BOOK OF ABSTRACTS

aelinco

IULMA

VNIVERSITAT DE VALÈNCIA Departament de Filologia Anglesa i Alemanya

VNIVERSITAT DE VALÈNCIA Facultat de Filologia, Traducció i Comunicació

GENERALITAT VALENCIANA
Conselleria d'Educació, Investigació, Cultura i Esport

SPECIAL THEME: CORPUS APPROACHES TO DISCOURSE ANALYSIS

Organising Committee:

Miguel Fuster-Márquez
José Santaemilia-Ruiz
Carmen Gregori-Signes
Paula Rodríguez-Abruñeiras

Scientific Committee

The abstracts contained in this booklet are in principle in the form in which they were submitted; because text was sometimes corrupted by the transfer from text processing files to the online platform, editing was often necessary for consistency and intelligibility.
In the course of March 2019, authors were given the opportunity to edit their abstracts themselves. Yet, the end result may still not be fully stisfactory in all cases. This collection of abstracts represents the state of affairs on 1 April 2019. Later cancellations may be responsible for the presence of abstracts that do not correspond to anything in the final programme launched on the website on 13 May 2019.

The abstracts have been listed alphabetically in accordance with the lastname of the first author.

CILC2019 PROGRAMME (13 MAY 2019)

CONTENT

PLENARY LECTURES

Developing Resilience against Online Grooming: A Corpus and Discourse Studies Approach

Nuria Lorenzo-Dus

Applied Linguistics, Swansea University, UK

One of the main characteristics of the current digital era is a propensity – some argue, a compulsion – to share online: from selfies and other files to personal views and experiences. We are indeed said to have experienced an evolutionary 'sharing turn' (Grassmuck 2012) – one that has seen us move away from being rather passive consumers of media and into becoming active co-producers thereof (Jenkins 2014). This turn is underpinned by a cultural rhetoric that constructs digital sharing not only as pervasive but also – and primarily – as good for our wellbeing and identity (John 2018). At the same time, there are serious concerns about the risks of over-sharing digitally, including in terms of privacy infringement and outright shattering of the trust that we place in those with whom we share online. It is this latter, more sinister aspect of the sharing turn that I engage with in this talk. A digital context marked by criminality is selected to this end: online child sexual grooming. This is a communicative process of entrapment in which an adult uses language and other semiotic modes (e.g. images) to lure a minor into taking part in sexual activities online and, at times, also offline (Lorenzo-Dus et al 2016; Chiang and Grant 2018).

During my talk I will discuss the key results of a series of inter-related studies conducted under the auspices of the *Developing Resilience against Online Grooming* Project at Swansea University. The results emerge from an integration of corpus and discourse analysis of >600 online sexual grooming conversations (approx. 3.7 million words by convicted predators) into Machine Learning models. The corpus and discourse analysis results are also translated – through stake-holder co-creation methods – into practical measures aimed at detection and prevention of online child sexual grooming. Therefore, during my talk I will reflect upon how Corpus and Discourse Studies may complement – and occasionally be at loggerheads with! – the premises and methods used in other disciplines, as well as the expectations of stake-holders. Ultimately, my presentation will highlight the contributions that Corpus and Discourse Studies can make to collective endeavours to address societal challenges.

References

Chiang, E. and Grant, T. (2018). Deceptive identity performance: Offender moves and multiple identities in online child abuse conversations. *Applied Linguistics,* 1-25

Grassmuck, V. (2012) The sharing turn: Why we are generally nice and have a good chance to cooperate our way out of the mess we have gotten ourselves into, in W. Sützl *et al* (eds.) *Media, Knowledge and Education: Cultures and Ethics of Sharing,* 17-34, Innsbruck: Innsbruck University Press.

Jenkins, H. (2014) Rethinking 'rethinking convergence/culture', *Cultural Studies* 28, 267-297.

John, N. (2018) *The Age of Sharing*, Cambridge: Polity.

Lorenzo-Dus, N., *et al* (2016). Understanding grooming discourse in computer mediated environments. *Discourse, Context & Media*, 12: 40-50.

Using Corpus Data to Explore Discourses of Climate in the Past

Tony McEnery

Lancaster University

This talk will use corpus linguistics in concert with a number of other methods to explore whether we can explore physical reality in the past through a corpus. Previous projects I have been involved with have looked at how corpus data may allow access to social attitudes in the past, but in this talk I will look at how a corpus may allow us access to information about the environment by exploring the degree to which a corpus can tell us about extreme weather events in the past, in this case droughts and water shortages in the nineteenth century. In order to do this, one has to both overcome a series of issues with the available data itself and to establish, as far as is possible, some ground truth regarding past climate. In doing this, the work presented relies on close reading analyses undertaken by other researchers, using that qualitative work of hydrologists as the grounds on which the corpus based analysis gains credence through triangulation. Given the goal of the work, and the data available, triangulation was not simply a methodological option, it was a necessity.

The talk will review the data available for the study and, based on the difficulty of using that data, move to explore the methods used in concert with corpus analysis: close reading to establish data quality, a review of the available records of droughts and water shortages undertaken by hydrologists and geoparsing. Following the introduction of these methods, I will explore how closely the corpus analysis matches what is known about droughts in the nineteenth century before exploring some new droughts which may have been revealed through the corpus analysis.

That *Other* Kind of Cohesion. A CaDS Outline of *Evaluative* Cohesion.

Alan Partington

Bologna University

In this talk I wish to concentrate on the "assisted" and the "discourse" parts of the label *Corpus-assisted Discourse Studies*. So don't expect any innovative sophisticated uses of corpus techniques, just the tried and tested.

Most courses in linguistics and discourse will involve a lesson or two on *propositional* cohesion, anaphoric reference being recall of previously mentioned entities and notions and cataphoric cohesion being ways of moving the discourse forwards.

