: Rodopi, 132-157.

Hyland, K. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes 27*(1), 4-21.

Nini, A. 2015. *Multidimensional Analysis Tagger 1.3 – Manual.* <https://sites.google.com/site/multidimensionaltagger/> (accessed 4 December, 2021)

Nini, A. 2019. The multi-dimensional analysis tagger. In Berber Sardinha, T. & Veirano Pinto M.(eds.) *Multi-Dimensional Analysis: Research Methods and Current Issues*. London, New York: Bloomsbury Academic, 67-94.

Staples, S., Egbert, J., Biber, D., & Gray, B. 2016. Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*(2), 149-183.

Storer, N. W. 1967. The hard sciences and the soft: Some sociological observations. *Bulletin of the Medical Library Association 55*(1), 75-84.

Translating contrastive markers in journalistic texts: The case of the translation of Dutch *maar* into French by translation students and professional translators

Nathanaël Stilmant – *Université de Mons*

Keywords: *contrastive markers, journalistic texts, translation, students, professional translators, Dutch, French.*

As "one of the richest groups of discourse markers" (Josep Cuenca, Postolea, & Visconti, 2019: 3), contrastive markers have received considerable attention from various angles of study, such as their prosody (Petit, 2010), their frequency (Dupont, 2019), or their acquisition by language learners (Lee, 2020). However, their role in the translation process deserves further investigation. This study focuses on the translation of the most typical Dutch contrastive marker (Perrez, 2006), *maar*, into French, by translation students and professional translators, adopting a three-dimensional semantic view of the notion of contrast.

The first contrastive semantic category is concession, which links two ideas, the second of which has a greater argumentative weight as in "Bill studied hard, but he failed the exam" (Izutsu, 2008: 649). Following Adam (1990), we distinguish three subtypes of concessive *maar*: simple concessions (to which the above example belongs) based on the idea of a denied expectation, additive concessions as in "For the adventure, of course, but also for a cinema lesson" (Adam, 1990: 192), and non-verbal *maar* who act on the structure of the discourse rather than on its elements as in "But do something!" (Adam, 1990: 199). The second category is opposition, as in "John is tall but Bill is short" (Lakoff, 1971: 133), where *maar* highlights "the existence of a difference between two discourse segments" (Dupont, 2019: 43). The third category relates to the correction of a false element, as in "Pierre is not French, but Danish" (Birkelund, 2009).

This study consisted in asking 18 students in the final year of the Master's degree in Dutch translation at the University of Mons and 10 professional translators (according to the PACTE criteria, PACTE, 2008) to translate an authentic Dutch journalistic text containing several occurrences of *maar* from all the above mentioned semantic categories. The translations were then classified into three categories: literal translations (*maar* is translated as *mais* in French), translations by a marker other than *mais*, and unmarked relations, i.e., sentences without any marker (Corminboeuf, 2014).

The results showed that literal translations are most used by students for most semantic categories of *maar*. Simple concessive *maar*, additive concessive *maar*, and oppositive *maar* were translated literally by more than three quarters of the students. The translations of the corrective *maar* balance relatively well between literal translations and unmarked relations. The non-verbal *maar* are fundamentally different from the other categories, with almost no literal translations. Their translational strategies vary depending on whether they link elements of written discourse or elements from reported oral discourse (in this case, quotations): students generally translate the non-verbal *maar* from quotations with an unmarked relation, whereas they almost always translate the non-verbal *maar* from written discourse with a marker other than *mais*. Furthermore, these non-verbal *maar* from written discourse show the largest differences in translation strategies between the students and the professional translators who, unlike the students, overwhelmingly translate these *maar* literally.

We can thus observe that, when it comes to instinctively translating the most typical contrastive marker from Dutch into French, the strategies of students and professional translators are only partially similar, according to the semantic categories of this marker.

References

Adam, J.-M. 1990. Éléments de linguistique textuelle. Théorie et pratique de l'analyse textuelle. Liège: Mardaga.

Birkelund, M. 2009. Pierre n'est pas français mais danois. Une structure polyphonique à part. *Langue française,* 164(4), 123-135.

Corminboeuf, G. 2014. L'identification des relations de discours implicites: le cas de l'adversation. Paper presented at *4e Congrès Mondial de Linguistique Française – CMLF* 2014, Berlin.

Dupont, M. 2019. Conjunctive Markers of Contrast in English and French. From Syntax to Lexis and Discourse (Doctoral dissertation, Université Catholique de Louvain).

Izutsu, M. N. 2008. Contrast, concessive, and corrective: Toward a comprehensive study of opposition relations. *Journal of Pragmatics,* 40(4), 646-675.

Josep Cuenca, M., Postolea, S., & Visconti, J. (2019). Contrastive Markers in Contrast. *Discours*, 25, 3-41.

Lakoff, R. 1971. If's, and's and but's about conjunction. In Charles J. Fillmore & D. Terence Langndoen (Eds.), *Studies in Linguistic Semantics* (pp. 3-114). New York: Holt.

Lee, K. 2020. Chinese ESL writers' use of English contrastive markers. *English Language Teaching,* 32(4), 89-110.

PACTE. 2008. First results of a Translation Competence Experiment: 'Knowledge of Translation' and 'Efficacy of the Translation Process'. In J. Kearns (Ed.), *Translator and Interpreter Training: Issues, Methods and Debates* (pp. 104-126). London: Continuum.

Perrez, J. 2006. Connectieven, Tekstbegrip en Vreemdetaalverwerving. Een studie van de impact van causale en contrastieve connectieven op het begrijpen van teksten in het Nederlands als vreemde taal (Doctoral dissertation, Université Catholique de Louvain).

Petit, M. 2010. Discrimination prosodique et représentation du lexique: les connecteurs discursifs. *Études de linguistique appliquée*, 1(157), 75-93.

Relative constructions in long-term immigrants from the UK in Spain

Cristina Suárez-Gómez & Pedro Guijarro-Fuentes – *University of the Balearic Islands*

This presentation explores the knowledge of relative clauses (RCs) in a group of adult native speakers of English who were born and raised in an English-speaking country, but who migrated to Spain more than 10 years ago ('long-term immigrants'). There is evidence that the variety of the home language they acquire differs from the variety of that language as native (L1) in the country of origin (e.g., Potowski 2018; see also Valdés 2000). Long-term immigrants allow us to answer questions concerning language structure and language change, since the variety of these speakers (Polinsky and Scontras 2020), is generally shaped by factors associated with the local ecologies in which these varieties developed (Mufwene 2014). Among these factors we find: (i) cognitive constraints of L2 learning and (ii) language contact among interacting varieties. Far from being opposing factors, these should be viewed as complementary (Thomason 2001: 62; Winford 2009: 226), and they may interact. These long-term immigrants are very likely to incorporate in their linguistic repertoire features which are not recorded in the original variety and may be the result of learning mechanisms, such as the preference for simpler and attrited structures (see also Scontras et al. 2017). Regarding language contact, and as a result of language coexistence, it is expected that in their variety there will be "complex patterns of contact linguistics, including […] discoursal and syntactic change and accommodation" (Bolton 2006: 261; see also Thomason and Kaufmann 1988). Therefore, we assume that in these situations the grammar of the resulting varieties is affected and undergoes changes in comparison with the languages that enter into the contact process.

Taking into account previous studies on the development of RCs by L1 and adult English speakers including adult heritage speakers, this study aims to answer the following research questions:

RQ1. Does the selection of relative markers differ in the long-term immigrants' varieties with respect to Standard English and/or other Englishes that emerged in similar environments?

RQ2. Is this resulting grammar influenced by Spanish grammar?

RQ3. Is this resulting grammar influenced by structural phenomena associated with language attrition, which violates the expected form-meaning alignment, as is the case of avoidance of empty elements (e.g., zero relatives) and redundant elements (e.g., *wh-* relativizers), with the aim of favoring [+transparent] structures?

Based on the results of a corpus-based study which compares native varieties of English (BrE) with L2 varieties of English (Asian Englishes) as represented in the *International Corpus of English* and the results of a Grammaticality Judgement Task (GJT) of 90 items from examples drawn from the corpus-based study, we observe that long-term immigrants do not match their PDE monolingual native speakers in their knowledge of RCs, but rather show results closer to World Englishes speakers, in that they prefer the relative marker *that* and *zero* (over *wh-*words), which reinforces the transparency principle operating in scenarios of language contact and language acquisition.

We interpret such an outcome as an indication that RCs seem to be affected by processes of degradation or interrupted competence that take place in heritage language acquisition; at the same time, our findings provide strong support for the claim that relativization might be affected by the number years of exposure to an L2 or lack thereof to the L1.

References

Bolton, Kingsley. 2006. World Englishes today. In Braj B. Kachru, Yamuna Kachru, & Cecil L. Nelson (eds.), *The handbook of World Englishes*, 240–269. Oxford: Blackwell.

International Corpus of English. http://ice-corpora.net/ice.

Mufwene, Salikoko S. 2014. Language ecology, language evolution, and the actuation question. In Tor Afarli & Brit Maelhum (eds.) *Language contact and change: Grammatical structure encounters the fluidity of language*, 13-35. Amsterdam: Benjamins.

Polinsky, Maria & Gregory Scontras 2020. Understanding heritage languages. *Bilingualism: Language and Cognition* 23. 4–20

Potowski, Kim (eds.). 2018. The handbook of Spanish as a heritage/minority language. London: Routledge.

Scontras Gregory, Maria Polinsky, C-Y. Edwin Tsai & Kenneth Mai 2017. Cross-linguistic scope ambiguity: When two systems meet. *Glossa: A Journal of General Linguistics* 2. 1–28.

Thomason, Sarah G. 2001. *Language Contact*. Edinburgh: Edinburgh University Press.

Thomason, Sarah G. & Terrence Kaufmann. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley & Los Angeles: University of California Press.

Valdés, Guadalupe. 2000. *Spanish for native speakers*. Vol. 1. New York: Harcourt College Publishers.

Winford, Donald. 2009. The interplay of 'universals' and contact-induced change in the emergence of New Englishes". In Markku Filppula, Juhani Klemola & Heli Paulasto (eds.), *Vernacular universals and language contacts: Evidence from varieties of English and beyond*, 206–230. London: Routledge.

***Sonderbar, auffällig* vs. *curioso, peculiar*. El uso del *Corpus PaGeS* para profundizar en el uso y significado de adjetivos alemanes y españoles**

Irene Szumlakowski-Morodo – *Complutense University of Madrid*

Keywords: *semántica adjetivos, lexicología contrastiva alemán-español, corpus paralelos bilingües, didáctica de lenguas.*

El aprendizaje del significado y uso del vocabulario, especialmente de adjetivos de campos tan específicos como los adjetivos calificativos referidos a fenómenos poco frecuentes o extraños, como los alemanes *auffällig, sonderbar, seltsam, kurios, rar, ungewöhnlich, komisch* o los españoles *curioso, raro, llamativo, extraño, peculiar, chocante, sorprendente,* supone un gran reto para los aprendices de lenguas extranjeras en niveles superiores, a partir del nivel B2 del MCERL. Tanto el uso de diccionarios monolingües como bilingües resulta insuficiente para alcanzar una precisión en el significado de cada lexema y conocer en detalle su uso. En esta contribución hacemos una propuesta de trabajo con los ejemplos buscados y seleccionados en el Corpus PaGeS, un corpus paralelo bilingüe alemán-español, que permita que los aprendices puedan por sí mismos agrupar los lexemas adjetivales de ambas lenguas, establecer correspondencias más precisas especialmente en los casos de lexemas polisémicos y evitar también casos de "falsos amigos". El objetivo final es demostrar con esta aportación una de las potencialidades didácticas del corpus PaGeS, con aprendices de lenguas extranjeras y con profesionales de la traducción.

El uso de corpus en la didáctica de las lenguas se va consolidando cada vez más, pese a las dudas que aún plantea en la práctica (Bubenhofer 2011; Stuyckens/Brône 2009). Pese a la importancia que pueden dar los docentes al trabajo empírico con datos reales de la lengua, plantea algunas dificultades que llevan a muchos profesores de lenguas extranjeras a resistirse al trabajo con corpus y a no motivar a sus alumnos en este empeño (Lamy/Klaskov 2012). Pero hay también ventajas innegables, el trabajo con corpus potencia la autonomía del alumno en su aprendizaje (Zanin 2011) y fomenta una «actitud investigadora» (Tolchinsky, 2014: 14). El trabajo con corpus permite ilustrar con ejemplos auténticos las reglas gramaticales y las cuestiones contrastivas y es útil para ejemplificar la evolución de la norma, pues docentes y alumnos analizan juntos los datos reales del uso del lenguaje (Stuyckens/Brône 2009, Tolchinsky 2014).

En un primer momento se delimitará el grupo concreto de estos adjetivos que califican fenómenos poco frecuentes en ambas lenguas. El siguiente paso es recopilar las definiciones existentes en diccionarios monolingües y bilingües, así como en algunas obras enciclopédicas. Con ellas se pueden establecer las correspondencias más frecuentes y los casos que pueden llevar a posibles errores: *rar-raro*, *kurios- curioso*, *komisch-cómico.*

A continuación, se realizarán búsquedas específicas en el Corpus PaGeS (Doval 2018), aprovechando las diferentes herramientas, que permiten, por un lado, incluir o excluir tipos de textos, lo que desvela algunos usos especializados; y, por otro, hacer búsquedas exactas de las diferentes equivalencias, es decir, cuándo se ha traducido un lexema adjetival por otro. Por ejemplo, encontramos en PaGeS diferentes equivalencias para el adjetivo *peculiar*: *sonderbar, komisch, außergewöhnlich, seltsam, eigenwillig, speziell, ungewöhnlich*. Y se encuentran 8 equivalencias del adjetivo alemán *auffällig* por el adjetivo español *extraño*, mientras que hay 69 ejemplos en los que *extraño* equivale a *ungewöhnlich*. De los 987 ejemplos de *komisch* que recoge el Corpus PaGeS, sólo 91 ejemplos equivalen al adjetivo español *cómico.* Encontramos con más frecuencia otras correspondencias, como los adjetivos *extraños* (213 ejemplos) o *raro* (242 ejemplos).

Partiendo de la información recopilada se hará una selección de ejemplos interesantes para trabajar con los aprendientes, que permitan sacar conclusiones de todos estos lexemas adjetivales. Una posibilidad añadida es ofrecer vías concretas para que sean los alumnos quienes hagan directamente búsquedas en el corpus PaGeS.

Bibliografía

Bubenhofer, N. 2011. "Korpuslinguistik in der linguistischen Lehre: Erfolge und Misserfolge", *Journal for Language Technology and Computer Linguistics* 26/1, 141- 156.

Cartagena, N. / Gauger, H.-M. 1989. *Vergleichende Grammatik Spanisch - Deutsch.* Mannheim: Duden.

Doval, Irene 2018. "Corpus paralelos en la enseñanza de lenguas extranjeras: un ejemplo de aplicación basado en el corpus PaGeS." *CLINA*, vol. 4-2, 65-82. DOI: https://doi.org/10.14201/clina2018426582

Dudenredaktion s.a.: *Duden online.* URL: https://www.duden.de/woerterbuch. [enero 2023]

DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart, Berlin-Brandenburgischen Akademie der Wissenschaften. URL: https://www.dwds.de [enero 2023]

Lamy M-N. / Klarskov Mortensen H. J. 2012. «Using concordance programs in the Modern Foreign Languages classroom», Módulo 2.4 en *Information and Communications Technology for Language Teachers* (ICT4LT), ed. G. Davies, Slough, Thames Valley University, URL: http://www.ict4lt.org/en/en_mod2-4.htm [julio 2022].

PONS online Wörterbuch Deutsch-Spanisch. Pons Langenscheidt GmbH, Stuttgart. URL: *https://de.pons.com/-Übersetzung* [enero 2023]

Real Academia Española s.a.: *Diccionario de la lengua española*, 23.ª ed., [versión 23.6 en línea]. URL: https://dle.rae.es. [enero 2023]

Slaby, R., Grossmann, R. 1993. Diccionario de las lenguas española y alemana I. Español-Alemán. Barcelona: Herder.