But discourses cohere and, more importantly, are made coherent by readers and listeners *evaluatively*. This should perhaps come as no surprise given how good vs bad evaluation is

one the most basic human instincts and means of socialisation, as noted by Darwin (*The Descent of Man*, 1871), not to mention Shakespeare:

> The web of our life is of a mingled yarn, good and ill together: our virtues would be proud, if our faults whipped them not; and our crimes would despair, if they were not cherished by our virtues.
>
> (*All's Well That Ends Well, A*c IV, Scene III, 69-72)

The first inklings of all this were first glimpsed by Sinclair (1991) and Louw (1993) in their work on what is generally called semantic prosody, though a more descriptive term might be *evaluative prosody*. If we recall, semantic prosody was originally hypothesised through cases where it is absent or, better, when the normal regularities, the habits, of prosody are flouted or exploited by speakers often, but not always, for effect. The role of computers, corpora and concordancing in uncovering the regularities – and shedding light on the hows and whys of prosodic irregularities – is considered one of the crowning contributions of corpus linguistics to our understanding of language.

Evaluative cohesion/coherence can perhaps be viewed as an extension – both conceptually and textually - of the notion of semantic prosody and, likewise, it has always been hiding in plain sight. Most CL and CaDS work on evaluation and evaluative prosody has concentrated on narrow collocations/co-occurrences or, at most, a little beyond to the level of the phrase. Here then I wish to illustrate, via corpus-assisted analysis of a couple of whole texts, how we actually read/interpret/compose evaluations as we move through a text, hoping to underline how we (re)construct evaluations *holistically*, and not word by word or collocation by collocation or even phrase by phrase, but evaluative block by evaluative block, contrary to how we are sometimes taught. Otherwise we soon would get lost in a fog of evaluative mixed messages and disharmony.

Fantastic online comments and how to find them

Maite Taboada

Department of Linguistics, Simon Fraser University, Canada

I provide an overview of my current research on discourse and computational methods to analyze social media language. The first part of the talk will be devoted to outlining the two frameworks for this research: rhetorical relations and sentiment analysis. Rhetorical relations are the fundamental building blocks of discourse, connecting propositions to make coherent text. I will describe existing research on rhetorical relations and present a study on how relations are signalled by discourse markers and other linguistic devices. Then I introduce my work on sentiment analysis, and on the role that rhetorical relations and other contextual factors play in the interpretation of sentiment and opinion.

The second part of the talk is devoted to describing a current project, analyzing online news comments in terms of constructiveness and toxicity. Using a large corpus of comments, I describe how we have modelled constructiveness in terms of rhetorical and argumentation relations, and toxicity as a type of extreme negative sentiment.

PRESENTATIONS

Self-promotional *I* and *we* in political discourse – Spanish politicians in campaign speeches and debates

Ana Albalat-Mascarell

Departamento de Lingüística Aplicada, Universitat Politècnica de València, Spain

***Keywords**: political discourse; pronouns; self-mention; self-promotion; metadiscourse*

Abstract

One tactful strategy in political discourse is the deployment of personal pronouns for rhetorical and self-promotional purposes. Politicians often use the personal pronouns *I* and *we* to promote themselves and the political party they represent. These self- references can be considered metadiscoursal devices in that they reflect the speaker's assessment of the audience's needs and his or her desire to monitor the possible responses of others (Dafouz-Milne, 2008: Fu & Hyland, 2014; Hyland, 2005, 2010,

2015, 2017; Hyland and Tse, 2004; Ilie, 2003; Mur-Dueñas, 2007, 2011). Self- referential pronouns assist politicians in building an *Ethos* of political knowledge and expertise, thus shaping their relationship with their audiences and influencing electorates. Taking Hyland (2005) and Hyland & Tse (2004)'s model of metadiscourse as the analytical framework, this paper aims at quantitatively and qualitatively exploring the use of self-promotional *I* and *we* by Spanish politicians in two different genres of campaign talk, i.e., speeches and debates. My objectives are, first, to compare the frequency of use and distribution of the pronouns *I* and *we* performing a self-referential role in campaign speeches and in a debate; second, to examine the rhetorical functions of the most commonly used forms and relate them to the candidates' persuasive aims; third, to contrast the main differences in the use of these personal pronouns with the outcome of the election. To this end a corpus made up of the transcripts of the 2016 Spanish election campaign speeches and the election debate was compiled, in which four different politicians from the four main political parties in Spain (i.e., the PP, the PSOE, Ciudadanos and Unidos Podemos) seek to convey credibility, evoke authority and thus garner the electorate's support. In the methodology set for this study, the quantitative use of the personal pronouns was carried out with the tool *Metool* developed specifically to detect metadiscourse strategies. The results indicate that the Spanish politicians tend to use more self-referential pronouns in speeches than in a debate. Furthermore, significant differences were found in the distribution and the rhetorical functions of these self-mentions depending on whether they were used in a speech or a debate, depending on the speaker and more specifically, depending on the political affiliation and ideas associated with each speaker. It could be concluded that Spanish politicians publicize themselves and their political parties and colleagues in a very different fashion depending on whether they are giving a campaign speech to a cheering crowd or participating in a debate in front of their adversaries and the wider public. Nonetheless, the speakers' own communicative characteristics and perceived personas also play a role in the ways they project themselves onto their discourse and build a relationship with their audiences.

References

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.

Fu, X., & Hyland, K. (2014). Interaction in two journalistic genres: a study of interactional metadiscourse. *English Text Construction*, 7(1): 122-144.

Hyland, K. (2005). *Metadiscourse.* London, UK: Continuum.

Hyland, K. (2010). Metadiscourse: Mapping interactions in Academic Writing. *Nordic Journal of English Studies*, 9(2): 125-143.

Hyland, K. (2015). Metadiscourse. In K. Tracy (Ed.), *The International Encyclopedia of Language and Social Interaction*, (pp. 1-11). Hoboken, USA: John Wiley & Sons, Inc.

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

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal.

Applied Linguistics, 25(2): 156-177.

Ilie, C. (2003). Discourse and metadiscourse in parliamentary debates. *Journal of Language and Politics*, 2(1): 71-92.

Mur-Dueñas, P. (2007). I/we focus on…: A cross-cultural analysis of selfmentions in business management research articles. *Journal of English for Academic Purposes*, 6: 143-162.

Mur-Dueñas, P. (2011). An intercultural analysis of metadiscourse features in research articles written in English and in Spanish. *Journal of Pragmatics*, 43: 3068-3079.

Variación geográfica de los marcadores evidenciales en el español de España y de América. Un estudio sobre el corpus Ameresco

Marta Albelda Marco

Hispanic Studies Department, Univerisity of Valencia

Abstract

El presente trabajo forma parte de una investigación mayor en la que se está llevando a cabo un mapa geográfico y genérico de los marcadores discursivos evidenciales (*por lo visto, al parecer, se ve (que), dizque*, etc.) del español de España y de Hispanoamérica. El objetivo de esta comunicación es presentar el comportamiento dialectal de los elementos evidenciales que se encuentran en una fase avanzada de gramaticalización en español, a partir de un estudio de conversaciones coloquiales. En concreto, se estudia

(1) la frecuencia de uso en cada zona geográfica de cada uno de ellos, y (2) los tipos de

evidencialidad (sensorial, inferencial y reportativo) que cada marcador expresa en los diversos dialectos estudiados.

Las zonas estudiadas para la presente investigación corresponden a las variedades que hasta la fecha están recogidas, transcritas y volcadas en el corpus Ameresco (América Español Coloquial, www.esvaratenuacion.es). Asimismo, en algunos casos, se complementará con datos obtenidos del corpus de entrevistas PRESEEA (Proyecto para el Estudio del Español de España y de América, http://preseea.linguas.net). En total se han analizado en torno a cinco millones de palabras.

Para el reconocimiento y extracción de los casos, se han recogido todos los casos de los marcadores evidenciales, bien en usos parentéticos o bien en usos fijados o cuasi- fijados gramaticalmente (Traugott y Trousdale 2013). Para el reconocimiento del valor evidencial, se ha seguido una posición funcional de la categoría, de acuerdo con Boye (2010), Hassler (2010) y Kotwica (2017).

Los resultados muestran diferentes frecuencias en la distribución de los distintos marcadores. En primer lugar, se observa una llamativa ausencia del marcador *al parecer* en el registro coloquial, el uso casi exclusivamente americano de *dizque*, y un uso más amplio, a la vez que heterogéneo de *por lo visto* y *se ve que*. En general, se ha obtenido que el evidencial *se ve que* presenta mayor frecuencia en el registro informal que *por lo visto*. También se ha observado que, en general, *se ve que* ofrece mayor frecuencia de uso que *por lo visto*. Por último, se ha obtenido que los modos de conocimiento que expresan son más reportativos en *dizque* y *por lo visto*, y más inferenciales en *se ve que*.

Referencias

Albelda Marco, M. (2016): "La expresión de la evidencialidad en la construcción se ve (que)", Spanish in Context 13.2, p. 237–262.

Anderson, Ll. (1986): "Evidentials, Paths of Change and Mental Maps: Typologically Regular Asymmetries", en Evidentiality: The Linguistic Coding of Epistemology. Advances in Discourse Processes, ed. por CHAFE, W. y NICHOLS, J., Berlin: Mouton de Gruyter, p. 273–312.

Boye, K. (2010): "Evidence for what? Evidentiality and scope", STUF Language Typology and Universals 63/4, p. 290–307.

Cornillie, B. (2016): "Las lecturas evidenciales de los verbos (semi)auxiliares en español", en González Ruiz, R., Izquierdo Alegría, D. y Loureda Lamas, O.

(eds.): La evidencialidad en español: teoría y descripción, p. 227--- 250.

Estellés Arguedas, M. y Albelda Marco, M. (2017): "On the dynamicity of evidential scales. Pragmatic indirectness in evidentiality as a rhetorical strategy in academic and political discourse", en Figueras, Carolina y Cabedo, A. (eds.): Perspectives on Evidentiality in Spanish: Explorations across genres, Amsterdam: John Benjamins.

Fernández Jaén, J. (2012): Semántica diacrónica cognitiva de los verbos de percepción física del español. Tesis Doctoral. Alicante: Universidad de Alicante.

Fuentes Rodríguez, C. (2009): Diccionario de conectores y operadores del español actual, Madrid: Arco/Libros.

González Ramos, E. (2016): "Por lo visto y al parecer: evidencialidad y restricción del

compromiso con la verdad de un contenido enunciado", en GONZÁLEZ RUIZ, R., IZQUIERDO ALEGRÍA, D. y LOUREDA LAMAS, O. (eds.): La evidencialidad en español. Teoría y descripción, Madrid: Iberoamericana/Vervuert, p. 129–152.

Hassler, G. (2010): "Epistemic modality and evidentiality and their determination on a deictic basis: the case of Romance languages", en DIEWALD, G. y SMIRNOVA, E. (Eds). Evidentiality in European Languages: the lexical--- grammatical distinction, p. 223–248.

Izquierdo Alegría, D. (2016): Alcance y límites de la evidencialidad. Aspectos teóricos y propuesta de análisis aplicada a un conjunto de adverbios evidencialoides del español. Tesis doctoral. Pamplona: Universidad de Navarra.

Jansegers, M. (2017): Hacia un enfoque múltiple de la polisemia. Un estudio empírico del verbo multimodal 'sentir' desde una perspectiva sincrónica y diacrónica, Berlin: de Gruyter.

Kotwica, Dorota (2017) "From Seeing to Reporting: Grammaticalization of Evidentiality in Spanish Constructions with ver ('to see')". En Marín--- Arrese, Juana I. et al. (eds.) Evidentiality and Modality in European Languages. Discourse--- pragmatic perspectives, Peter Lang, Bern pp. 87--- 109.

Marcos Sánchez, M. (2016): "Subjetividad e intersubjetividad en los marcadores evidenciales formados por el verbo de percepción visual ver en español", en GONZÁLEZ RUIZ, R., IZQUIERDO ALEGRÍA, D. y LOUREDA LAMAS, O. (eds.): La evidencialidad en español. Teoría y descripción, Madrid: Iberoamericana/Vervuert, p. 153–178.

Traugott, E. C. y G. Trousdale (2013): Constructionalization and Constructional Changes, Oxford University Press.

PRESEEA (2014--- 2017): Corpus del Proyecto para el estudio sociolingüístico del español de España y de América, Alcalá de Henares: Universidad de Alcalá. Accesible en línea: http://preseea.linguas.net. Consultado: marzo 2017.