Slaby, R., Grossmann, R., Illig, C. 1991. Diccionario de las lenguas española y alemana II. Alemán-Español. Barcelona: Herder.

Sommerfeldt, K-E., Schreiber, H. 1974. *Wörterbuch zur Valenz und Distribution deutscher Adjektive.* Leipzig: VEB Bibliographisches Institut.

Stuyckens, G. / Brône, G. 2009. "Brauchbarkeit von Korpora des geschriebenen Deutsch für DaF-Lehrende. Eine Fallstudie", *Deutsch als Fremdsprache* 46,1, 3-9.

Tolchinsky, L. 2014. "El uso de corpus lingüísticos como herramienta pedagógica", *Textos de didáctica de la lengua y de la literatura,* 65, 9-17.

Zanin, R. 2011. "Korpusinstrumente im Umkreis des Lernens", en Abel, A./ Zanin, R. (eds.), *Korpora in Lehre und Forschung*, Bozen, Bozen University Press, 103-128. DOI: 10.13124/9788860460950.

A variationist approach to subject omission: Null and overt subjects in eight varieties of English

Iván Tamaredo Meira – *Complutense University of Madrid*

Keywords: *null subjects, subject omission, web-based language, GloWbE, World Englishes, variationist linguistics.*

English is a language in which the subjects of finite clauses are obligatory (Dryer 2013). However, as shown by recent research (AUTHOR A, Schröter and Kortmann 2016) and by *The Electronic World Atlas of Varieties of English* (Kortmann et al. 2020), referential and non-referential subject omission, as in examples (1) and (2), respectively, do occur in this language and are in fact much more common than previously thought, particularly in non-standard varieties (e.g. AUTHOR B; Schröter 2019).

(1) No oil showing so he put 3 litres in to bring it up to half way on the dipstick. **Drove** to Richmond and the next morning he checked the oil and found it way over full. (*GloWbE*, AU B, exploroz.com)

(2) **Seems** the political leader who is responsible for the father's death is trying to pass legislation to legalize ganja […]. (*GloWbE*, JM G, jamaicans.com)

The present contribution adopts a variationist approach and examines the occurrence of null subjects, thus comparing them with the overt subject variant, in a set of eight varieties of English in different evolutionary stages of Schneider's (2007) Dynamic Model. Referential and non-referential overt subject pronouns are illustrated in (3) and (4), respectively.

(3) **He** goes to bed confident of success and without any worry of "disappointing" us should a different outcome occur. (*GloWbE*, CA G, drykids.info)

(4) **It** seems that his call to save Pakistan is irrelevant and is for me sadly indicative of the criminal apathy that Pakistan and her citizens are sadly synonmous with. (*GloWbE*, PK B, blog.otherpakistan.org)

Data was retrieved from the *Corpus of Global Web-Based English* (*GloWbE*; Davies 2013), the largest corpus available for the study of language variation with 1.9 billion words from 20 English-speaking countries (Davies and Fuchs 2015). In particular, the Australian, Bangladeshi, Canadian, Indian, Jamaican, Nigerian, Pakistani, and Singaporean national components of *GloWbE* were investigated. A random sample of 3rd person null and overt pronominal subjects in initial clause position was extracted from the corpus and annotated for the following factors:

- Variety: stage in the Dynamic Model and L1 vs. L2 status.
- Verb: verb tense and verb semantics.
- Genre: general websites vs. blogs.
- Referential status: referential vs. non-referential null subjects.
- Pronominal form: s/he vs. it.
- Referential (dis)continuity: reference maintenance from previous clause vs. partial switch vs. full switch.
- Persistence: previous subject is null vs. a pronoun vs. a full noun phrase.

The alternation between null and overt pronominal subjects was examined across the levels of the aforementioned extra- and intra-linguistic variables by means of predictive modelling and dimensionality reduction techniques (e.g. Levshina 2015). Preliminary results suggest that null subjects are more frequent in varieties that are more advanced in Schneider's Dynamic Model than in those in earlier phases of development. In addition, extra-linguistic variables do not account for the distribution of null and overt subjects, so intra-linguistic variables, in particular, persistence, pronominal form and reference, are the strongest determinants of

null subjects. Therefore, this study sheds light on the distribution of null and overt pronominal subjects in World Englishes and on the factors that have an influence on the choice between these competing variants.

References

Davies, Mark. 2013. Corpus of Global Web-Based English.

Davies, Mark and Robert Fuchs. 2015. "Expanding Horizons in the Study of World Englishes with the 1.9 Billion Word Global Web-based English Corpus (GloWbE)." *English World-Wide* 36 (1): 1–28.

Dryer, Matthew S. 2013. "Expression of Pronominal Subjects." In Matthew S. Dryer and Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Kortmann, Bernd, Kerstin Lunkenheimer, and Katharina Ehret, eds. 2020. *The Electronic World Atlas of Varieties of English*. Zenodo.

Levshina, Natalia. 2015. How to Do Linguistics with R: Data Exploration and Statistical Analysis. Amsterdam/Philadelphia: John Benjamins.

Schneider, Edgar. 2007. *Postcolonial English: Varieties Around the World*. Cambridge: Cambridge University Press.

Schröter, Verena. 2019. Null Subjects in Englishes: A Comparison of British English and Asian Englishes. Berlin/Boston: De Gruyter.

Schröter, Verena, and Bernd Kortmann. 2016. "Pronoun Deletion in Hong Kong English and Colloquial Singaporean English." *World Englishes* 35 (2): 221–241.

A corpus-based account of resonance and engagement in Chinese doctor-patient interaction: A western vs traditional Chinese medicine divide

Vittorio Tantucci[1] & Carmen Lepadat[2] – *Lancaster University[1] – Roma Tre University[2]*

Resonance in interaction involves speakers/writers re-using (parts of) the utterances of their interlocutors (Du Bois 2014). When resonance is creative, speakers/writers engage with other people's language to express something new. Persistent creative resonance is a key indicator of interactional engagement. Conversely, consistent absence of it underpins interactional detachment, which is distinctive of ASD speech (Author(s) 2022a). This study tackles creative resonance in Chinese-doctor patient interaction. We sourced 60 conversations, including 1415 utterances from the Chunyu Yisheng platform of online medical consultations and compared the speech of Western medicine doctors (WMD) with Traditional Chinese medicine doctors (TCMD).

Doctors' engagement is central in medical communication, with focus on rapport building, small talks, and patient-centeredness (Jin et al 2022). Palliative care textbooks refer to 'active listening', backchanneling and repetitions of patients' words (Jenkins et al. 2021). We propose that creative resonance is key for disease assessment, as it is not limited to repetition, but involves engagement as joint creation of knowledge (Author(s) 2002b) between specialist and patient. We provide a multifactorial analysis of medicine type, gender, turns and words' count, peripheral particles of intersubjectivity, overt acknowledgement of interlocutors' speech, illocutionary force and creative resonance. A mixed effects linear regression showed that TCMD's textual engagement with patients is higher than in WMD. Creative resonance in TCMD's speech correlates with overt relevance acknowledgment and marked intersubjectivity. We also discovered that TCMD's language is inherently directive, whilst WMD's speech is more assertive. This suggests that TCM is distinctively lifestyle-oriented and geared towards advice-giving (Yip 2020). Conversely, WM favours etiological assessment and following prescription. TCM involves a holistic approach to the body with its social and natural environment (Lu et al. 2004:1854), with stronger emphasis on harmonious interaction (Spencer-Oatey 2005). This is reflected in the pragmatics of TCMD in contrast with WMD.

References

Du Bois, J. W. 2014. Towards a dialogic syntax. *Cognitive linguistics,* 25(3), 359-410.

Jenkins, L., Parry, R., & Pino, M. 2021. Providing opportunities for patients to say more about their pain without overtly asking: A conversation analysis of doctors repeating patient answers in palliative care pain assessment. *Applied Linguistics,* 42(5), 990-1013.

Jin, Ying; Younhee Kim, Andrew P Carlin 2022. Co-Topical Small Talk: Troubles-Telling in Traditional Chinese Medical Encounters. *Applied Linguistics*, Volume 43, Issue 3, June 2022, Pages 493-516, https://doi.org/-10.1093/applin/amab057.

Lu, A. P., Jia, H. W., Xiao, C., & Lu, Q. P. 2004. Theory of traditional Chinese medicine and therapeutic method of diseases. *World journal of gastroenterology: WJG,* 10(13), 1854.

Spencer-Oatey, Helen. 2005. (Im)Politeness, face and perceptions of rapport: unpackaging their bases and interrelationships. *Journal of Politeness Research* 1(1), 95-119.

Yip, J. W. 2020. Directness of advice giving in traditional Chinese medicine consultations. *Journal of Pragmatics* 166, September 2020, Pages 28-38.

How the pragmatics of engagement is changing in British English interaction: Resonance across generations in the BNC1994 and the BNC2014

Vittorio Tantucci[1] & Aiqing Wang[2]

Lancaster University[1] – University of Liverpool[2]

Resonance in interaction underpins speakers/writers re-using (parts of) the utterances of their interlocutors (Du Bois 2014). When resonance is creative, speakers/writers engage with other people's language to express something new. Persistent creative resonance is a key indicator of interactional engagement and reciprocity at talk (Author(s) 2002a). Conversely, consistent absence of it underpins interactional detachment, which is distinctive of ASD speech (Author(s) 2002a; Du Bois et al. 2022b). This study tackles creative resonance in naturalistic interaction across different generations of British speakers (15-44 vs over 60) in the years 1994 and 2014. We controlled for gender, context and intra and inter-generational exchanges, analysing 1200 turns at talk. We found that older generations after 2014 show a significant increase in the way they creatively 'resonate' with their interlocutors. This is remarkably evident for female speakers. What this entails is that, different from what is often assumed, older generations creative engagement at talk is remarkably higher than how it used to be 20 years before, whilst younger speakers do not show a significant increase in the way they creatively resonate with their peers. We also discovered that class is a key factor for the rise of interactional engagement from 1991 to 2014, as BNC social categories 'C' and 'D' show significantly higher levels of resonance than upper class populations, in turn tagged as 'A' and 'B'. We finally found that interaction characterised by resonance after 2014 has become less diverse in terms of illocutionary force, with assertions becoming by far the most frequent speech act.

All in all, this large-scale study suggests that British interaction underwent a significant increase of dialogic engagement and creativity in older generations and lower social class populations. This may indicate that, one the one hand, older speakers may be less detached from topics considered 'at issue' from other generations, perhaps due to a facilitated access to technological media of information. On the other hand, this trend may also show an increase of literacy level of populations from lower social strata in the transition from 1991 to 2014. This study finally indicates that assertions are by far the preferred form of speech act that used when engagement is at play. This correlation has been growing significantly in the BNC2014.

References

Du Bois, J. W. 2014. Towards a dialogic syntax. *Cognitive linguistics*, 25(3), 359-410.

Du Bois, J. W., Hobson, R. P., & Hobson, J. A. 2014. Dialogic resonance and intersubjective engagement in autism. *Cognitive Linguistics*, 25(3), 411-441.

Urban stigmas: A corpus-assisted discourse study

Jenny Tarvainen – *University of Jyväskylä*

Keywords: *corpus-assisted discourse studies, stigmas, discourse prosody, NLP, segregation.*

In this presentation, I will discuss stigmas related to neighborhoods of Helsinki, the capital of Finland, in social media discussions. I define stigmas as negative representations that are (re)created by discourses (see Hall 1999). In this study, discourses are defined according to Fairclough (2003): discourses reflect the world view and ideologies of the language user and they reflect and reconstruct the power relations of the society. A negative representation may cause stigmatization that complicates the residents' lives and the positive development of the neighborhood (Wacquant 2008). Stigmatization can occur when a neighborhood is discriminated against due to their perceived features such as socioeconomic or ethnic composition (Musterd et al. 2008; Wacquant 2008). Stigmatization may lead to segregation, which means a strong differentiation between neighborhoods (THL 2020). I will analyze what kinds of discourses are related to these neighborhoods to grasp a better understanding of the discursive processes that are related to socio-spatial segregation.

For the research data, I will use a corpus compiled from one of the most visited social media platforms in Finland, Suomi24 ('Finland24'; City Digital Group 2021), which has more than 2.1 million visitors per month (FIAM). The corpus contains circa 4 billion words including all the discussions on the platform between years 2001 and 2020 (Meta-Share). This kind of authentic language usage data is exceptionally large. Therefore it is optimal for corpus assisted discourse studies (CADS), which has emphasis on the frequently occuring phenomena.

I will study the stigmas using CADS, since it combines quantitative and qualitative approaches (see Partington, Duguid & Taylor 2013) – both statistical evidence and deep understanding of the research object. I will complete a collocation analysis (see e.g. Sinclair 1991) of two neighborhoods with a positive reputation and of two neighborhoods with a negative reputation. These collocates will then be grouped to semantic categories which form the base for discourse analysis. A closer examination is carried out by studying discourse prosodies, which in this study are recurring associations between the place names and the related semantic groups (see Stubbs 2001, Hoey 2005). The final discourse analysis consists of close reading of these semantic groups' cotexts.

I have completed a preliminary sentiment analysis of the Helsinki neighborhood names using built-in polarity annotation of the corpus. It seems that the negative contexts are more dominant than positive contexts in general, neutrals being the most frequent. However, one neighborhood can be discussed both in a negative and in a positive manner. The next step will be to choose two neighborhoods, one with mainly positive priming and one with a negative one, for a closer examination. I will choose these neighborhoods based on this preliminary study and on previous studies on the matter (e.g. Erola, Kallio & Vauhkonen 2017).

This presentation relates to a research paper in progress that will serve as a part of my doctoral dissertation at the University of Jyväskylä. The study combines corpus linguistics, natural language processing (NLP), and urban studies. The societal aim of the broader study is to reveal the linguistic mechanisms of socio-spatial segregation, and the academic aim is to utilize a set of methods to complete an automatic semantic analysis of linguistic big data.

Bibliography

City Digital Group 2021. *The Suomi24 Sentences Corpus 2001–2020*, Korp version [text corpus]. Kielipankki [Language Bank of Finland].

Erola, J., Kallio, J. & Vauhkonen, T. 2017. *Ylisukupolvinen kasautuva huono-osaisuus Turussa ja muissa Suomen suurissa kaupungeissa.* [Intergenerational accumulation of social disadvantages across generations in Turku and other big cities of Finland]. Turun kaupunki.

Fairclough, N. 2003. Analysing discourse: textual analysis for social research. London: Routledge.

FIAM. Finnish Internet Audience Measurement: *Tulokset* [Results]– Comscore MMX MP. Accessed 16 June 2020. Retrieved from https://fiam.fi/tulokset/

Hall, S. 1999. *Identiteetti* [Identity]. Tampere: Vastapaino.

Hoey, M. 2005. Lexical Priming. A New Theory of Words and Language. London: Routledge.

Meta-Share. *The Suomi 24 Sentences Corpus 2001-2020.* http://urn.fi/urn:nbn:fi:lb-2021101525 Accessed Jan 13th 2023.

Musterd, S., Andersson, R., Galster, G., & Kauppinen, T. 2008. Are immigrants' earnings influenced by the characteristics of their neighbours? – Environment and Planning A, 40 (4) s. 785–805.

Partington, A., Duguid, A. & Taylor, C. 2013. *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS).* Studies in Corpus Linguistics 55. Amsterdam: John Benjamins.

Sinclair, John 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, Michael 2001. Words and phrases. Corpus studies of lexical semantics. Oxford: Blackwell Publishers.

THL = Terveyden ja hyvinvoinnin laitos [Finnish institute for health and welfare] 2020. *Yhdyskuntasuunnittelu* [Urban planning].

Wacquant, L. 2008. Urban outcasts. A comparative sociology of advanced marginality. Malden: Polity Press.

From the corpus DISTRIBUCOR to the dictionary DISTRIBUDICC: Creating a specialized dictionary for translators using *Sketch Engine*, *Lexonomy*, and *OneClick Dictionary*

Cristina Toledo-Báez – *University of Málaga*

Despite the improvements in Lexicography and Terminography and the vast amount of specialized dictionaries, translators still face a major obstacle when dealing with specialized translation: the lack of appropriate terminological and lexicographical resources that meet their needs and enable them to acquire expert knowledge. In response to the scarcity of terminological resources, translators are compelled to create their own material using, a task for which they need to make use of the latest lexicographical tools.