Samper, J. A. (1998): Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico, Las Palmas: Universidad de Las Palmas de Gran Canaria--- ALFAL.

VV.AA. (en línea): Corpus Ameresco. Proyecto Es.VaG.Atenuacion. Accesible en línea: http://esvaratenuacion.es.

"Like she'd gone totally into like Sunderland mode like": a corpus-based analysis of LIKE in Northeast and Southeast English

James Algie

Theoretical and Applied Linguistics, University of Cambridge, United Kingdom

***Keywords**: sociolinguistic variation; discourse markers; corpus linguistics; language change*

Abstract

The use of 'like' as a discourse marker (hereon referred to as LIKE) has attracted an extraordinary amount of attention as part of the increasingly popular study of discourse markers in recent decades (Tagliamonte 2005). There remains academic disagreement regarding its meanings and functions, its sociolinguistic distribution and its historic development. This ubiquitous feature of spoken English is pervasive in varieties in at least four continents (Kortmann & Lunkenheimer 2013) and the extent of its multi-functionality is "perhaps unprecedented in any modern language" (D'Arcy 2017: 2).

The present study is unique in developing a comparative analysis of LIKE across two regional varieties of British English. The study aims to discover if, how and why Northeast English (NE-Eng) and Southeast English (SE-Eng) differ in terms of LIKE usage, attempting to offer qualitative sociolinguistic, pragmatic and historical explanations for the quantitative patterns identified, and to determine what these patterns can tell us about LIKE more generally.

The study makes use of the Spoken British National Corpus 2014 (Love, *et al.* 2017) to perform a quantitative analysis of the two varieties, measuring relative frequencies of LIKE according to regional dialect, gender, age and clausal position.

Clause-final LIKE is found to be much more common in NE-Eng than in SE-Eng, though still relatively rare in both varieties. Further, it is found to be capable of performing similar functions as LIKE in other positions, namely discourse organisation, hedging and focus. It also has unique functions that cannot be accomplished by other LIKE types, serving to counter objections and inferences and to request or provide clarification. One would expect that SE-Eng speakers use one or more other discourse elements to perform such functions, but to establish which elements would require further quantitative study.

These findings add to evidence that clause-final LIKE is in decline (Bartlett 2013), in contrast to LIKE in other positions, though this decline is at a considerably more advanced stage in SE-Eng. Theories of geographical diffusion (Trudgill 1984), as well as evidence of overlapping function, support the notion that clause-final LIKE is related to other LIKE types, which in fact developed from it as part of a process of grammaticalization. This supports D'Arcy (2017) in contradicting previous claims that clause-initial and clause-medial LIKE originated in counterculture movements of New York City in the 1960s. The evidence presented here adds weight to the notion that these movements did not represent the birth of a new type of LIKE, but rather drew on features that had been occurring at low rates at least as far back as the early nineteenth century.

The data reveals that LIKE usage is clearly biased towards younger speakers, regardless of clausal positioning, an effect which has not changed in recent decades. It is therefore argued that LIKE is an age-graded feature that is primarily undergoing individual change, while generational change is secondary. This, and the finding that LIKE may be

marginally preferred by female speakers, is explained by LIKE's functioning as a sociolinguistic identifier. Young (and possibly female) speakers use LIKE as a marker of social membership, a characteristic that is reinforced by sometimes skewed social perceptions of the feature.

References

Bartlett, J. (2013). "Oh I just talk normal like": A corpus-based, longitudinal study of constituent-final like in Tyneside English. *Newcastle Working Papers in Linguistics*, 19(1), 1-21.

D'Arcy, A. (2017). *Discourse-pragmatic variation in context: Eight hundred years of LIKE*. Amsterdam, Netherlands: John Benjamins.

Kortmann, B. and Lunkenheimer, K. (eds.) (2013). *The Electronic World Atlas of Varieties of English.* Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology.

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319-344.

Tagliamonte, S. (2005). So who? Like how? Just what? Discourse markers in the conversation of young Canadians. *Journal of Pragmatics*, 37, 1896- 1915.

Trudgill, P. (1984). *On dialect: Social and geographical perspectives*. Oxford, United Kingdom: Blackwell

Economy on Twitter. Linguistic variation in positive and negative tweets

Ángela Almela[a]; Gema Alcaraz-Mármol[b]; Jésica López-Hernández[c]

[a]Department of English, Universidad de Murcia, Spain; [b]Department of Modern Languages, Universidad de Castilla-La Mancha, Spain; [c]Department of Computer Science and Systems Engineering, Universidad de Murcia

***Keywords**: Syntax; economy; financial corpus; social networks; sentiment polarity.*

Abstract

Social networks have become an important source of linguistic data which can be analyzed from different perspectives. In the last decade, many studies have focused on the type of language found in these social networks (Grieve et al., 2010; Summer et al., 2012; Markovikj et al., 2013; Ortigosa, Martin, & Carro, 2014; Saif et al., 2016, to name but a few). Twitter is one of the best known networks with over 200 million users, who can send and read short messages called tweets. Through the analysis of what is published on Twitter, researchers can observe emerging trends in speech, in popular

fields like politics (e.g., Mancera & Pano, 2013), tourism (e.g., Marrese-Taylor et al., 2014), and branding (e.g., Ghiassi et al., 2013). Accordingly, the study of linguistic variation in social media is a fairly recent research field full of possibilities (Bamman, Eisenstein, & Schnoebelen, 2014). One of the areas which has achieved an explosive growth on social networks is financial information.

In this framework, the present study performs the analysis of syntactic constructions in an ad-hoc corpus comprising 4,400 Spanish economic tweets, originally compiled for the study presented in García-Díaz et al. (2018), albeit not fully exploited yet. The supervised machine learning method developed by the authors in that study attempted to detect the polarity of financial tweets by means of the UMTextStats tool, mainly exploring lexico-morphological and semantic features in 1,000 positive and 1,000 negative tweets. Three were the main sources for the corpus data: (1) official government accounts like @_minecogob; (2) reputed economists; and (3) official Twitter accounts from news websites with financial information such as @ElPais_economia. It is worth noting that the tweets were examined and manually classified as positive or negative by students from Economics and Business fields so as to ensure reliability.