Following in the footsteps of Bartolomé-Díaz & Frontini's work (2020), the aim of our study is to create a specialized bilingual (English-Spanish) dictionary about exclusive distribution agreements using two main elements: on the one hand, the dictionary writing system Lexonomy (Měchura, 2017) and the function OneClick Dictionary (Jakubíček et al., 2021), both available at the lexicographical and corpus manager Sketch Engine (Kilgarriff et al., 2014) and, on the other hand, the bilingual (English-Spanish) comparable virtual corpus DISTRIBUCOR (Authors, 2021). This corpus of exclusive distribution agreements contains contracts written in the diatopic varieties of English from United States of America and Spanish from Spain. In addition, DISTRIBUCOR was semiautomatically compiled using Sketch Engine and its representativeness was determined thanks to the ReCor tool (Seghiri, 2017). The result of our study is the dictionary DISTRIBUDICC, a specialized bilingual 250-entry dictionary containing the six following grammatical categories: nouns (179), different types of phrases (26), verbs (17), adjectives (12), prepositions (12) and adverbs (4).

The combination of a virtual comparable bilingual corpus, Sketch Engine, Lexonomy and OneClick Dictionary helps to create a terminological resource, specifically a specialized dictionary, which is quick, at zero cost and of great quality. Its use is intended to be beneficial for Translation and Interpreting lecturers, students, and practitioners as well as for linguists, lawyers and international trade experts.

References

Bartolomé-Días, Z & Frontini, F. 2020. Building a domain-specific bilingual lexicon resource with Sketch Engine and Lexonomy: Taking Ownership of the Issues. In I. Kernerman, S. Krek, J. P. McCrae, J. Gracia, S. Ahmadi & B. Kabashi. (Eds.), *Proceedings of the Globalex Workshop on Linked Lexicography. Language Resources and Evaluation Conference (LREC 2020)* (pp. 62-68). Marseille: European Language Resources Association.

Jakubíček, M., Kovář, V. & Rychlý, P. 2021. Million-Click Dictionary: Tools and Methods for Automatic Dictionary Drafting and Post-Editing. In Z. Gavriilidou, L. Mitits, L. & K. Spyros (Eds.), *EURALEX XIX. Congress of the European Association for Lexicography. Lexicography for Inclusion. Book of Abstracts* (pp. 65-67). Komotini: SynMorPhoSe Lab, Democritus University of Thrace.

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

Měchura, M. 2017. Introducing Lexonomy: An Open-Source Dictionary writing and publishing system. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (Eds.), *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference* (pp. 19-21). Brno: Lexical Computing.

Seghiri, M. 2017. Metodología de elaboración de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus para la traducción de manuales de instrucciones de televisores. *Babel,* 63(1), 43-64.

Toledo Báez, Cristina 2021. Compilación semiautomática con Sketch Engine de un corpus ad hoc comparable bilingüe (inglés-español) de contratos de distribución exclusiva (DISTRIBUCOR). In E. Sartor (2021), *Los corpus especializados en la lingüística aplicada: enseñanza y traducción* (pp. 15-41). Verona: Colección Pliegos Hispánicos.

Is fake news more evaluative? Comparing appraisal expressions across fake and genuine news in English

Radoslava Trnavac[1] & Nele Põldvere[2] – *National Research University Higher School of Economics[1] – University of Oslo[2]*

Keywords: *disinformation, linguistic cues of fake news, appraisal theory, qualitative and quantitative analysis.*

Fake news has become an important topic of research in a variety of disciplines including corpus linguistics. The main strength of corpus approaches to fake news is that we have access to carefully designed datasets of both fake and genuine news, which can be used to identify salient linguistic features of fake news (Asr and Taboada, 2019). Moreover, the choice of the linguistic features is grounded in existing linguistic theory, which can inform our analyses in important ways (Grieve and Woodfield, forthcoming).

The goal of this paper is to compare the use and distribution of evaluative features across fake and genuine news in English using qualitative and quantitative corpus linguistic techniques. The qualitative analysis, grounded in Appraisal Theory—ATTITUDE, ENGAGEMENT and GRADUATION (Martin and White, 2005)—sheds new light on how and why evaluative meanings are construed differently in fake news compared to genuine news. The Appraisal expressions were identified manually in the corpus texts, followed by classification into the Appraisal categories and their subcategories (e.g., CONTRACTIVE vs. EXPANSIVE functions of ENGAGEMENT). The quantitative analysis involved the implementation of a chi-square test to measure the association between the evaluative features, on the one hand, and fake and genuine news, on the other.

The corpora were extracted from two sources. The first corpus is based on the writings of The New York Times journalist Jayson Blair, who was fired from the newspaper for fabricating news stories in the early 2000s (Grieve and Woodfield, forthcoming). The corpus contains approximately 56,000 words, with samples of both fake and genuine news. The second corpus is based on the writings of seven other journalists who have been identified by various fact-checking services to have written both fake and genuine news. This corpus contains approximately 13,000 words from online news websites. On the one hand, the corpora are similar in the sense that both are controlled for well-known confounding variables such as authorship. On the other hand, they differ in terms of genre (mainstream newspaper vs. alternative news websites) and the journalists' motivation to lie. While Blair's motivation was personal, the other journalists' motivation was more ideological. In this way, the corpora contain different types of fake news, which may or may not lead to differences in terms of evaluation.

The preliminary results showed interesting differences between the two corpora, and between fake and genuine news. So, is fake news more evaluative? This is certainly the case in the Jayson Blair corpus. Moreover, Blair's fake news is characterized by a greater use of ATTITUDE, which might be due to his lack of access to basic information about the news events (the *who*, *what*, *when* and *where* of journalism), since he never attended them. These results are slightly different in the other corpus, where fake news is also more evaluative, but where the key Appraisal category is ENGAGEMENT instead. In particular, the journalists seem to draw heavily on the CONTRACTIVE function of ENGAGEMENT to fend of alternative viewpoints and to present their ideology as 'true' and 'valid'. Thus, evaluation provides important cues for fake news detection, but more work is required to understand the effect of genre and motivation on the linguistic outcomes.

References

Asr, Fatima and Maite Taboada. 2019. Big Data and quality data for fake news and misinformation detection. *Big Data & Society*. https://doi.org/10.1177/2053951719843310.

Grieve, Jack and Helena Woodfield. forthcoming. *The language of fake news*. Cambridge Elements in Forensic Linguistics. Cambridge University Press.

Martin, Jim and Peter White. 2005. *The language of evaluation: Appraisal in English*. Palgrave Macmillan.

A corpus-assisted discourse analysis of vague language in sustainability reports

Syamimi Turiman & Siti Aeisha Joharry – *Universiti Teknologi MARA*

Keywords: *sustainability reports, vague language, corpus-assisted discourse analysis, corporate reporting.*

There has been growing interest in studying corporate discourse in recent years (e.g. Kapranov 2016; Jin, 2022). Sustainability reporting is the disclosure and communication of environmental, social and governance goals and the company's progress toward achieving them. Among the benefits of sustainability reporting are improved corporate reputation and consumer confidence. To achieve these benefits, vague language can be used to strategically communicate with their stakeholders and the public. While it is acknowledged that vague language has been studied in terms of its usage and communicative function, there is a need to look into how it is used in sustainability reports, given the attention by many large corporations towards achieving the Sustainable Development Goals (SDGs). This study examines the vague language employed in sustainability reports, taking as a specific case the non-financial reports produced by Petroliam Nasional Berhad (PETRONAS), Malaysia. A total of 10 sustainability reports published between 2007 and 2018 were collected and compiled into a corpus. In this paper, the use of vague language in the sustainability reports is explored via a corpus-assisted approach. Using #Lancsbox 6.0, a list of frequent words is generated to first identify the expressions related to vague language based on the list/items identified in the literature (e.g. Cheng & O'Keeffe, 2015; Li, 2017; Jin 2022). Following this, a close inspection of the concordance lines of the identified vague language was done to observe the context where vague language occurs, as well as to identify the functions for employing it in the sustainability reports. Findings revealed that there are three major types of vague language in the sustainability reports, namely quantity (e.g. *more than*, *many, various*), degree (e.g. *important, significant, effective*), and time (e.g. *early, recent, often*). Moreover, the use of vague language is associated with communicative functions such as to enhance persuasion and give the right amount of information. From the point of view of corpus linguistics and grammar, this study demonstrates how the use of corpus linguistics techniques complements the examination of vague language in sustainability reporting.

References

Cheng W. & O'Keeffe, A. 2015. Vagueness. In K. Aijmer & C. Rühlemann (eds.), Corpus pragmatics: A handbook (pp. 360-378). Cambridge University Press

Jin, B. 2022. A corpus-assisted study of vague language in corporate responsibility reports of the cosmetics industry. *Ibérica, 43*, 77-102.

Kapranov, O. 2016. Corpus Analysis of Discourse Markers in Corporate Reports Involving Climate Change. In A.M. Ortiz & C. Pérez-Hernández (eds). *CILC2016. 8th International Conference on Corpus Linguistics, Vol 1*, pages 216 -227.

Li, S. 2017. A corpus-based study of vague language in legislative texts: Strategic use of vague terms. *English for Specific Purposes, 45*, 98-109.

A Fala: Corpus-based minority language grammar

Miroslav Vales – *Technical University of Liberec*

Keywords: *A Fala, minority language, grammar description, diatopic variation, primary data.*

When we talk about corpora, we frequently think of large sets of data containing tens or even hundreds of millions of tokens. However, in the field of minority and less described languages we do not have these tools and even a small corpus is sometimes a great achievement. The objective of the presentation is to describe a corpus and a *Database of A Fala* (Valeš 2021a), a minority language of Extremadura, Spain, and its usage for a new project of a grammatical description of the language. The corpus was created for the purpose of lexicographical description and it resulted in the first *Dictionary of A Fala* (Valeš, 2021b). The corpus contains 225,000 tokens documented in 156 texts. It has been compiled from transcribed recordings, which contributed with 49%, and published and unpublished texts written in one of the three varieties of A Fala, which contributed with the remaining 51%.

We can find partial grammatical descriptions of A Fala in Frades Gaspar (2000), Rey Yelmo (1999), Costas González (1992), Álvarez Pérez (2014) or Castro Piñas (2016). These descriptions include phonology, lexicology, some aspects of morphology, nevertheless, they either concentrate on one grammatical issue or they consider only one of the three varieties of the language, and in general, they are quite unreliable. For this reason, the grammar of the language remains undescribed.

The aim of the current project is to enlarge the original *Database* (Valeš 2021a) with a new set of recordings and to use it for the first complex grammatical description of the language. The focus of the grammar will be non-prescriptive in order to display the rich diatopic variation. The language has three varieties, according to the three villages where it is spoken, and the corpus enables identification of the phenomena related to each variety and it also registers social stratification. For example, the definite article plurals are different for each variety and in one of them we can find two competing forms. The grammar will make use of the corpus to expose which of the competing forms is more frequent and if there are some further conditions for the selection of each.

As half of the corpus comes from recordings, it offers rich data for phonological analysis. The drive engine, FLEx database, is semi-directly connected with the real recordings transcribed in ELAN, and for this reason it is easy to find examples and verify sounds and phonemes that are specific of A Fala; for example, the system of sibilants, which is more complex in comparison with Spanish. The oral part of the corpus is large enough to examine this set of phonemes in detail, including their diatopic and diastratic stratification.

The phonological and grammatical description and the enlargement of the corpus should help this minority language to create further resources, for example, didactic materials, because up to now A Fala remains an extremely under-resourced language.

Bibliography

Álvarez Pérez, Xosé Afonso. 2014. Correspondencias léxicas entre A Fala de Cáceres e o portugués. *Estudos de Lingüística Galega*, vol. 6, p. 5-27.

Castro Piñas, Fortunato. 2016. Más noticias sobre el pronombre enclítico al participio en la lengua del valle de Jálama o Xálima. *Limite: Revista de Estudios Portugueses y de la Lusofonía*, vol. 10, no. 1. p. 41-62.

Costas González, Xosé-Henrique. 1992. Breve caracterización das Falas do val do río das Ellas. *Cadernos de lingua*, vol. 6, p. 85-107.

Frades Gaspar, Domingo. 2000. *Vamus a falal*. 2nd edition. Mérida: Editora regional de Extremadura.

Rey Yelmo, Jesús. 1999. *La Fala de San Martín de Trevejo: O mañegu*. Mérida: Editora regional de Extremadura.

Valeš, Miroslav. 2021a. *A Fala Database: version 02, Sep. 2021*. Minde: CIDLeS. Available at: http://cidles.eu/projects/fala-outputs/

Valeš, Miroslav. 2021b. *Diccionariu de A Fala: lagarteiru, mañegu, valverdeñu*. Minde: CIDLeS. Available at: http://cidles.eu/projects/fala-outputs/

Literatura y revolución: aproximación a la literatura cubana post-1959 desde la lingüística de corpus

Danilo Orlando Vargas Nardiz – *Universidad Autónoma de Querétaro*

Palabras clave: *análisis de agrupación jerárquica, Cuba, lectura distante, migración.*

La literatura, como cualquier otra forma de arte, es parte de la identidad de una nación. En ella se recogen las costumbres, puntos de vista, temores y esperanzas de un pueblo. Gracias a los libros somos capaces de conocer y tal vez, entender la forma de vida de una determinada sociedad en una época determinada. La literatura se nutre del imaginario colectivo y los autores, por lo general y de diversas maneras, expresan sus puntos de vista particulares sobre lo que sea que este aconteciendo en esa sociedad, en ese momento.

Esta investigación emplea metodología propia de la lingüística de corpus para plasmar el enfoque de *lectura distante* propuesto por Franco Moretti (2013). Este enfoque se utiliza para analizar un corpus de narrativa cubana contemporánea. Así, el objetivo de este trabajo es identificar diferencias temáticas en la narrativa cubana previa a 1959 y la posterior a esta fecha. Se toma como referencia 1959 porque es el año del triunfo de la revolución cubana. Para ello se diseñó un corpus de 45 textos de narrativa (cuento y novela) cubana contemporánea, que se estudió con técnicas propias de la lingüística de corpus.

El Corpus de Narrativa Cubana Contemporánea (CNCC) tiene 4,956,173 de casos y 169,345 tipos. Para el estudio se decidió dividir el corpus en dos secciones: una contiene 19 textos (correspondiente a los textos publicados antes de la revolución) y otra con los restantes 26. La selección de obras y autores se llevó a cabo tras consultar con un panel de expertos en literatura cubana. Solo uno de los 23 autores revisados, Alejo Carpentier, contribuye a ambas secciones del corpus.

Se realizaron varios análisis, de corte cuantitativo y cualitativo en cuatro fases. En primer lugar, se realizaron tres análisis de palabras clave: cada una de las dos secciones se contrastaron con el CREA y posteriormente entre ellas. Después, estas palabras clave se agruparon manualmente para identificar los temas característicos de cada período. Así, se identificaron temas como la migración, la etnicidad, la tierra, la revolución y el género. Uno de los hallazgos más interesantes resultó ser la aparición del tema *migración* durante el segundo período, el cual no fue de interés para los escritores cubanos del período prerrevolucionario. Esto parece estar condicionado por los cambios sociales económicos y políticos que se generaron a partir de 1959. A continuación, a partir de las palabras clave vinculadas al fenómeno de la migración, se llevó a cabo un análisis cualitativo de líneas de concordancia y se identificaron una serie de sentimientos –no siempre positivos–: ansiedad, miedo, incertidumbre, alivio y tristeza, entre otros. Finalmente, realizó un análisis de agrupación jerárquica (Cantos Gómez, 2013) que permite evaluar cuantitativamente qué tan bien discriminan estos sentimientos las dos secciones del corpus. El dendrogama generado señala unos resultados muy prometedores por la precisión en la clasificación de los textos.

Todo esto confirma la pertinencia de combinar métodos propios de la lingüística de corpus con enfoques propios de los estudios literarios y culturales.