Unlike the previous study, the methodology employed did not only involve semantics but also morphosyntax. Specifically, the underlying linguistic principle is that speakers unconsciously use and repeat morphosyntactic features in measurable different ways. Therefore, the frequency of these features in language output exhibits a range of variability, which is speaker-dependent, which has been used in the field of computational authorship attribution to recognize intra-writer versus inter-writer variation to differentiate authors from each other. Specifically, the absence, presence or quantitative deviation of these structures have proven indicative of authorship (Chaski, 2005; 2012; 2013). Accordingly, the main aim of the present study was the exploration of the potential of psycholinguistic categories and morphosyntactic patterns for the polarity of financial tweets. The preliminary findings show that there are essential differences in the two opposite poles, namely positive and negative tweets. From the promising results yielded, we discuss the clustering of the texts across the corpus, specific language patterns, and some implications for the automatic processing of tweets for research.

References

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender and variation in social media. *Journal of Sociolinguistics,* 18, 135-160. doi: 10.1111/josl.12080

Chaski, C. E. (2005). Who's at the Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence,* 4(1).

Chaski, C. E. (2012). Author Identification in the Forensic Setting. In L. Solan & P. Tiersma (Eds.), *The Oxford Handbook of Forensic Linguistics* (pp. 333-372). Oxford, UK: Oxford University Press.

Chaski, C. E. (2013). Best Practices and Admissibility in Forensic Author Identification. *Journal of Law and Policy*, 21(2), 332-376.

García-Díaz, J. A., Salas-Zárate, M. P., Hernández-Alcaraz, M. L., Valencia-García, R., & Gómez-Berbis, J. (2018). Machine Learning Based Sentiment Analysis on Spanish Financial Tweets. In A. Rocha, H. Adeli, L. P. Reis, & S. Costanzo

(Eds.), *Trends and Advances in Information Systems and Technologies. Advances in Intelligent Systems and Computing 746* (pp. 305-311). New York, NY: Springer.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40, 6266-6282.

doi:10.1016/j.eswa.2013.05.057

Grieve, J., Biber, D., Friginal, E., & Nekrasova, T. (2010). Variation among Blogs: A Multi-dimensional Analysis. In A. Mehler, S. Sharoff, G. Rehm, & M. Santini (Eds.), *Genres on the Web: Corpus Studies and Computational Models*, pp. 89-123. New York, NY: Springer. doi: 10.1007/978-90-481-9178-9_14

Mancera, A. & Pano, A. (2013). Nuevas dinámicas discursivas en la comunicación política en Twitter. *Círculo de Lingüística Aplicada a la Comunicación*, 56, 53-80.

Markovikj, D., Gievska, S., Konsinski, M., & Stillwell, D. (2013). Mining Facebook Data for Predictive Personality Modeling. *Computational Personality Recognition: AAAI Technical Report WS-13-01* (pp. 23-26).

Marrese-Taylor, E., Velásquez, J. D., & Bravo-Márquez, F. (2014). A Novel Deterministic Approach for Aspect-Based Opinion Mining in Tourism Products Reviews. *Experts Systems with Applications*, 52, 7892-7902.

doi:10.1016/j.eswa.2014.05.045

Ortigosa, A., Martín, J., & Carro, R. (2013). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behaviour*, 30, 1-15. doi:10.1016/j.chb.2013.05.024

Pollard, C. & Sag, I. (1994). *Head-driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

Summer, C., Byers, A., Boochever, R. & Park, G. (2012). Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. *11th International Conference on Machine Learning* (pp. 386-393). doi: 10.1109/ICMLA.2012.218

Saif, H., He, Y., Fernández, M., & Alani, H. (2016). Contextual Semantics for Sentiment Analysis of Twitter. *Information Processing and Management*, 52, 5- 19. doi:10.1016/j.ipm.2015.01.005

A Corpus-aided Critical Discourse Analysis of the Saudi Women's Social Media Campaign

Nouf Alotaibi

School of English, University of Sheffield, United Kingdom

***Keywords**: Critical Discourse Analysis; Corpus Linguistics; Cultural injustice; Twitter data; Saudi women*

Abstract

The role of social media has been crucial in encouraging and organising the revolutions across North Africa and the Middle East (i.e. Arab Spring in 2011) to change unjust and inequitable socio-political conditions. Twitter, particularly, has been perceived as an influential social platform in Saudi Arabia, where the law prohibits the formation of political parties, civil society organisations, or marches. Saudi women's rights activists discovered the value of social media sites during the Arab Spring and started to demand sociopolitical change in their lives. The #EndMaleGuardianshipSystem campaign is one campaign that has provoked heated debate on social media and in Saudi society. This critical discourse analysis research aims to investigate how the Twitter genre intersects with the socio-political changes taking place in the country. To date, scholarly work done on Saudi women has not drawn connections between patterns of gender inequality in Saudi society and sociological questions of female misrecognition. In her theory of social justice, the feminist writer Nancy Fraser (1995) distinguished socioeconomic injustice from cultural injustice, wherein the former is rooted in the political-economic construction of a society, such as exploitation, economic marginalisation, and deprivation, and the latter - including cultural domination, misrecognition, and disrespect

- is rooted in the social paradigms of representation, interpretation, and communication. In practice, however, both are intertwined and lead to overlapping demands. In this campaign against the male guardianship system, demands for cultural change intermingle with economic change because Saudi women are not recognised, and they are economically marginalised because of the male-guardianship system. The cultural injustices generate cultural devaluation and denial of legal rights, which in turn cause socioeconomic injustices, like the denial of rights to private economic means without a male-guardian's permission. Thus, this research combines a corpus-based approach with textual analytical methods, Halliday (2014) and Van Leeuwen (2008), in order to investigate a feminist Saudi campaign on Twitter calling for an end to the male-guardianship system, the findings of which are then compared with the (pro-guardianship) counter-campaign. Two corpora of 3000 tweets were compiled, then divided into female campaigners and female opponents, in order to identify the representation of Saudi women and their actions. The findings showed that female campaigners represented Saudi women as 'victims' under this system and treated as minors and second-class citizens. Secondly, whereas Saudi women are textually passivated by the female campaigners, they are assigned to active roles by the female opponents. Female opponents of the campaign who find the guardianship system supportive of Saudi women, tend to use surprisingly sexist language while representing themselves and other Saudi women.