Bibliografía

Cantos Gómez, P. 2013. *Statistical Methods in Language and Linguistic Research*. Equinox Pub. Limited.

Moretti, F. 2013. *Distant reading*. Verso Books.

Creation and application of a self-built corpus in teaching Chinese as a foreign language to Spanish teenagers: A case study

Lili Wang & María Teresa Cáceres-Lorenzo – *University of Las Palmas de Gran Canaria*

Keywords: *self-built corpus, AntConc, teaching Chinese as a foreign language; Spanish teenagers.*

As computer technology continuously progresses, corpus-based investigation is increasingly used in the language classroom. Studies have found that corpora can be useful and directly applicable in L2 classrooms (Hunston, 2002; Aston, 2004; Breyer, 2009; Kennedy, 2014; Friginal & Cox, 2022). Corpus-based research and its application are critical signs of language digitization. In teaching Chinese as a foreign language, the research on classroom teaching through corpus is still at the initial stage (Cui & Zhang, 2011).

The overall objective of this work is the following: create a corpus using the corpus tool *AntConc* for Spanish teenagers who study Chinese as a foreign language at the Confucius Institute of the University of Las Palmas de Gran Canaria (CI-ULPGC) and apply it to the classroom to improve their ability to use vocabulary at level A2.

To this end, we have to answer the following research questions: What are the main theoretical frameworks within the European context that support the creation of a Chinese corpus for secondary school students? How do Chinese teachers use the corpus tool *AntConc* to build a corpus for the classroom? Does our Chinese self-built corpus help to improve our students' communication skills?

Common European Framework of Reference for Language (CEFR, Council of Europe, 2001), ICT Competency Standards for Teachers (UNESCO, 2008), *Common Digital Competence Framework for Teachers (CDCFT, INTEF, 2017)* and DDL (data-driven learning) are the main documents used as theoretical references for this work. The CEFR is the framework document that shows the general language policy project within the European context; the document of UNESCO in 2008 stipulates that teachers and students must utilize technology effectively for today's classroom; the CDCFT is a working document for the digital training of teachers, which describes the different levels of essential knowledge, skills, and attitudes necessary to be digitally competent (A1-C2), and DDL refers to an approach to learning vocabulary with linguistic data; DDL has become the inevitable choice of teaching development in the new era (Johns, 1991; Qi & Zhang, 2022). In addition, our research used an investigation-action methodology divided into the following steps: planning, action, observation, and reflection (Vidal Ledo & Rivera, 2007).

Our study subjects were 42 middle school students in 2015-2016. The process of our investigation includes the following: the need to improve the YCT vocabulary (Muñoz, 2014; Cáceres-Lorenzo, 2015; Griffiths, 2015); the creation of *Corpus of Chinese for Spanish Students (CCSS)* with AntConc (Laurence, 2023); the creation of classroom activities through self-built corpus CCSS; the initial evaluation of our subjects with a pretest; the creation of activities and their applications in the classroom; a corresponding posttest that indicates the degrees of improvement; and finally, the analysis of the results.

The results showed a significant improvement between the pretest and posttest and also led to the feasibility of using a self-built corpus to improve communication skills at the YCT-A2 level. According to our case study, we hope this research is an empirical contribution to teaching Chinese as a foreign language through self- built corpus.

Bibliography

Anthony, L. 2023. *Laurence Anthony's AntConc.* https://www.laurenceanthony.net/software/antconc/

Aston, G., Bernardini, S. & Stewart, D. 2004. *Corpora and language learners*. Amsterdam: John Benjamins Publishing Company.

Breyer, Y. 2009. Learning and teaching with corpus: Reflections by student teachers. *Computer Assisted Language Learning, 22*(2), 153-172.

Cáceres-Lorenzo, M. T. 2015. Teenagers learning Chinese as a foreign language in a European Confucius Institute: the relationship between language learner strategies and successful learning factors. *Language Awareness, 24*(3), 255-272.

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Cui, X. y Zhang, B. 2011. Global Chinese learner corpus construction plan. 崔希亮,张宝林.全球汉语学习者语料库建设方案.语言文字应用. *Language,* 100-108.

European Commission. 2010. *Europe's Digital Competitiveness Report*. Luxembourg: European Commission. http://ec.europa.eu/information_society/digital-agenda/documents/edcr.pdf.

Friginal, E. & Cox, A. 2022. Corpus uses in language teaching. In *Handbook of Practical Second Language Teaching and Learning* (pp. 161-172). Routledge

Johns, T. F. 1991. From Printout to Handout: grammar and Vocabulary Teaching in the context of Data-driven Learning in Johns. *English Language Research Journal,* 4, 27-45.

Griffiths, C. 2015. What have we learnt from good language learners?". *ELT Journal*, *69*(4), 425-433.

Hunston, S. 2002. *Corpus in Applied Linguistics*. Cambridge: Cambridge University Press.

INTEF. 2017. Common Digital Competence Framework for Teachers. Madrid: INTEF.

Kennedy, G. D. 2014. *An introduction to corpus linguistics*. London: Routledge.

Muñoz, C. 2014. Exploring young learners' foreign language learning awareness. *Language Awareness*, *23*(1-2), 24-40.

Qi, Y., Wang, L. & Zhang, Y. 2022, October. Research on the Design of Data-Driven Teacher Support. In *Proceedings of the 14th International Conference on Education Technology and Computers* (pp. 330-336).

UNESCO. 2008. ICT Competency Framework for Teachers. Paris: UNESCO.

Vidal Ledo, M. & Rivera Michelena, N. 2007. Investigación-acción. *Educación Médica Superior, 21*(4), 0-0.

Creating an institution-specific academic wordlist for an industrial engineering bachelor's programme

Claudia Wunderlich – *Technical University of Applied Sciences Würzburg-Schweinfurt*

Keywords: *English for Specific Purposes, English for Academic Purposes, Academic word-list, subject-specific academic wordlist, institution-specific academic wordlist, DIY corpus, industrial engineering.*

Creating subject-specific word lists from specialized DIY corpora has proved to be highly relevant for learning and teaching English for Specific Purposes (ESP) including English for Academic Purposes (EAP,cf Ward 2009, Nation 2016, Coxhead 2018). This paper presents first results of a corpus study into the language relevant for an industrial engineering programme with a focus on engineering in an EMI (English Medium of Instruction) bachelor's programme in Central Europe. The international students enrolling in the programme are required to provide proof of B2 general English upon enrolment, but lack knowledge of the specific technical and academic language. While engineering word lists have been created in the past (for example Hsu, 2014; Mudraya, 2006; Todd, 2017; Ward, 2009), a list consisting specifically of the academic vocabulary in this programme is needed. The relevance of institution-specific word lists has also been demonstrated (cf Veenstra & Sato 2018). The aim is to equip students with the linguistic means to succeed in their studies and understand the literature needed to research and write their bachelor's thesis in English.

The corpus compiled and its yields is to be used for data-driven learning and creating corpus-informed materials for the English courses to optimize the learning of the relevant words and lexical bundles. A DIY corpus of 2,130715 tokens of written academic language is analysed on the basis of of Coxhead (2000) using frequency, range, as well as expert opinion as selection criteria and also comparing the final list with Gardner and Davies' (2013) to identify the word families specific to the academic and technical language of industrial engineering. The corpus consists of journal articles from relevant international journals of the discipline reflecting the topics currently central to the discipline such as robotics, industrial Internet of things, manufacturing, supply-chain management, and AI grouped into ten sub-corpora. The study confirms the previous findings according to which subject-specific academic word lists deviate significantly from the above-cited general academic word lists and even general engineering word lists. To achieve better reading comprehension in industrial engineering, this newly created list should be preferred, which contains 550 word families overall and is divided into ten sub-lists. Together with separate lists of acronyms (such as ERP, JIT, CAD), collocations, and lexical bundles consisting of up to five consecutive words, these sub-lists provide the basis for a task-based English course with a lexical syllabus and enable the systematic acquisition of the most relevant academic vocabulary.

References

Coxhead, A. 2000. A New Academic Word List. *TESOL Quarterly*. 34/2, 213-238. https://doi.org/10.2307-/3587951.

Coxhead, A. 2018. Vocabulary and English for Specific Purposes Research: Quantitative and qualitative perspectives. London: Routledge.

Gardner, D. & Davies, M. 2013. A New Academic Vocabulary List. *Applied Linguistics*, 35. 305-327. https://doi.org/10.1093/applin/amt015.

Hsu, W. 2014. Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes*, 33, 54–65. https://doi.org/10.1016/j.esp.2013.07.001.

Mudraya, O. 2006. Engineering English: A lexical frequency instructional model. *English for Specific Purposes*, 25, 235–256. https://doi.org/10.1016/j.esp.2005.05.002.

Nation, I. S. P. 2016. Making and Using Word Lists for Language Learning. Amsterdam: John Benjamins.

Todd, R. W. 2017. An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes,* 45, 31–39. https://doi.org/10.1016/j.esp.2016.08.003.

Veenstra, J. & Sato, Y. 2018. Creating an Institution-Specific Science and Engineering Academic Word List for University Students. *The Journal of Asia TEF*, 15/1, 148-166. 10.18823/asiatefl.2018.15.1.10.148.

Ward, J. 2009. A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, 28, 170–182. https://doi.org/10.1016/j.esp.2009.04.001.

Translating *I mean* on social media: A corpus-based analysis of the use of *I mean* from English to Chinese during the Covid-19 pandemic in Taiwan

Yu-Che Yen – *National Chengchi University*

Keywords: *I mean, discourse marker, translation, corpus linguistics, Covid-19.*

In the interactional conversation, the speaker applies the discourse marker *I mean* to utter their intention. During the Covid-19 pandemic, the use of social media is rising high. Kiesling (2020) finds that the discourse marker *I mean*, as a stance-taking marker depending on the previous proposition to process the following conversation, is used to intensify the speaker's utterance on social media. However, these intensified statements are constructed in various syntactical forms (also see Thompson, 2002). Interlocutors may find it hard to follow the conversation immediately since the syntax of propositions of *I mean* can be varied, not to mention how to translate the speaker's utterance into different languages. In most cases, EFL learners may turn to the help of Google Translation to understand the speaker's utterance. Therefore, to understand the relationship between functions of *I mean* and propositions in translating *I mean*, this paper collects the use of *I mean* on CrowdTangle, an analytical tool designed by Meta to help researchers understand what is happening across social media during the Covid-19 pandemic (2020/01/02~2023/01/12).

Among 1111 posts, 477 tokens of *I mean* are pulled from CrowdTangle. To understand how propositions affect *I mean* in translation, this study manages to categorize propositions into two forms: (1) Statement (S), and (2) Question (Q). Also, to explore how functions of *I mean* affect the translation, this study inputs the collected data to Google Translation. The data output is compared in each category to understand the relationship between the translation and functions of *I mean*. Functions of *I mean* are categorized according to past studies (e.g., Crible, 2017; Tree & Schrock, 2002; Schiffrine, 1987; Thompson, 2002) into four categories: clarification (C), utterance (U), expansion (E), and fixed expression (F). Results show that among all the propositions before and after *I mean*, SS appears to be the most frequent use of propositions in constructing *I mean* (53.25%). Also, *I mean* is frequently equivalently translated into "*wodeyisishi*" (**我的意思是**) in Chinese (84.5%). As to functions of *I mean*, *I mean* is often equivalently translated into "*wodeyisishi*" in Chinese (71.06%). However, in fixed expression, the result shows that *I mean* is typically equivalently translated into "*wodeyisi*" (**我的意思**) in Chinese (43.53%). It indicates that in both English and Chinese syntax, there are no further words, phrases, or clauses following behind the fixed expression of using *I mean*. Therefore, there is no need to add "*shi*" (**是**) to bring out more information comes after, since "*shi*" in Chinese is a predicate to introduce further information as *I mean* does. Considering that Thompson (2002) asserts that three-dimensional uses of *I mean* (also see Schiffrin, 1987), *I mean* should be translated into different semantic or syntactic forms regarding propositions and functions to make the conversion explicit. In short, the findings show that EFL learners can explore the other uses of propositions before and after *I mean* to mark different conversational purposes. Also, if heavily relying on the use of Google Translation, EFL learners may not learn to differentiate the discourse relations of using *I mean*, nor can they understand how to translate them into the needs of the conversation.

Bibliography

Crible, L. 2017. Discourse markers and (dis) fluency in English and French_ Variation and combination in the DisFrEn corpus. International Journal of Corpus Linguistics, 22(2), 242-269.

Kiesling, S. F. 2020. Investment in a model of stancetaking: I mean and just sayin'. *Language Sciences*, 82, 101333.

Schiffrin, D. 1987. *Discourse markers* (No. 5). Cambridge University Press.

Thompson, S. A. 2002. "Object complements" and conversation towards a realistic account. Studies in Language. *International Journal sponsored by the Foundation "Foundations of Language"*, 26(1), 125-163.

Tree, J. E. F., & Schrock, J. C. 2002. Basic meanings of you know and I mean. *Journal of Pragmatics*, 34(6), 727-747.

POSTERS

Fases en la elaboración de AMERLEX: americanismos léxicos en las lenguas españolas e inglesa documentados en textos sobre América anteriores a 1700

María Teresa Cáceres-Lorenzo, Yaiza Santana-Alvarado & Anabel Mederos-Cedrés

University of Las Palmas de Gran Canaria

Palabras clave: *corpus, americanismos léxicos, siglos áureos.*

El descubrimiento de América impulsa la publicación de textos que registran lo que acontece en el Nuevo Mundo mediante nuevas designaciones que se incorporan progresivamente a las lenguas europeas. Si bien España protagoniza la aventura americana, Inglaterra, deseosa de emular los éxitos españoles, relata sus propias expediciones a territorios ya colonizados y traduce con interés textos españoles, lo que conlleva el conocimiento y posterior incorporación de este vocabulario para designar la nueva realidad. Las nuevas voces americanas no solo se documentan en distintas tipologías textuales que se caracterizan por el propósito general de informar con veracidad sobre el descubrimiento, exploración, conquista y colonización de América, sino que pasan a formar parte del inventario de entradas de diccionarios españoles e ingleses.

El objetivo de esta comunicación es explicar las distintas fases (planteamiento teórico, realización, resolución de problemas y ejecución) que se han seguido para la elaboración de una base de datos en línea y en abierto que hemos denominado AMERLEX (americanismos léxicos en las lenguas españolas e inglesa documentados en textos sobre América anteriores a 1700- Proyecto PID2019-104199GB-I00).

La investigación llevada a cabo en la elaboración de este corpus sigue una metodología de fundamentación documental que se basa en la recopilación de los siguientes datos de los americanismos identificados: lema, grafías, obra, año de publicación, número de edición, tipología textual, autor y su origen, categoría gramatical, área léxica, lengua de procedencia, muestra en textos españoles e ingleses, definición en diccionarios de la época, etc.

Los resultados de AMERLEX en cifras indican que se han recopilados 1.720 lemas y 4.015 grafías. Toda esta información se ha ejemplificado con 14.471 citas textuales. Las tipologías textuales ordenadas por número de lemas son: a) crónica de indias: 6.487; b) descripción natural: 3.324; c) prosa didáctica: 1.195; d) libro de viajes: 800; e) relación geográfica: 694; f) obra en verso: 654; g) prosa religiosa: 571; h) diccionario: 528; i) tratado médico: 149; j) texto político: 41; k) teatro: 24; y l) carta: 4.

Esperamos que esta contribución sea un aporte teórico-práctico a las preguntas planteadas en esta sección.

Bibliografía

Asociación de Academias de la Lengua Española. 2010.*Diccionario de Americanismos*. Santillana.

Bertolotti, Virginia y Company Company, Concepción. 2022. "Corpus diacrónicos del español en las Américas". *Lingüística de corpus en español* / coord. por Giovanni Parodi Sweis, Pascual Cantos Gómez, Chad Howe, págs. 45-58.

Cáceres Lorenzo, María Teresa. 2020. Fundamentación textual en el Corpus del español del siglo XXI (CORPES) del americanismo obsolescente. Variables de un estudio de caso. *Revista signos: estudios de lingüística*, 102, págs. 144-169.

Corbella, Dolores; Fajardo, Alejandro y Langenbacher-Liebgott, Jutta (eds.), 2018. *Historia del léxico español y Humanidades digitales,* Peter Lang.

Creación de un corpus previo al registro en la web de anotación: el caso del antillanismo 'canoa'

Anabel Mederos-Cedrés & Miguel Ángel Rodríguez-Falcón

University of Las Palmas de Gran Canaria

Palabras clave: *corpus, web de anotación, americanismos, antillanismos, lexicografía.*

El objetivo de esta investigación es presentar la creación de un corpus previo en la metodología seguida en la creación de una base de datos en abierto en el Proyecto PID2019-104199GB-I00 sobre la vitalidad de los americanismos léxicos en las lenguas española e inglesa documentados en textos de los siglos XVI y XVII-AMERLEX-DATABASE (Proyectos de I+D+i – PGC). Dicho corpus previo se realizó antes del registro de los datos en la web de anotación con los datos extraídos de otros corpus: Léxico Hispanoamericano (1493-1993) (Boyd Bowman 2003); Diccionario de Americanismos (DA 2010); Nuevo diccionario histórico de la lengua española (DHLE 2013); Diccionario de la Lengua Española (DRAE/DLE, distintas ediciones); Corpus Electrónico del Español Colonial Mexicano (COREECOM 2013); y Corpus Diacrónico y Diatópico del Español de América (CORDIAM 2015) de la Academia Mexicana de la Lengua. AMERLEX debía contener ejemplos textuales de los objetivos propuestos. La anotación de corpus o la creación de marcas con información era necesaria para demostrar o probar una determinada teoría lingüística con respecto a los americanismos, y tanto era así que el registro del corpus era fiel reflejo de la teoría que se estaba describiendo o probando (Ide y Pustejovsky 2017)

El propósito de este nuevo corpus es averiguar qué antillanismos, a través del estudio de caso de *canoa*, han pervivido en la evolución histórica de las hablas americanas, y cuál ha sido el grado de difusión panhispánica o regional según las fuentes académicas en la sincronía actual (Quirós García y Ramírez Luengo 2015; Cáceres Lorenzo 2022). La vitalidad de una lexía indígena va unida a otras nociones como adaptación, adopción ortográfica, morfológica y semántica (Alvar 1975) y a la creación de nuevos significados en su proceso diacrónico de transmisión (Sala et álii 1982). Las conclusiones sobre la incorporación de las voces antillanas al español provienen de investigaciones parciales y dispersas (determinadas crónicas, textos de localidades muy concretas, etc.), pero, al mismo tiempo, en la última década existe una proliferación de nuevos corpus que persiguen obtener resultados acordes a la naturaleza de los escritos indianos (Rivarola 2012; Bertollotti y Company 2014) a la incorporación de las variables diacrónica, diastrática, diafásica y diatópica.

La efectividad de un corpus como herramienta de investigación lingüística reside en gran medida en el tipo de búsqueda que permite hacer y en los datos que proporciona (Davies 2009). En nuestro caso, el citado corpus previo tuvo las siguientes funciones: 1) formar parte del marco teórico de la tarea que se quiere resolver y del objetivo del proyecto; 2) registrar datos de la bibliografía; 3) ser útil para crear un modelo de anotación; y 4) ser una información teórica necesaria en la interacción entre ingenieros y lingüistas.

Los resultados obtenidos sobre este estudio de caso, en que *canoa* presenta más de once significados con sus respectivas variantes, supone una evidencia necesaria para la creación de un corpus en el que se informe sobre el reconocimiento de la integración y vitalidad de las voces antillanas en la lengua española.

Referencias

Alvar, M. 1975. *España y América cara a cara*. León: Bella Época.

Bertolotti, V. y Company, C. 2014. El corpus diacrónico y diatópico del español de América (CORDIAM). Propuesta de tipología textual. *Cuaderno de la ALFAL*, (6), pp. 130-148.

Boyd Bowman, P. 2003. Léxico hispanoamericano 1493-1993. En Ray Harris-Northall and John J. Nitti (eds). Nueva York: *Hispanic Seminary of Medieval Studies*. Consultado en https://textred.spanport.lss.wisc.edu/lexico_hispanoamericano/.

Cáceres Lorenzo, M. T. 2022. Hispanic-American dialectology in the 16th century. Penetration of Americanisms in Nicolas Monardes' *Historia Medieval. Dialectologia et Geolinguistica*, 30, (1), pp. 115-130.

[CORDIAM] Academia Mexicana de la Lengua. 2015. Corpus Diacrónico y Diatópico del Español de América. Consultado en https://www.cordiam.org.

[COREECOM] Grupo de Estudio del Español Colonial Mexicano. 2013. Corpus Electrónico del Español Colonial Mexicano. Arias, B. (Coord.). IIFL-UNAM. Consultado en https://www.iifilologicas.unam.mx/coreecom/index.php?page=inicio&men=1.

[DA] Asociación de Academias de la Lengua Española. 2010. *Diccionario de americanismos*. Consultado en https://www.asale.org/damer/.

Davies, M. 2009. Relational databases as a robust architecture for the analysis of word frecuency. En Archer, D. (ed.), *What's in a Wordlist? Investigating Word Frecuency and Keyword Extraction*. Londres: Ashgate, pp. 53-68.

[DHLE] Real Academia Española. 2013. *Diccionario histórico de la lengua española*. Consultado en https://www.rae.es/dhle/.

[DLE] Real Academia Española 2014. *Diccionario de la lengua española* (23.ª ed.). Consultado en https://dle.rae.es/.

Ide, N. y Pustejovsky, J. 2017. *Handbook of Linguistic Annotation*. Springer Netherlands.

Quirós García, M. y Ramírez Luengo, J. L. 2015. Observaciones sobre el léxico del español de Yucatán (1650-1800). *Revista de filología española*, 95, (1), pp. 183-210.

Rivarola Rubio, J. L. 2012. Los corpora en el estudio histórico del español de América (un corpus documental del español en el Perú de los siglos XVI y XVII): *Reflexiones y perspectivas. Actas del VIII Congreso Internacional de Historia de la Lengua Española* (pp. 391-396). España: Santiago de Compostela.

Sala, M., Munteanu, D., Neagu, V. y Sandru-Olteanu, T. 1982. *El español de América*. Bogotá: Instituto Caro y Cuervo.

Sinclair, J. 1996. *Preliminary Recommendations on Corpus Typology*. EAGLES. Consultado en http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

An annotated English-Mandarin code-switching corpus for sociolinguistics research and language technologies

Priya Rajeev & Hongchen Wu

Georgia Institute of Technology

Keywords: *informal text-based data, online community, young bilingual professionals, sociolinguistic analysis.*

Code-switching refers to the phenomenon where speakers incorporate speech from more than one language into one utterance. We present in this paper the design of an annotated text-based English-Mandarin code-switching corpus with data collected from an online bilingual community and some initial sociolinguistics analysis alongside it.

Existing English-Mandarin code-switching corpora primarily detail speech-based data and are often not open-access (Li et al. 2012; Liu et al. 2015). A recently published English-Mandarin corpus by Calvillo et al. (2020) covers informal text-based data; however, this data was collected from online forums for Chinese students at only three universities in the United States, with just four topics covered (housing, secondhand goods, experience sharing, and ride sharing). Given the existing English-Mandarin corpora, our goal was to build a text-based corpus with data collected from a widespread source. Thus, this paper presents informal text-based English-Mandarin data across ten topics collected from 1point3acres, a global online forum for English-Mandarin bilinguals. Our objective is for this data to be used in language technology research and to further examine the dynamics within the English-Mandarin bilingual community.

To gather data, we developed a web-scraping script to visit each of the topic pages displayed on the home screen. We collected the first page of threads from each of the ten topic pages on two dates: first on November 29th, 2022, then again on December 9th, 2022. For each thread, we collected the thread title, the thread message, the thread topic, and the date posted. After removing duplicates, we had a total of 522 threads. We conducted data cleaning processes by removing the words related to log-in prompts and formatting, and separating threads into sentences. We then used the spaCy program (Honnibal et al., 2020) for word segmentation and part-of-speech (POS) tagging. Manual evaluation and translation of the code-switched parts are in progress. The data is annotated on four levels: topic of discussion, word segmentation, POS tagging, and proportion of code-switching. At this moment, the dataset includes 193,582 words, with 162,088 Mandarin words and 31,494 English words.

The topics with the highest rates of code-switching were USA-based job referrals, data science, and China-based job referrals. For posts related to US-based job referrals and data science, we theorize that the motives for code-switching may be referential in nature due to lack of Mandarin words (Appel & Muysken, 1987). However, we conjecture that posts relating to China-based job referrals may code-switch at higher frequencies for directive reasons (Appel & Muysken, 1987), in order to keep job-related information confidential. In addition, further examination of our data found that speakers were most likely to code-switch adpositions and proper nouns. While a high rate of code-switching for nouns has been discussed in prior literature, the finding that adpositions are code-switched more frequently is novel.

Our next step is to further annotate the data and add to the corpus by including a variety of code-switching sources. Our goal is for our corpus to be beneficial for many various applications, such as demographic research on the English-Mandarin bilingual community, code-switching predictive models, and bilingual language technologies.

References

Appel, R., and Muysken, P. 1987. *Language Contact and Bilingualism*. London: Edward Arnold.

Calvillo J., Fang L., Cole J., Reitter D. 2020. Surprisal Predicts Code-Switching in Chinese-English Bilingual Text. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP '20)*. doi: 10.18653/v1/2020.emnlp-main.330.

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python. doi: 10.5281/zenodo.1212303.

Li, Y., Yu, Y., and Fung, P. 2012. A Mandarin-English Code-Switching Corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC '12).

Lyu, D-C., Tan, T-P., Chng, E-S., Li, H. 2010. Mandarin–English code-switching speech corpus in South-East Asia: SEAME. *Language Resources and Evaluation*. doi: 10.1007/s10579-015-9303-x.

The problem of balance and representativeness in the *Electronic Corpus of 17th- and 18th-century Polish Texts*

Ewa Rodek – *Institute of Polish Language Polish Academy of Sciences*

Keywords: *historical corpus, Polish Baroque and Enlightenment, overrepresentation of topic.*

The second edition of the Electronic Corpus of 17th- and 18th-century Polish Texts (korba.edu.pl) has fully revealed the problems a historian of language has to face when selecting texts for the Polish diachronic corpus. In my presentation I will focus on the issue of preserving the principle of representativeness and balancing the corpus counting (after the completion of the work) 25M tokens from the Polish Baroque and Enlightenment periods. I will present how much the shape of the corpus is influenced by the specificity of a given epoch in the history of language.

The main problem of preserving the representativeness of the linguistic material is the monothematic nature of Polish literature from the years 1680-1740. Due to the domination of the bookselling market by monastic colleges, the vast majority of writing production concerned religious subjects, or utility texts such as calendars were printed.

Another problem is the fact that a large part of the 17th century sources remained in manuscripts. This is related to the issue of the exclusivity of Baroque literature. Its centres became noblemen's courts, which meant that it was consumed in close circles, there was little need for print, and thus a significant part of the literary legacy remained in manuscripts. In addition, heterogeneous genre forms proliferated (*silva rerum*, called also *miscellanea*, *collectiva*, *etc.*), which pose a major methodological problem for corpus creators as well. They are a significant part of the culture of the time, but very difficult to include in a corpus of texts. Manuscripts pose several problems: it is difficult to establish their exact dating, as well as their originality, and it is embarrassing that often the main criterion for accepting a manuscript into a corpus is its legibility, rather than its subject matter, author or addressee.

The Age of Enlightenment, on the other hand, brings a significant revival of literature. Mature, extensive treatises, multi-volume works appear. This poses a new problem, as it is necessary to adjust the size of the text samples in relation to the source material from the earlier period, but at the same time to take into account as fully as possible the new lexis coming into use in the second half of the 18th century.

Another issue was the low geographical diversity of available texts. The monastic printing houses at the major academic centres (Warsaw, Cracow, Vilnius) were most active, which meant that some regions were over-represented in the literature, while others (Greater Poland, Podlasie, Pomerania) were under-represented.

In preparing the Baroque corpus, we decided that our role was to strive sensibly for a balance between particular sub-periods, geographical regions or thematic sections. We have attempted to record the phenomena highlighted here through an extensive system of metadata, including the designation of genre and subject matter for each text, release date, region of origin, and literary type, including the mixed type (mixed epic and lyric).

References

Adamiec D. 2015: Kryteria doboru tekstów do „Elektronicznego korpusu tekstów polskich z XVII i XVIII w. (do 1772 r.)". *Prace Filologiczne* LXVII, p. 11-20.

Biber D. 1993: Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4), p. 243-257 (https://doi.org/10.1093/llc/8.4.243).

Chachulska B., Górski R. L. 2005: Korpusy komputerowe języków słowiańskich. *Studia z Filologii Polskiej i Słowiańskiej* 40, p. 483-507.

Gruszczyński W., Adamiec D., Bronikowska R., Wieczorek A. 2020: Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. – problemy teoretyczne i warsztatowe. *Poradnik Językowy* 8, p. 32–51.

Gruszczyński W., Adamiec D., Bronikowska R., Kieraś W., Modrzejewski E., Wieczorek A., Woliński M. 2022: The Electronic Corpus of 17th- and 18th-century Polish Texts. *Language Resources and Evaluation* 56, p. 309–332.

Krinková Z. 2018: Historical corpus linguistics and Spanish: the state of the art and current problems. *Časopis pro moderní filologii*, 100(1), p. 60-79.

Kytö M. 2011: Corpora and historical linguistics. *Revista Brasileira de Linguistica Aplicada*, 11(2): 417–457 (https://doi.org/10.1590/S1984-63982011000200007).

Ejemplo de construcción de subcorpus específico de americanismos hispanizados: el caso del vocabulario en *Historia general de las conquistas del Nuevo Reino de Granada* (1676)

Miguel Ángel Rodríguez-Falcón & María Teresa Cáceres-Lorenzo – *University of Las Palmas de Gran Canaria*

Palabras clave: *corpus, americanismos léxicos, siglos áureos.*

La investigación que se va a presentar forma parte de una serie de trabajos vinculados al proyecto AMERLEX (PID2019-104199GB-100), cuya finalidad principal es realizar una base de datos en línea y en abierto que recoja el vocabulario americano existente en una selección de textos españoles e ingleses sobre América publicados en los siglos XVI y XVII.

Las investigaciones del léxico de la época se han centrado frecuentemente en el análisis de los textos redactados en lengua española para inventariar los indigenismos y las palabras patrimoniales hispánicas que comienzan un proceso de americanización. Los trabajos académicos de dialectología histórica del español americano giran en torno, pues, a los corpus o conjuntos de datos lingüísticos, palabras diferenciales de América, que aparecen en determinados textos que proporcionan evidencia empírica. La relevancia de la fundamentación documental para la recuperación de los americanismos utilizados ha sido demostrada por diversos especialistas de la historia de la lengua española, pero la abundancia de documentos impide que la labor de recopilar testimonios textuales se concluya, por lo que se deben escoger mediante criterios relacionados con las referencias diatópicas, diacrónicas, diafásicas y diastráticas, así como el reflejo de la oralidad y la tradición discursiva. Para la realización de los corpus acerca de América, algunos autores proponen, en virtud de la función comunicativa que asumen y sin que existan fronteras nítidas entre ellas, cuatro tradiciones discursivas o clases textuales: cartas, textos jurídicos, crónicas y textos administrativos.

El objetivo de nuestro estudio es la creación de un subcorpus de americanismos léxicos para poner de manifiesto los que ya se empleaban en el siglo XVII. Por ello, se han detectado y extraído las voces, tanto indígenas como españolas americanizadas, presentes en una obra escrita en dicha centuria por el obispo Lucas Fernández de Piedrahíta, *Historia general de las conquistas del Nuevo Reino de Granada* (1676). Con el póster científico en elaboración se pretenden mostrar, concretamente, los americanismos hispanizados que contiene este texto de carácter descriptivo.