References

Fraser, N. (1995). From redistribution to recognition? Dilemmas of justice in a 'post- socialist' age. *New Left Review, 212*, 68-93.

Halliday, M. A. K. (2014). *Halliday's introduction to functional grammar* (4th ed.). London: Routledge.

Van Leeuwen, T. (2008). *Discourse and practice: New tools for critical discourse analysis*. Oxford: Oxford University Press.

Corpora, historical sociolinguistics and variational 'lies'

Carolina P. Amador-Moreno

Department of English, University of Extremadura, Spain

***Keywords**: historical sociolinguistics; variational corpus linguistics; Irish English; embedded inversion*

Abstract

Embedded inversion is often included in the catalogue of features resulting from the contact between Irish (Gaelic) and English in Ireland. While Hickey (2012) does not include it among the areal features of Irish English when discussing areality in the British isles, this feature has been viewed as a legacy of Celtic influence that has perhaps survived due to low salience.

This paper investigates the occurrence of embedded inversion in Irish English across time, by looking at patterns of use in CORIECOR, the *Corpus of Irish English Correspondence*, which contains emigrant letters written to and by Irish emigrants from 1760 to 1940. The letters were sent mainly between Ireland and other countries such as the United States, Canada, Great Britain, New Zealand, Australia, and Argentina, and therefore provide an empirical base for studies of historical change.

The paper will report on the occurrence of embedded inversion in CORIECOR and results will be compared with those found in other historical corpora such as the *Corpus of Historical American English* (COHA) and the *Corpus of Oz Early English* (COOEE), which provide data from two of the destinations of Irish emigrants. Other existing corpora covering roughly the same period from older varieties include the *Corpus of Late Modern English Texts* (CLMET) and the *Corpus of Modern English Prose*, which provide information on its attestation in British English, and the *Corpus of Modern Scottish Writing* (CMSW), for Scottish English. Its robustness in IrE in comparison with the British English and the Scottish English data will be discussed, raising the question of whether corpora can provide answers to variational questions. The analysis appears to raise questions over the general 'northern'/Celtic trait of this feature, but it may provide some answers in relation with how this feature was transported and with the connection between the varieties of English that emerged in the New World in the colonial period (Tagliamonte 2013: 18), and more specifically with the influence of Ulster English in some regions of North America (Montgomery and Robinson 2000: 40).

From a sociolinguistic angle, the study confirms that this feature was already strong enough by the 1830s, when public education was introduced in Ireland, and the type of patterns that emerge from the analysis indicate that, while showing no clear formal preferences in terms of matrix verbs of question types yet, this feature was being used by educated as well as uneducated speakers at the time the language shift from Irish (Gaelic) to English was under way. This was also a period of feature loss in the context of Irish English, which means that embedded inversion was not prescribed against.

References

Hickey, R. (2012). English in Ireland. In Kortmann, B., Traugott, E. C., & Hickey, R.

(Eds.) *Areal features of the anglophone world*. 79-107. Berlin/ Boston: De Gruyter.

Montgomery, M. & Robinson, Ph. (2000) Ulster: a Linguistic Bridge to North America. *The Journal of Scotch-Irish Studies 1*, 40-60.

Tagliamonte, S. (2013) *Roots of English*. Cambridge: CUP.

Trust drivers across scam genres

Patrizia Anesa[a]; Ismael Arinas Pellón[b];

[a]Department of Foreign Languages, Literatures, and Cultures, University of Bergamo, Italy; [b]Departamento de Lingüística Aplicada a la Ciencia y a la Tecnología, Universidad Politécnica de Madrid, Spain.

***Keywords**: Scams, cognitive judgment errors, persuasion strategies, fraud detection*

Abstract

Scams are unethical communication practices whose presence online represents a dramatic risk for the victims both from a financial and emotional perspective. This study aims to observe what persuasion strategies are common to different types of scams, by focusing on a multi-genre corpus compiled to this end. The corpus includes advance-fee scams and romance scams. They were collected between 2004 and 2018, and the language used is English.

In particular, this analysis considers what type of discursive devices are employed to attract the victims' interest and subsequent trust, to the extent that they become unable to recognize the linguistic signs which represent clear manifestations of a fraud. Indeed, persuasion is a core aspect of human interaction. While it can be used positively to promote improved behavior, it can also be a manipulative tool, as happens in the case of fraudsters (Ferreira 2016: 29).

The different types of scams draw considerably on errors of judgment. Thus, Lea et al.'s taxonomy including both motivational and cognitive judgment errors (2009: 25-34) is used in this study to identify the persuasive processes adopted by the scammers. The categories analyzed include, inter alia, visceral influences, reduced motivation for information processing, likening and similarity, reciprocations, as well as commitment and consistency.

The analysis shows that fraudulent emails and messages can be successful despite their manifestly suspicious appearance, because of their ability to implement persuasion strategies and to fulfill certain credibility requirements. Indeed, regardless of the type of scam, they display an effective narrative structure within a recognizable genre and, at the same time, they create an illusion of intimacy, sincerity and urgency. The artful combination of these aspects leads the victims (who display a very heterogenous background) to believe in the authenticity of the messages received. Subsequently, once a relation with the victim has been established, scammers can continue to exploit the victim's errors of judgment.

This study draws on previous research on scam messages (see Arinas Pellón et al. 2005) and aims to gain a finer understanding of how discursive strategies are employed for fraudulent purposes across genres. Moreover, the results obtained can be used for the definition of linguistic models which can be implemented in the detection of adversarial phishing/fraudulent strategies. In particular, future work on this topic is also aimed to leverage data from this paradigm to develop two applications which are aimed at detecting the potential danger of business and romance scams and, thus, can contribute to the development of systems capable of neutralizing these messages. The data obtained will also be used for the creation of guidelines to be circulated especially on sensitive sites (e.g. dating sites) to help end-users to be more knowledgeable about phishing attempts.