Para llevar a cabo la contribución hemos utilizado, además de la mencionada obra, compuesta de cinco libros disponibles en la web, el *Breve diccionario de colombianismos* de la Academia Colombiana de la Lengua y, máxime, el *Diccionario de la lengua española* y el *Diccionario de americanismos* de la Real Academia Española y la Asociación de Academias de la Lengua Española.

Se sigue en nuestra investigación, al igual que en el proyecto AMERLEX, una metodología de fundamentación documental, basada en la recopilación de los siguientes datos de los americanismos identificados: lema, grafías, obra, año de publicación, número de edición, tipología textual, autor y su origen, categoría gramatical, área léxica, lengua de procedencia, muestra en textos en español, definición en los diccionarios académicos, etc.

Deseamos que este trabajo constituya un aporte teórico-práctico a las preguntas que se formulan en la sección de lexicología y lexicografía basadas en corpus.

Bibliografía

Arias Álvarez, B. y Hernández Mendoza, J. A. 2013. Importancia de la incorporación de los parámetros diastráticos y diafásicos en la elaboración del corpus electrónico del español colonial mexicano. *Scriptum Digital, 2*, 5-20.

Asociación de Academias de la Lengua Española 2010. *Diccionario de americanismos*. Santillana.

Bertolotti, V. y Company, C. 2014. El corpus diacrónico y diatópico del español de América (CORDIAM). Propuesta de tipología textual. *Cuaderno de la ALFAL, 6*, 130-148.

Cáceres Lorenzo, M. T. 2013. Nuevos datos sobre el uso de voces del fondo hispánico tradicional en textos españoles del siglo XVI. *Onomázein, 27*, 135-143.

Cáceres Lorenzo, M. T. 2017. Taxonomía de los documentos del siglo XVI: las Relaciones Geográficas de Indias para un corpus sobre americanismos léxicos. *Estudios Filológicos, 59*, 31-46.

Cáceres Lorenzo, M. T. 2020. Fundamentación textual en el Corpus del español del siglo XXI (CORPES) del americanismo obsolescente. Variables de un estudio de caso. *Revista Signos, 53*(102), 144-169.

Hernández, E. 2012. En torno a la selección y la edición de documentos para un corpus histórico de textos del español americano. En M. J. Torrens Álvarez y P. Sánchez-Prieto Borja (Coords.), *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos* (pp. 260-269). Peter Lang USA.

Kabatek, J. 2013. ¿Es posible una lingüística histórica basada en un corpus representativo? *Iberoromania, 77*, 8-28.

Parodi, C. 2008. Lingüística de corpus: una introducción al ámbito. *Revista de Lingüística Teórica y Aplicada, 46*, 93-119.

Real Academia Española y Asociación de Academias de la Lengua Española 2014. *Diccionario de la lengua española*. Espasa.

Enseñar los colores a estudiantes plurilingües C1-C2 a través de un texto cronístico del siglo XVI

Yaiza Santana-Alvarado & Kim Tate-Pérez

University of Las Palmas de Gran Canaria

Palabras claves: *americanismos léxicos, Joseph de Acosta, historia natural y moral de las Indias, estudiantes plurilingües.*

La presente investigación forma parte del Proyecto: Americanismos léxicos en las lenguas españolas e inglesa documentados en textos sobre América anteriores a 1700: AMERLEX (PID2019-104199GB-IOO), que se desarrolla en la Universidad de Las Palmas de Gran Canaria. Este proyecto nos ha llevado al análisis del léxico de obras consideradas productos culturales del siglo de Oro en relación con las necesidades de aprendizaje de estudiantes plurilingües cuyo nivel de español se refiere a C1-C2 según el Marco Común Europeo de Las Lenguas, motivando así el estudio de los colores (creaciones españolas frente a herencia románica) en el aprendizaje de ELE en estos estudiantes, a través de un estudio de caso del siglo XVI: Historia natural y moral de las Indias (1590), de Joseph de Acosta, (jesuita antropólogo, historiador y profesor universitario). Espejo Muriel en su obra: *Los nombres de color en la naturaleza: estudio onomasiológico*, afirmaba que si hiciéramos una comparación del estado del léxico cromático del periodo en cuestión con el contemporáneo, observaríamos la aparición y desaparición de términos o cambios semánticos, y podríamos reconstruir el proceso de formación de este campo léxico. Ante esto se ha elaborado una investigación manual en la que hemos realizado un método de búsqueda exhaustivo en la obra y a través del Corpus del Diccionario Histórico de la lengua española sobre aspectos referidos a los nombres de los colores a lo largo de la historia de la lengua española, desde el español medieval para darle respuesta a nuestras preguntas de investigación: ¿qué colores predominan en el texto *Historia natural y moral de las Indias* en el que se abordan temas cosmográficos, biológicos, botánicos y geográficos?, y en dicho léxico cromático, ¿cuáles son de herencia medieval o de nueva creación?, y ¿qué finalidad en la metodología didáctica promueve el uso de dicho texto en el aprendizaje del español?

Para ello se ha diseñado una investigación cuantitativa y cualitativa en tres fases: (a)localización de los colores predominantes en la obra seleccionada; (b)registro de léxico de origen medieval o de nueva creación; (c)análisis cuantitativo de los resultados obtenidos. En cuanto a los primeros resultados obtenidos, los colores predominantes en la obra son: el negro, azul, rojo, amarillo y verde, en el que un 35% corresponde al color negro, seguido del color verde; un 30%; el color azul, un 20%; el color rojo, un 10% y, finalmente el color amarillo, un 5%.

Pensamos con los resultados aportar materiales para la elaboración de material didáctico con productos culturales del periodo áureo.

Bibliografía

Cáceres-Lorenzo, María Teresa. 2021. Herencia medieval y neologismo cromático de textos historiográficos en el siglo de Oro según el Corpus Diacrónico del español, *Neophilologus*,105:521-538, DOI: 10.1007/s11061-021-09689-3.

Corpus del Diccionario histórico de la lengua española. En línea. Disponible en https://apps.rae.es/CNDHE-/view/inicioExterno.view;jsessionid=31397AB68CB08D05C7A9AD39462E3904.

Duncan, M. 1975. Color words in medieval Spanish. En S. Beardsley et al. (eds.), *Studies in honor of Lloyd A. Kasten* (págs. 53-71). Madison: Hispanic Seminary of Medieval Studies.

Enguita Utrilla, J. M. 2004. *Para la historia de los americanismos léxicos.* Frankfurt am Main: Peter Lang.

Espejo Muriel, M. 1990. Los nombres de color en la naturaleza: estudio onomasiológico. Granada: Universidad.

Frago Gracia, J. A 1999. Historia del español de América. Textos y contextos. Madrid: Gredos.

Lapesa, R. 2000. *Historia de la lengua española.* Madrid: Gredos.

Stala, E. 2011. *Los nombres de los colores en el español de los siglos XVI-XVII.* Alicante: Biblioteca Virtual Miguel de Cervantes: http:/www.cervantesvirtual.com/obra/los-nombres-de-los-colores-en-el-español-de-los-siglos-xvi-xvii/. Consultado el 20 de marzo de 2020.

Reframing metaphors in discourse on COVID-19 and climate change: A corpus- based analysis of media representations of two intersecting global issues

Giovanni Tucci – *Università degli Studi di Bari Aldo Moro*

Keywords: *corpus, metaphors, COVID-19, climate; media.*

This research study draws on corpus linguistics to investigate how the intersection between COVID-19 and climate change is currently portrayed in the media. It has been argued that the severe coronavirus crisis has been only a foretaste of the long-term calamitous consequences the world will suffer in the near future due to climate change (Fuentes *et al.* 2020). Clearly, these two existential threats share similarities, as they are viewed as two transboundary phenomena that expand in space and time.

In this project, a language analysis has been carried out considering a corpus of tweets and a corpus of articles, retrieved from *The Guardian* and *The Sunday Times*, which address the synergistic interaction between climate change and COVID-19. In a second step, the data have been examined with the software *WordSmith Tools* 7.0 (Scott 2019), employing a mix of quantitative and qualitative analysis. The language investigation has been conducted at multiple levels, combining a phraseological perspective with an analysis of metaphors related to the topic under investigation.

Based on the assumption that metaphors are widespread in human communication (Lakoff and Johnson 1980), thus having profound implications (Musolff 2006), as they hardly promote neutral representations of reality (Atanasova 2022), this study aims to shed light on how the link between these overlapping issues is framed through the use of specific metaphorical constructions. The research is still in its early stages, and predictably, the preliminary findings show that colour and war metaphors prevail in environmental discourse connected with COVID-19.

It should be pointed out that the concept of "green recovery" has been abundantly exploited in the past, especially in the aftermath of the 2008 financial crisis to argue for climate-friendly economic stimulus packages (Shaw and Nerlich 2015). Accordingly, in the analysed corpora, the pandemic tends to be represented as a challenge to "go green", which may force countries to rethink contemporary capitalism, fuelling a "Green Industrial Revolution" or a "Green New Deal". Furthermore, the language investigation has unearthed a considerable number of instances in which war metaphors predominate. Broadly speaking, it is not surprising that war metaphors are ubiquitous in discourse, as they perfectly lend themselves to expressing definitive, unambiguous outcomes, such as victory or defeat, conveying a sense of urgency to act (Karlberg and Buell 2005). Hence, climate change and COVID-19 end up being personified as two enemies to beat at any cost.

Yet, as Sontag (1989) asserts in his seminal work on cancer, arguments have been made against the use of war metaphors, especially when discussing health issues, since militaristic and violent language would help stigmatise people who fail to recover from an illness, thus being blamed for "not fighting hard enough". Apparently, negative metaphors exert tremendous power over people, who have what Thibodeau *et al.* (2019) call a "negativity bias". It follows that further investigation is still required to assess the existence and the impact of reframing mechanisms involving those metaphors, which may offer alternative ways of representing the link between COVID-19 and climate change, so as to ultimately inspire proactive attitudes in line with the principles of the #ReframeCovid initiative (#ReframeCovid, online: https://sites.google.com/view/reframecovid/home?authuser=0), launched by two Spanish researchers in April 2020, which promotes the spread of non- war related metaphors, studied to unite people in difficult times.

References

Atanasova D. 2022. "How Constructive News Outlets Reported the Synergistic Effects of Climate Change and Covid-19 Through Metaphors", *Journalism Practice*, 16/2-3: 384-403.

Fuentes R., Galeotti M., Lanza A., Manzano B. 2020. "COVID-19 and Climate Change: A Tale of Two Global Problems", *Sustainability*, 12:1-14.

Karlberg M., Buell L. 2005. "Deconstructing the 'War of all Against All': The Prevalence and Implications of War Metaphors and Other Adversarial News Schema in TIME, Newsweek, and Maclean's.", *Journal of Peace and Conflict Studies*,12/1: 22–39.

Lakoff G., Mark J. 1980. *Metaphors We Live by*, Chicago, IL, Chicago University Press. Musolff A. 2006, "Metaphor Scenarios in Public Discourse", *Metaphor and Symbol* 21/1. Scott M. 2019, *WordSmith Tools* 7.0, Lexical Analysis Software Limited.

Shaw C., Nerlich B. 2015. "Metaphor as a Mechanism of Global Climate Change Governance: A Study of International Policies, 1992–2012.", *Ecological Economics*, 109: 34–40

Sontag S. 1989. *Illness as Metaphor: AIDS and Its Metaphors*, New York: Picador/Farrar, Straus and Giroux.

Thibodeau P.H., Teenie Matlock T., J. Flusberg S.J. 2019. "The role of metaphor in communication and thought", *Language and Linguistics Compass*, 13/5:

Websites

#ReframeCovid, online: https://sites.google.com/view/reframecovid/home?authuser=0.

SEMINARIO SOBRE COMPILACIÓN DE CORPUS

Corpus compilation workshop: How to quickly compile a corpus using *R*

Daniel Granados-Meroño – *University of Murcia*

Keywords: *corpus compilation, format conversion, corpus annotation, POS, R.*

As the name of the discipline makes clear, any research study related to corpus linguistics requires a corpus to work with. The corpus is the ultimate source of data in our research, therefore sooner or later the researcher will need to either find one corpus available fitting their researching needs (which, unfortunately, in many cases is not the case) or to compile a new one. This might increase considerably the time that the linguist needs to complete their study since the task of compiling a corpus is usually very time-consuming. It is necessary to consider what purpose our corpus (general or specialised, diachronic or synchronic) serves, and structure it in terms of representativity (millions of words or just hundreds of words) (Llisterri & Torruella Casañas, 1999; Marín & Camino, 2012). Once all these decisions are made, the researcher needs to look for the texts which will conform their corpus and then, manage to convert them into processable texts that corpus tools (such as automatic annotators or corpus query tools) are able to work with. This last step is especially difficult and is the one we are trying to make easier with this workshop.

R and Python are programming languages extensively used by many scholars from different scopes of study, but it has greatly gained interest among scholars in Linguistics with the quick evolution of Computational Linguistics research, closely related to the emerging disciplines of Natural Language Processing, Deep Learning and AI. However, we can use programming languages for other tasks such as corpus compilation, and this workshop will show participants how to do it. Even if we already have very useful statistical software packages such as SPSS or corpus tools such as AntConc or Lancsbox, R is a very powerful and versatile tool that can help us create much more efficient and organised studies.

In this 1-hour workshop, participants are going to learn how to use R to:

- Convert simultaneously, in a matter of seconds, several PDF files containing thousands of words into .txt files (with a reliable result);
- Create a corpus file in R to calculate basic statistical measures such as the TTR:
- Annotate automatically with POS tags more than 1 million-word corpora in less than an hour.

Participants will need to bring their laptops, the last version of R and RStudio installed, and a user-level knowledge of R or similar programming languages. For the last task (annotation with POS tags using SpacyR) you are recommended to work with a 8 GB RAM laptop and you need the last version of Python installed.

R packages required: 'pdftools', 'readtext', 'reticulate', 'dplyr', 'quanteda', 'SpacyR', 'stringr' (Benoit et al., 2018, 2020, 2021; Kalinowski et al., 2023; Ooms [aut & cre, 2023; Wickham et al., 2023; Wickham & RStudio, 2022)

References

Benoit, K., Matsuo, A., & Council (ERC-2011-StG 283794-QUANTESS), E. R. 2020. spacyr: Wrapper to the "spaCy" "NLP" Library (1.2.1). https://CRAN.R- project.org/package=spacyr

Benoit, K., Obeng, A., Watanabe, K., Matsuo, A., Nulty, P., & Müller, S. 2021. *readtext: Import and Handling for Plain and Formatted Text Files* (0.81). https://CRAN.R- project.org/package=readtext

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105-/joss.00774

Kalinowski, Tomasz [ctb, cre], Ushey, Kevin [aut], Allaire, J. J. [aut], RStudio [cph, fnd], Tang, Yuan [aut, cph], Eddelbuettel, Dirk [ctb, cph], Lewis, Bryan [ctb, cph], Keydana, Sigrid [ctb], Hafen, Ryan [ctb, cph], Geelnard,

Marcus [ctb, cph] (TinyThread library, http://tinythreadpp.bitsnbites.eu/) 2023. *Reticulate: Interface to Python* (1.28). https://CRAN.R-project.org/package=reticulate + https://rstudio.github.io/reticulate/

Llisterri, J., & Torruella Casañas, J. 1999. Diseño de corpus textuales y orales. *Filología e informática: nuevas tecnologías en los estudios filológicos,* 1999, ISBN 84-89790-41-8, págs. 45-81, 45–81. https://dialnet.unirioja.es/servlet/articulo?codigo=595883.

Marín, M. J., & Camino, R. 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. https://doi.org/10.18172/jes.184.

Ooms, J. [aut. & cre. 2022. *pdftools: Text Extraction, Rendering and Converting of PDF Documents* (3.3.3). https://CRAN.R-project.org/package=pdftools

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D., Posit, & PBC. 2023. *dplyr: A Grammar of Data Manipulation* (1.1.0). https://CRAN.R-project.org/package=dplyr

Wickham, H., & RStudio. 2022. *stringr: Simple, Consistent Wrappers for Common String Operations* (1.5.0). https://CRAN.R-project.org/package=stringr

ABSTRACTS RECEIVED IN APRIL 2023

PAPERS

POSTER

Medical Discourse Translation During COVID-19:
A Case Study of Translating Medical Discourse into Arabic

Asmaa Alduhaim – Gulf University for Science & Technology

Keywords*: COVID-19, neologism, Arabicization***,** *borrowing, medical translation.*

"Scientific and technical translation has always played a pivotal role in disseminating knowledge." (Krein-Kühle, 2003, p. 1). Scientific translation is now, quite clearly, an important field in translation in general, and requires both accuracy and knowledge of the field. Medical discourse has always been challenging for translators with regard to translating terminologies from one language to another. The challenge increases when the source text and the target language fundamentally differ, as in the case of Arabic and English. This occurs every day in parallel with the spread of diseases and pandemics that lead eventually to pharmacological discoveries and, consequently, the creation of new medical terminologies.

This research aims to tackle the problems of translating medical terms from English to Arabic, particularly terms related to the COVID-19 pandemic. It first examines medical terminology in English, considering how such terms are coined or created, and later sheds light on the importance of neologism in medical discourse and how untranslatable some terms are. Although English maintains itself as the *lingua franca* of science and medicine, many researchers have examined the importance of translating medical terminology into other languages and creating new equivalents instead of borrowing foreign words directly. This dominance of the English language in the medical field inevitably results in English terminology such as *AIDS*, *virus*, *bacteria*, and *influenza* being used in other languages, including Arabic. According to Hassan (2017), scientific translation is a crucial channel of knowledge dissemination, but which is extremely scarce in Arabic, and which is not necessarily keeping pace with the explosion of global knowledge. Perhaps the most likely reasons behind the difficulty in translating medical terminology into Arabic are as follows: 1) Arabic is not flexible with regard to borrowing morphemes or words from other languages; 2) many such terms are the product of the West, and therefore carry their inventors' Western names; and 3) there is no agreement within the Arab regions regarding the use of these terms (Elmgrab, 2011). Thus, this study focuses on a number of recently coined and newly created words that were recently incepted due to the emergence of the newly discovered disease, COVID-19. These terms are associated to some greater or lesser extent with this Coronavirus, and are used in various platforms such as newspapers, news broadcasts, Twitter and, of course, medical reports.

The data collected consists of a number of terminologies associated with Coronaviruses, such as *Covidiot*, or has been widely used during the pandemic, for instance, *infodemic'*. The research aims to examine the origin of the term, and how it was firstly coined in English. The analysis will later compare its various Arabic translations and examine them. It will highlight the different methods used to translate the terms such as Arabicization, transliteration, and description. The study concluded that the majority of these terminologies are translated using a descriptive method, or Arabicization. The study further highlights the importance of creating a consistent medical terminology base in the Arabic region for translators to refer back to.

Bibliography

Elmgrab, Ramadan Ahmed. 2011. Methods of Creating and Introducing New Terms in Arabic. *IPEDR-International Proceedings of Economics Development and Research*, *26*, 491-500.

Hassan, Bahaa-eddin. 2017. Translating Scientific Terminology: Examples from the Arabic versions of Two International Magazines. *Mediterranean Journal of Social Sciences*, *8*(2), 183-183.

Krein-Kühle, Monika. 2003. Equivalence in Scientific and Technical Translation. A Text-in- Context-based Study, PhD Thesis, University of Salford.

A corpus-based bilingual glossary for translation in the legal domain

Assunta Caruso – University of Calabria

Keywords: *legal terminology, legal translation, corpora, bilingual glossary.*

The use of corpus linguistics in terminology work and technical translation has been largely advocated and adopted by scholars over the years. The law is rich in specialized terminology, and legal translation is acknowledged as a daunting and time-consuming task (do Céu Bastos, 2020). Many corpora have been created to facilitate the process of translating legal texts. Fan and Xunfeng (2002), for example, documented the use of a bilingual corpus of Chinese and English law to assist Translation students at a Hong Kong university, while Pontrandolfo (2012) focused on English, Spanish and Italian corpora and research methods in legal translation studies. The advances in corpus linguistics and the wide use of corpora in terminology and translation have also led to corpus-based terminology projects for translation purposes.

The objective here is to present a specialized, synchronic, corpus-based bilingual glossary (English- Italian) of legal terms belonging to a variety of sub-domains (human rights, global terrorism, intellectual property, illegal trafficking) created by third-year Law majors taking an English for Specific Purposes course at the University of Calabria, Italy, which was then used and evaluated by Linguistic Mediation students at the same university.

The Law students, therefore, compiled and analyzed specialized corpora which led to the promotion of data-driven discovery learning, the enhancement of legal English lexicon acquisition and the development of a bilingual glossary of legal terms. Terminological resources, especially those that include comparative information retrieved from reliable sources and which describe concepts in their natural context, are essential to the quality of a translation and particularly important in areas such as legal translation. As stated by Diaz Torres (2021), legal translation is considered to be one of the most challenging activities when it comes to mediating between two or more culturally-different parties. The glossary is based on a corpus of authentic, authoritative texts including laws, regulations and treaties. Corpus compilation and term extraction were carried out in the following stages: (i) research into the various specialized sub-domains; (ii) parallel and/or comparable corpora compiled from the selected texts; (iii) set of texts from the corpus processed by means of the freeware software *AntConc* in order to obtain a full list of words present in the corpus; (iv) candidate terms selected and used to extract relevant terms; (v) glossary of legal terms developed using a variety of functions available in the corpus analysis toolkit.

Surveys carried out by Durán-Muñoz (2010, 2012), have pointed out that translators prefer digital bilingual to monolingual resources, that they need clear definitions and require domain specification and context. Therefore, extensive conceptual information was included such as definitions and explanatory contexts as well as information such as synonyms, collocations and phraseology.

Bibliography

Bastos, M. do C. 2020. Legal Terminology for Translators: Company Law. A Bilingual Corpus- Driven Project. *POLISSEMA – Revista De Letras Do ISCAP*, (20), 64–86.

Dias Torres, D. M. 2021. Glosario bilingue de términos legales basado en corpus para perfeccionar la traduccion de textos legales. *Pedagogia y Sociedad*, 24 (61), 174-193. Retrieved from: http://revistas.uniss.edu.cu/index.php/pedagogia-y-sociedad/article/view/1199

Durán-Muñoz, I. 2010. Translator's Needs into account: A Survey on Specialised Lexicographical Resources. In S. Granger & M. Paquot (Eds.), *eLexicography in the 21st century. New Challenges, new applications. Proceedings of eLex 2009*, 55–66. Presses Universitaires de Louvain.

Durán-Muñoz, I. 2012. Meeting translators' needs: translation-oriented terminological management and applications. *The Journal of Specialised Translation*, 18, 77–92. Retrieved from http://www.jostrans.org/issue18/art_duran.pdf

Fan, May & Xu, Xunfeng 2002. An evaluation of an online bilingual corpus for the self-learning of legal English. *System.* 30, 47-63.

Pontrandolfo, G. 2012. Legal Corpora: an overview. *Rivista Internazionale di Tecnica della Traduzione*, 14, 121-136.

Identifying parliamentary sub-genres/sub-registers across languages and cultures: a case study on the ParlaMint corpora

Anna Kryvenko[1] & Petya Osenova[2] – Institute of Contemporary History / NISS[1] & Bulgarian Academy of Sciences / Sofia University[2]

Keywords: *corpora of parliamentary records, metadata, parliamentary sub-genres and sub-registers, contrastive studies, Bulgarian, Ukrainian.*

Parliamentary discourse as a genre/register of political discourse has been an under-researched area in linguistics, with a traditional bias toward the English language (Ilie 2010). Research on parliamentary discourse was not even included into Barbieri and Wizner's (2019) major survey of important register and genre studies published over the past few decades. The recent growing number of parliamentary corpora in languages other than English opens new avenues and poses new challenges for genre/register studies, especially from a contrastive perspective.

This contribution aims to present a case study on those features of corpus design and compilation that are relevant for identifying differences and similarities in sub-genres/sub-registers of parliamentary discourse in corpora of parliamentary proceedings. The data come from the ParlaMint family of comparable specialized corpora of parliamentary records for 31 European countries and regions[6]. The transcript files commonly represent individual session days and are segmented into utterances, i.e. stretches of spoken language produced by individual speakers. The ParlaMint corpora are annotated within a dedicated TEI schema[7]. The schema is genre (and register) neutral, yet rich in metadata related to speakers, parties and utterances in addition to the linguistic annotation. Parliamentary records are seen here as a special speech related text-type (after Culpeper and Kytö 2010).

However, while the ParlaMint corpora are unified in their format and types of metadata included (Ogrodniczuk et al., 2022), they differ in their content with respect to sub-genre characteristics, which has implications for contrastive analysis. First, the authors discuss the limitations of the contrastive sub-genre studies of parliamentary discourse based on corpora of parliamentary proceedings. These limitations are due to: a) access only to the 'frontstage' performances (Wodak 2009, p.4) as manifested in parliamentary records, which do not embrace all types of parliamentary activities; b) a degree of variation in records of plenary sittings across corpora (confinement to a particular type of debates, inclusion or exclusion of ceremonies, etc.); c) variation in parliamentary transcripts (manually/automatically corrected, repertoire of transcriber comments); and d) variation among underlying entities (unicameral vs. bicameral parliaments, different parliamentary systems). The authors also reflect on constraints in the compilation of the ParlaMint corpora due to problems with the availability of some data and metadata.

Second, the authors examine the utility of speaker metadata and transcriber comments for identifying parliamentary sub-genres and sub-registers. In particular, they use Ilie`s (2015) typology of parliamentary sub-genres as a starting point and report on variation in chairpersons' conventional patterns of self-identifying sub-genres (in terms of Biber 2012: 193) in the records. From a register perspective, they account for situational variation with respect to speaker roles (MPs, state officials, guests) and incident elements (vocal markings like laughter or kinetic markings like clapping). The Bulgarian and Ukrainian ParlaMint corpora are featured in this contribution in comparison with the British ParlaMint corpus.

Finally, the authors draw their preliminary conclusions about sub-genre/sub-register variation in the Bulgarian, Ukrainian and British ParlaMint corpora based on the featured discussed and outline further research from a contrastive perspective.

[6] https://www.clarin.eu/parlamint
[7] https://clarin-eric.github.io/parla-clarin

References

Barbieri, F., & Wizner, S. 2019. Appendix I. Annotations of major register and genre studies. In D. Biber, & S. Conrad, *Register, Genre, and Style* (2nd ed., pp. 318-349). Cambridge University Press.

Biber, D. 2012. Register and Discourse Analysis. In Handford, M., & Gee, J.P. (Eds.), *The Routledge Handbook of Discourse Analysis* (1st ed., pp. 191-208). Routledge.

Culpeper, J., & Kytö, M. 2010. *Early Modern English Dialogue: Spoken Interaction as Writing*. Cambridge University Press.

Ilie, C. 2010. Identity co-construction in parliamentary discourse practices. In C. Ilie (Ed.), *European Parliaments under Scrutiny: Discourse Strategies and Interaction Practices* (1st ed., pp. 57–78). John Benjamins.

Ilie, C. 2015. Parliamentary discourse. In: K. Tracy, C. Ilie, & T. Sandel (Eds.), *The International Encyclopedia of Language and Social Interaction*. Wiley-Blackwell. https://doi.org/10.1002/9781118611463.wbielsi201

Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešic, N., Çöltekin, Ç., Kopp, M., & Meden, K. 2022. ParlaMint II: The Show Must Go On. In *Proceedings of the LREC 2022 ParlaCLARIN III,* Workshop on Creating, Enriching and Using Parliamentary Corpora (pp. 1–6). http://www.lrec-conf.org/proceedings/lrec2022/workshops/ParlaCLARINIII/2022.parlaclariniii-1.0.pdf

Wodak, R. 2009. *The Discourse of Politics in Action: Politics as Usual* (1st ed.). Palgrave.

***It takes two to tango, SO TO SPEAK*: A corpus-informed study of phraseology markers and breakers in English and Polish**

Łukasz Grabowski[1] & Piotr Pęzik[2] – *University of Opole*[1] – *University of Łódź*[2]

Keywords: *phraseology markers, phraseology breakers, novelty markers, formulaic language.*

One of the challenges faced by corpus phraseology, lexicography, natural language processing or machine translation, is to devise better methods of identifying phraseological expressions (henceforth PEs). In this paper, we focus on a special class of lexical devices which are used naturally to signal prefabricated language in texts which can be helpful in the task of verifying the status of a potential PE.

The criteria used to extract instances of linguistic prefabrication from texts are varied, as they include formal-linguistic, frequency-driven or distributional and psycholinguistic ones, which also correspond with various perspectives on formulaic language. (Wray 2002, 2007; Wood 2015). However, it is also possible to identify PEs using a distinct class of linguistic markers found in the immediate co-text or punctuation marks found in near proximity to instances of PEs, a phenomenon which has not been comprehensively explored by linguists in Anglophone countries. For example, Chlebda (2010), Ruiz-Gurillo (2015) argue that there are many orthographic and linguistic devices, including punctuation marks (e.g., quotation marks), routine or conversational formulas (e.g., *as they say, so-called, so to speak, (the) proverbial n*), that can function as textual indicators of PEs.

Thus, in this corpus-informed cross-linguistic study, we focus on 'phraseology markers' (PMs), which are recurrent and fixed word combinations used to demarcate instances of linguistic prefabrication. Using selected corpora of general and spoken English (*so to speak, so to say, as it were*) and Polish (*by tak rzec, że się tak wyrażę, że tak powiem*), we study the use and discoursal functions of three pairs of formally, semantically and pragmatically similar phraseology markers and attempt to determine the amount of prefabricated language demarcated by those linguistic items in spoken and written texts.

The findings revealed the studied phraseology markers perform two opposing functions with respect to the use of formulaic language. On the one hand, they are used to mark conventional, prefabricated expressions that contribute to the formulaicity of a speaker's utterances. On the other hand, we have identified contexts where they mark expressions or phrases that represent unusual, unconventional, idiosyncratic phrasings unattested or rarely used in native texts. Hence, paradoxically, the phrasings usually referred to as phraseology markers occasionally perform the function of 'phraseology breakers', where the fixed or canonical form of a phraseological unit is modified yet it is still possible for speakers or writers to intuitively trace back to its very source, i.e. the canonical form. In addition, those phrasings may also be used to signal novel and idiosyncratic, unconventional word combinations, rarely or never attested in language corpora, and hence, we may also refer to them as 'novelty markers'. Finally, taking the English phrase *so to speak* as a case in point, we focused on its pragmatic function of phraseology marker and attempted to measure how much prefabricated language that English phrase marks when used in the Spoken component of the COCA 2020 corpus. It was found to be in the range of 14-25% of its total number of occurrences (a confidence interval at the level of 95%).

Overall, we believe that the study findings cast new light on the pragmatic functions of so-called phraseology markers, which can also come in useful as a complementary method for measuring the degree of phraseological prefabrication in texts.

References

Chlebda, W. (2010). Nieautomatyczne drogi dochodzenia do reproduktów wielowyrazowych. In: W. Chlebda (Ed.), *Na tropach reproduktów: w poszukiwaniu wielowyrazowych jednostek języka*. Opole; Wydawnictwo Uniwersytetu Opolskiego, 15-35.

Ruiz-Gurillo, L. (2015). "Phraseology of humor in Spanish: Types, functions and discourses". *Lingvisticae investigaciones: International Journal of Linguistics and Language Resources*, 38(2): 191-212.

Wood, D. (2015). *Fundamentals of Formulaic Language*. London: Bloomsbury.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic language. Pushing the boundaries*. Oxford: Oxford University Press.