References

Arinas Pellón, I./Gozalo Sáinz, M. J./González González, T. (2005). Nigerian Letters, Dutch Lottery and Teaching an ESP Genre. In L. Sierra/E. Hernández (Ed.). Lenguas para fines específicos (VIII) Investigación y enseñanza. Alcalá de Henares: Universidad de Alcalá de Henares (pp. 89-96).

Lea, S./Fischer, P./Evans, K. (2009). The Psychology of Scams: Provoking and Committing Errors of Judgment. Exeter: Office of Fair Trading.

Ferreira, A. (2016). Persuasion in Scams. In Jakobsson, M. (Ed.) Understanding Social Engineering Based Scams. New York: Springer-Verlag (pp. 29-47).

A Comparative Corpus-Based Study of Discourse Markers in Native and Non-native Spoken Discourse

Asuman Aşık [a]

[a]*Department of English Language Teaching, Gazi University, Turkey*

***Keywords**: Discourse markers, Non-native spoken corpus, Pragmatic competence*

Abstract

Discourse markers (DMs) are among the essential frequent units of spoken discourse used by the speakers to enrich their speech for several purposes. DMs are acknowledged as a multifunctional and pragmatic class of words which provide smooth and efficient speech production and interaction (Aijmer, 2002; Schiffrin, 1987). Due to their unique properties in spoken discourse, previous studies found a strong relationship between DMs and communicative and pragmatic competence and their pedagogical significance in language teaching through comparative corpus-based studies (Fung and Carter, 2007; Müller, 2005). In this respect, the study attempts to contribute to the field by identifying the DMs of Turkish non- native speakers in their spoken discourse so as to provide implications for language acquisition regarding discourse-pragmatic competence in English. Thus, the research objectives are (i) to investigate the frequency level of the DMs used by Turkish non-native speakers of English, and

(ii) to compare the non-native use of DMs with the native use in spoken discourse. To this end, a research corpus was compiled with 20 senior university student presentations in Turkey. For comparison, a sub-corpus including student presentations within Michigan Corpus of Academic Spoken English (MICASE) corpus was used. The research is a corpus-based study based on quantitative and qualitative analysis through descriptive statistics and AntConc to display the occurrences and distribution of DMs in the discourse through lexical size and frequency counts. The results indicate that the frequency of DMs in the non-native spoken corpus, 11.15 %, is lower than the one in MICASE with 21.46 % in the total corpus size. Furthermore, the types of DMs used by non-native and native speakers show similarities and some significant differences. The most frequent DMs used by native speakers of English such as and, like, so, but, just, then, okay, you know, really, yeah, I mean are also used by Turkish non-native speakers but less frequently. However, particular DMs such as kind of/kinda, right, i think, basically, well and cuz in native spoken discourse are not used at all by non-native speakers. Furthermore, the results show overuse of uhh and yes and underuse of well by non-native speakers, which can be interpreted as the non-native speakers are not aware of the multi-functional features of DMs such as hesitation, denoting thinking process, searching for the right word and fillers. The study provides some implications to increase discourse-pragmatic competence in nonnative spoken discourse. Since the variety and the range of DMs used by Turkish non-native speakers are limited in and confined to particular items, their awareness should be raised towards the variety and functions of DMs through both explicit and implicit teaching, language observation, problem-solving, and cross-language comparisons (Fung and Carter, 2007) and language samples from daily conversations of native speakers (Hellermann and Vergun, 2007).

References

Aijmer, K. (2002). English Discourse Particles. Amsterdam: John Benjamins.

Fung, L. & Carter, R. (2007). Discourse markers and spoken English: native and learner use in pedagogic settings. Applied Linguistics, 28(3), 410-439Hellermann, J. & Vergun, A. (2007). Language which is not taught: The discourse marker use of beginning adult learners of English. *Journal of Pragmatics, 39*(1), 157-179.

Müller, S. (2005). Discourse markers in native and non-native English Discourse.

Amsterdam/Philadelphia: John Benjamins.

Schiffrin, D. (1987). *Discourse Markers*. Cambridge: Cambridge University Press

A Comparative Corpus Study of Nonverbal Elements in Native and Non-Native Russian Speech

Ekaterina Baeva [a]

[a] *English Language Department for Philology and Art, Saint Petersburg State University, Russia*

***Keywords**: colloquial Russian, corpus data, speech corpus, nonverbal elements, hesitation phenomena, language acquisition*

Abstract

Each utterance consists of elements of three types. Verbal elements are the essence of the semantic dimension of a text; they are known to bear the principal meaning of an utterance. Roughly verbal elements help structure the message without actually being linked to its meaning; they are auxiliary parts of speech, parenthetical words, and discourse units (Kiseleva, Paillard, 2003; Dobrovol'skij, Pöppel, 2015). Nonverbal elements can be found almost in any type of spoken speech of any given speaker, and it has been acknowledged that they often help determine one's fluency in a foreign language (De Jong, Bosker, 2013; Götz, 2013).

Nonverbal elements include hesitation pauses, voice qualifiers (whispering, breathing, hoarse voice etc), affect bursts (laughing, sobbing, coughing etc), tongue clicks, lip-smacks, noisy air intakes, and others (Bogdanova-Beglarian, Filyasova, 2018). Nonverbal elements are not directly related to the utterance content; nevertheless, they are, indeed, rather frequent in everyday speech and thus essential for its understanding and hands-on decoding, not only for scholars, but, most importantly, for instructors and learners.

The study highlights the benefits of corpus study nonverbal elements in oral discourse in Russian as L1 and L2. The aim of the research was to track various nonverbal elements in spontaneous Russian of native and non-native speakers, to define their functions and to compare them in L1 and L2.