Evaluation as co-selection and discursive construction of Covid-19 pandemic in American media discourse: A diachronic perspective

Meijuan Liu & Min Dong – *University of Beijing*

Keywords: *co-selection view of evaluation; discursive construction of Covid-19 pandemic; American media discourse; diachronic study.*

Previous evaluation studies have been conducted via Appraisal System, of which attitude system and affect system are frequently explored. They mainly examine the Evaluative Category itself based on evaluative resources such as adverbials, adjectives, and nouns. Therefore, this study adopts an evaluative co-selection view, arguing that appraisal is instantiated by co-selections of both attitudinal target and attitudinal lexis. Meanwhile, this study takes the American media as a case study to explore features of the discursive construction of the COVID-19 pandemic discourse at different stages based on the co-selection patterns between the evaluative parameters and target parameters. This study learns from the Appraisal System and adopts a parameter-based framework to build the analytical framework of evaluative parameters, which includes parameters such as Capacity and Evidentiality. Each parameter is further divided into positive and negative poles. Meanwhile, target parameters are formulated based on the semantic classification of high-frequency words retrieved from the diachronic corpus of COVID-19 pandemic news discourse in American mainstream media, namely DCCOVID-19NDAm corpus. All the sentences containing the word "coronavirus" are extracted and 10 percent of them are taken as a sample. These sample sentences are imported into the corpus tool MAXQDA and evaluative parameters and target parameters act as the annotation codes. These sentences are annotated by two linguists at the same time.

Major findings of the present research are as follows.

Firstly, in terms of targets, generally speaking, the total number of objects evaluated in stages 1, 2, and 3 is much higher than that of stages 4 and 5. This is explained by the truth that the first three phases witness the sudden outbreak of the COVID-19 pandemic and its unexpectedly transmissible variants, and accordingly, the media compete to cover coronavirus-related topics to capture people's attention. By the fourth stage, coronavirus has become a normal problem, and reports related to coronavirus have lost their novelty. As a result, journalists are drawn to other issues, such as the Russia- Ukraine war, and fewer news pages are allocated for the pandemic. Specifically speaking, first, the most evaluated target changes across different stages. Covid-19 spread and its prevention is the most dominant target in stage 1. Stages 2 and 3, instead, are dominated by politics inside America. The nature of the pandemic has changed from a public health emergency of international concern to a political event. The American government and former president Trump manipulate the disease to win people's support and to demonize China for taking responsibility for the disease. Second, the frequency of research peaks in stage 2 which begins in April 2020. Thus, it can be concluded that firstly, as more money and efforts are invested in investigations, the scientific world has learned about the virus, produced powerful products, and identified ways to prevent and treat the disease; secondly, the media often cite scientific studies to enhance their objectivity and credibility.

Regarding the evaluative parameters, pos-capacity, pos-tenacity, pos-evidentiality, pos-security, neg-satisfaction, neg-happiness, neg-normality, and neg-security are mostly connected to different targets. Among them, neg-normality frequently evaluates the mutation and disastrous effect caused by the coronavirus. Pos-evidentiality closely links with scientific research. Pos-capacity is used to appraise people's behaviors. Other parameters are commonly employed to convey people's feelings or opinions toward an entity or a proposition. To be specific, the abnormality of Covid-19 is evaluated the most in stage 1. Stage 2, instead, uses pos-capacity the most frequently to depict the American government and politicians. It is inferred that though American media claim to be objective

and independent, they are the voice of the government and help the American politicians shirk their responsibilities for the outbreak of disease, maliciously deem China the culprit of the pandemic, make China a scapegoat, and build an image of a capable government. Meanwhile, they create an irresponsible image of China. Furthermore, the semantic realization of evaluative parameters also has prominent features. Taking the evaluative parameter of capacity as an example, it is mostly linked with activity verbs such as "address, build, and ship". Meanwhile, EPs of happiness, satisfaction, security, and tenacity are realized by mental verbs and verbal verbs such as "anger, accuse, blame, criticize and claim". EP pos-evidentiality is commonly represented by relational verbs such as "suggest and show", and verbal verbs such as "explain and report".

Secondly, the discursive construction of the COVID-19 pandemic across different stages of news reports in American mainstream media is discussed based on the distribution of co-selection patterns between targets and evaluative parameters in the DCCOVID-19NDAm corpus. Images of the COVID-19 pandemic, companies and ordinary people, and the American government are constructed. First of all, the COVID-19 pandemic is metaphorized as an external, ruthless, and lethal killer who is more frightening than the contagious SARS. Some American media are promoting "the weaponization of the virus" and portraying the coronavirus as a Chinese-made weapon, which is completely contrary to the facts. Meanwhile, the vaccine is deemed helpful and acts as a shield against the attack of coronavirus. Chinese-made coronavirus supplies such as masks and vaccines are offered to many other countries. Second, American media construct responsible images of companies and researchers because they work hard to prevent and cure the disease. They are calm and brave fighters against the disease. Third, a vivid image of the public is built, that is, always struggling in the face of suffering. Fourth, the images of the American government are multi-faced and mixed feelings are expressed by the media and people. American governments are portrayed as capable and far-sighted leaders and the victim of coronavirus with China being demonized by Trump's so-called "Wuhan coronavirus". Moreover, the Trump administration has stirred up much more controversy than the Biden administration because Trump has wrongly downplayed the seriousness of the coronavirus. In contrast, Biden has held a consistent policy on coronavirus. In addition to all the positive evaluations, some negative aspects are reported. For example, in stage 3, the evaluation of Covid spread shows that China has largely cut the number of confirmed cases, while the United States has not successfully dealt with the disease, leaving its people still affected by the disease. Next, the U.S. government discriminates against some specific groups, including Black Americans and the LGBT community, by limiting their access to basic media.

Using corpus linguistics in interpreting studies:
A research project on simultaneous interpreting at the United Nations

Monika Stögerer – *University of Vienna*

Since the relevance of corpus linguistics for translation studies as an emerging scientific discipline in the early 1990s was highlighted by Baker (1993), corpus-based approaches became soon afterwards centre of interest also in interpreting studies (Shlesinger 1998). Since then, corpus-based interpreting studies (CIS) hold an enormous potential for interpreting research, even though it was also observed critically due to a lack of profound theory (Setton 2011, pp. 34-35).

While in the beginning, smaller, manually transcribed and analysed corpora were collected by individual researchers, nowadays, large, machine-readable interpreting corpora often administered by a research team, exist. Such corpora of authentic interpreting situations exist nowadays for different interpreting modes such as simultaneous and consecutive interpreting (e.g. EPIC Corpus (Sandrelli & Bendazzoli 2005), FOOTIE Corpus (Sandrelli 2012), or CorIT Corpus (Falbo 2012).

In the context of international organisations, simultaneous interpreting has been investigated so far mostly within the European institutions thanks to publically available audio material. Interpreting in other international institutions, such as the UN bodies, has not received a lot of attention from corpus-based interpreting studies until now. In this research project, a bilingual parallel corpus of authentic source speeches as well as their simultaneous interpretations was built from scratch. For this purpose, 64 English source speeches with a minimum length of six minutes each and the respective simultaneous interpretations into French were transcribed, aligned and POS-tagged using different software tools. This newly compiled *UNIC* Corpus was then analysed, guided by the principles of a descriptive approach (Toury 1995). Therefore, in this product-oriented research project, the goal has been set to gain insight in characteristics of simultaneous interpreting in UN conferences. For this purpose, the software Sketch Engine was used. Analysis was conducted via different features for corpus query included in the tool. Among others, the keyword extraction and the parallel concordance tool proved to be more than useful. Methodologically, linguistic features like the syntactic position of the French adverb were observed under the strong influence of English source texts. Considering that interpreting as a communicative act has to be viewed within its individual context, results gained from this project were compared with large reference corpora of French language (CRFC (Siepmann et al. 2017), one of its subcorpora consisting of spoken language provided by Dirk Siepmann, and the UN Parallel Corpus included in Sketch Engine) to detect characteristics proper to French language use within the UN context.

References

Baker, Mona 1993. Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (eds) *Text and Technology: In Honour of John Sinclair.* Amsterdam: John Benjamins, 233–250.

Falbo, Caterina 2012. *CorIT (Italian Television Interpreting Corpus): classification criteria.*

Sandrelli, Annalisa 2012. Introducing FOOTIE (Football in Europe): Simultaneous interpreting in football press conferences. In F. Straniero Sergio & C. Falbo (eds) *Breaking Ground in Corpus-Based Interpreting Studies.* Frankfurt: Peter Lang, 119– 153.

Setton, Robin 2011. Corpus-based Interpreting Studies (CIS): Overview and prospects. In A. Kruger, K. Wallmach & J. Munday (eds) *Corpus-based Translation Studies: Research and Applications.* London/New York: Continuum, 33–75.

Shlesinger, Miriam 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta* 43 (4), 486–493.

Siepmann, Dirk; Bürgel, Christoph & Diwersy, Sascha 2017. The corpus de reference du français contemporain (CRFC) as the first genre-diverse mega-corpus of French. In*: International Journal of Lexicography* 30 (1), 63-84.

Toury, Gideon 1995. *Descriptive Translation Studies – and beyond.* Amsterdam: John Benjamins.

Breach of *pacta sunt servanda*: The AUKUS agreement and evaluation in newspaper discourse

Radoslava Trnavac[1] & Encarnación Hidalgo Tenorio[2]
The National Research University Higher School of Economics[1] – University of Granada[2]

Keywords: *AUKUS, Corpus-assisted Critical Discourse Analysis, Sentiment Analysis, Appraisal Theory*

The AUKUS agreement is a trilateral military partnership signed by the US, the UK and Australia on 15 September 2021, to protect a rules-based international order and preserve security in the Indo-Pacific. As a result of the agreement, Australia violated the norm *pacta sunt servanda* (agreements must be kept) by scuttling its $90 billion deal with the partially French state-owned Laval Group; subsequently, it triggered the deepest diplomatic crisis between France and Australia since France undertook nuclear testing in the Pacific (Stauton and Day 2022: 1).

Against this backdrop, in this paper we intend to analyze evaluative language in the news items on AUKUS published in Asian and Anglo-speaking countries from September 15, 2021, until October 31, 2021. In pursuit of our goal, we focus on the manifestation of inequality and power structures in international politics, combining both corpus-driven and corpus-based analyses. The corpus-driven approach is aimed toward the identification and classification of keywords into main themes. On the other hand, the corpus-based approach is grounded in Sentiment Analysis alongside the exhaustive inspection of evaluative patterns in texts. We produce the automated Sentiment Analysis with *Lingmotiff* (Moreno-Ortiz 2017) and *SEANCE* (Crossley, Kyle, and McNamara 2017) to get the text's overall sentiment score, the proportion of sentiment vs. non-sentiment items, the quantity and the type of polarity items, and the quantity of social order and affective items. In line with Coffin and O'Halloran (2005, 2006), we combine the interpretation of evaluation in the entire corpus, with the intensive Appraisal analysis of only a small portion of the texts. For the manual annotation, we employ *UAM CorpusTool* (O'Donnell 2008) with a redefined version of Martin and White's (2005) scheme (see Bednarek 2009).

Both components of our corpus share some similar results regarding topic: They speak about the same issues, such as politics, countries/regions/cities, technology, and quality; and, in general, they all refer to AUKUS rather positively. Nevertheless, the Anglo sub-corpus discusses different political positions regarding the agreement and possible outcomes, and is more dialogic than the Asian sub-corpus, which is reflected in its prominent use of modality markers. Additionally, in the Anglo sub-corpus, intensification and emotions are significantly more widely distributed, which generates more dramatic discourse around the agreement, while opinion dominates in the Asian sub-corpus. Furthermore, whilst in the Asian sub-corpus the features of trust and propriety are more relevant, the Anglo sub-corpus focuses on judgement, in particular international actors' tenacity, and abounds in social order terms, such as need and rectitude. Finally, as for what entities are most frequently appraised in each, in the Anglo sub- corpus it is alliances, politicians and the Chinese challenge to the US's power; as for the Asian sub-corpus, it is focused on the political cooperation between East Asia countries, and attempts to balance their geostrategical position between the US and China.

References

Coffin, C. and K. O'Halloran. 2005. "Finding the Global Groove: Theorising and Analysing Dynamic Reader Positioning Using Appraisal, Corpus, and a Concordancer." *Critical Discourse Studies* 2 (2): 143–163.

Coffin, C. and K. O'Halloran. 2006. "The Role of Appraisal and Corpora in Detecting Covert Evaluation." *Functions of Language* 13 (1): 77–110.

Bednarek, M. 2009. "Language Patterns and Attitude." *Functions of Language* 16 (2): 165–192.

Crossley, S.A., Kyle, K. and D. McNamara. 2017. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behaviour Research Methods* 49: 803-821.

Martin, J. R. and P. R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Basingstoke: Palgrave Macmillan.

Moreno-Ortiz, A. 2017. Lingmotif: Sentiment Analysis for the Digital Humanities. *Proceedings of the EACL 2017 Software Demonstrations*, Valencia, Spain, April 3-7 2017, pp. 73-76.

O'Donnell, M. 2016. UAM CorpusTool http://www.corpustool.com/.

Stauton, E. and B. Day. 2022. Australia-France Relations after AUKUS: Macron, Morrison and Trust in International Relations. *Australian Journal of International Affairs*, 1-8.

Metodología para crear un corpus paralelo de informes financieros: conversión, limpieza y alineamiento

Sofía Roseti, Yanco Torterolo, Blanca Carbajo Coronado & Antonio Moreno Sandoval
Universidad Autónoma de Madrid

Keywords: *compilación, corpus paralelo, corpus bilingüe español-inglés, dominio financiero.*

En este trabajo mostramos una metodología para crear un corpus paralelo español-inglés a partir de informes en formato PDF. El objetivo final es poder alinear las oraciones correspondientes de cada lengua, en un texto limpio de errores.

La extracción y estructuración de la información siempre supone uno de los mayores problemas a la hora de compilar un corpus que procede de diferentes fuentes y formatos. Precisamente, uno de los formatos más problemáticos es el PDF. Especialmente en los casos donde se presenta la información en varias columnas y con una organización no lineal que dificulta la extracción del texto de manera automática. Además, pueden incluir cuadros con texto que interrumpen el texto principal. Otros elementos que conviene tener en cuenta son las notas al pie, el número de página, encabezamientos, gráficas, etc.

Este es el caso de los informes financieros anuales en PDF de las principales empresas del IBEX 35, que son la fuente del corpus paralelo español-inglés. Estos informes tienen una manera muy visual de presentar la información, pero impide la extracción automática del texto.

En primer lugar, mostraremos la solución encontrada al problema: la manipulación del propio PDF previa a su exportación al formato TXT. Para ello, empleamos Adobe Acrobat Pro, con la función "Editar PDF". La idea central es quitar aquellos elementos que obstaculicen el procesamiento automático de la narrativa principal y reorganizar los elementos textuales de tal manera que sigan una diagonal imaginaria que nace en la esquina superior izquierda y se dirige hacia la esquina inferior derecha. Se empieza borrando los elementos no necesarios de las páginas (encabezamientos, número de página) para continuar eliminando figuras, tablas y secciones concretas que no contienen narrativa; solo datos numéricos. Una vez eliminadas las partes prescindibles del texto, hay que restructurar los bloques de contenido de manera escalonada (rompiendo las columnas dobles o triples) y colocando el texto secuencialmente mediante el reajuste de los bloques de texto. En otras palabras, se maqueta de nuevo el documento para producir una versión en una única columna de texto corrido.

Una vez producida la conversión PDF a TXT hay que realizar la limpieza de los textos, pues el resultado ha pasado por un OCR que genera ruido: eliminación de espacios entre palabras, partición inadecuada de las oraciones, reconocimiento erróneo de caracteres, etc. Mostraremos diferentes expresiones regulares en Python para solucionar estos problemas.

Con los textos limpios, en español e inglés, procedemos al último paso: el alineamiento de las oraciones. Usamos la aplicación de Trados, que nos obliga a partir los textos de más de 3000 oraciones para procesarlos por partes que luego se tienen que unir. Importamos la alineación a la memoria de traducción y exportamos el TMX resultante, para finalmente convertir el TMX en un CSV, lo que permite usarlo de manera tanto monolingüe como bilingüe.

Corpus Linguistics in the Digital Era: Genres, Registers and Domains
La lingüística de corpus en la era digital: géneros, registros y dominios

14th International Conference on Corpus Linguistics - May 10 - 12, 2023
XIV Congreso Internacional de Lingüística de Corpus – Oviedo, 10 a 12 de mayo de 2023