The data was obtained from the two modules of the Corpus of the Russian language: the corpus of Russian everyday speech "One Day of Speech" (the ORD corpus) (Asinosvky et al., 2009) and "Balanced Annotated Text Collection" (SAT) (Bogdanova-Beglarian et al., 2017). The ORD corpus (1250 hours of recordings, 128 respondents) contains spontaneous speech of native speakers, residents of St. Petersburg, for the major part everyday dialogues and polylogues, which were collected using the method of continuous daily speech monitoring. SAT comprises experimental speech of both L1 and L2 speakers. The recordings are categorized into typical communicative scenarios of natural communication: reading, rendering, description of the image, story-telling. In addition to the speech of different professional groups of Russian speakers, SAT also includes several blocks of L2 Russian speech American, French, Chinese, and Dutch speakers. This focus group amounts to 153 speakers and comprises 772 monologue texts, with total duration of 30 hours.

In our research, we have found that all the types of nonverbal elements are inherent to spontaneous Russian speech, regardless of native or non-native speakers or their level of proficiency. Furthermore, in SAT, while comparing recording of the same

communication scenarios of different respondents, we could detect nonverbal elements of the same kind, roughly in the same positions. What's more, on the pragmatical level, these elements perform several functions, and at times simultaneously. Their most frequent function was found to be hesitation. However, nonverbal elements in both L1 and L2 often combine hesitation with a search function (when a speaker searches his mind for a word, an expression or an idea to continue or complete an utterance). They can also be a reflexive marker or a discursive marker of the speech start or finale.

References

Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., Sherstinova, T. (2009). The ORD speech corpus of Russian everyday communication «One Speaker's Day»: creation principles and annotation. In V. Matoušek, P. Mautner (Eds.), *TSD 2009*, *LNAI*, 57292009 (pp. 250-257). Berlin, Heidelberg.

Bogdanova-Beglarian, N.V., Filyasova, Yu.A. (2018). Discourse vs. pragmatic markers: a contrastive terminological study. *SGEM2018 Vienna ART Conference Proceedings*, 5, 123-130.

Bogdanova-Beglarian, N.V., Sherstinova T.Yu., Zaides, K.D. (2017). Korpus "Sbalansirovannaya Annotirovannaya Tekstoteka": metodika mnogourovnevogo analiza russkoj monologicheskoj rechi [Corpus "Balanced Annotated Text Collection": the methodology of multi-level analysis of Russian monological speech]. In D. A. Kocharov, P. A. Skrelin (Eds.), *Analiz russkoj razgovornoj rechi (ARZ-2017): Trudy sed'mogo mezhdisciplinarnogo seminara* [*Analysis of Russian spoken speech (ARZ-2017): VII Interdisciplinary Seminar proceedings*] (pp. 8-13). Saint Petersburg, Russia. (in Russ.).

De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In: *The 6th Workshop on Disfluency in Spontaneous Speech (DiSS)*, 17–20.

Dobrovol'skij, D., & Pöppel, L. (2015). Corpus perspectives on Russian discursive units: semantics, pragmatics, and contrastive analysis. In *Yearbook of Corpus Linguistics and Pragmatics* 2015, 223–241. Springer, Cham.

Götz, S. (2013). Fluency in native and nonnative English speech. *Studies on Corpus Linguistics*, Vol. 53. John Benjamins Publishing.

Kiseleva, K. L., Paillard, D. (eds.) (2003). Diskursivnye slova russkogo jazyka. Kontekstnoe var'irovanie i semantičeskoe edinstvo [Discourse words of Russian: contextual variation and semantic units]. Moskva: Azbukovnik. (in Russ.).

El uso didáctico de un corpus lingüístico en el aula de ELE: la Movida madrileña

Azucena Barahona Mora

Departamento de Estudios Ingleses, Universidad Complutense de Madrid, España
Palabras clave: *Lingüística del corpus; materiales auténticos; enseñanza-aprendizaje de ELE*

Abstract

La finalidad de esta propuesta es la presentación de un corpus elaborado a partir de canciones y películas destacadas del periodo de la Movida madrileña y que ha sido empleado en el aula de ELE con el objetivo de enseñar aspectos lingüísticos y culturales. Este corpus pertenece al tipo especializado y sincrónico, el cual, según Torruella y Llisterri (1999), Pearson (1998), Corpas Pastor (2002), Atkins, Clear y Ostler (1992), está orientado hacia una variedad lingüística y propósito en concreto, y a un tema, campo o tiempo determinados.

La elección del fenómeno de la Movida madrileña se debe, por un lado, a su notable contribución artística, política y social, lo que posibilita trabajar aspectos culturales y socioculturales, y por otro, a las muestras de lengua real que aparecen en los textos escogidos, que contienen un lenguaje diferenciado en el que sobresale el *cheli.* Este fue tan relevante que incluso se publicaron obras como el *Diccionario cheli,* Umbral (1983) y *El Tocho Cheli,* Márquez (1993). Autores como Rodríguez González (2002), Morant (2001) y Herrero Moreno (1989), entre otros, ya estudiaron la importancia del lenguaje juvenil y de contracultura.

En ocasiones, los profesores de lenguas extranjeras se encuentran con el problema de que los manuales graduados no cubren las necesidades de sus alumnos, por lo que deben buscar otros recursos. Para solventar estas carencias, los materiales auténticos se convierten en elementos indispensables. Según García Arreza, Segura Báez y Zamora López (1994), estos son aquellos que incluyen contenidos lingüísticos creados por y para hablantes nativos y no se han modificado. Un ejemplo de estos son las películas y canciones. Siguiendo las orientaciones didácticas y los criterios de selección para esta clase de materiales planteados por Álvarez Montalbán (2007), Cancelas y Ouviña (1998), García Mata (2003), y Pozzobon y Pérez (2010), entre otros autores, se puede afirmar que ofrecen datos genuinos y diversos y encierran un gran número de registros lingüísticos y variedades dialectales.

Por su parte, los corpus reflejan la lengua en contexto y evidencian el uso que hacen los hablantes de ella. Permiten acercar al alumnado información muy variada, no solo lingüística sino cultural, y el entorno real donde se habla. Resultan interesantes las aportaciones de Sinclair (1996), Aijmer (2009) y Cruz Piñol, (2012) sobre la aplicación de los corpus en la enseñanza.

En cuanto a la metodología de análisis del corpus, Wallis y Nelson (2001) proponen distintos sistemas; uno de ellos es la anotación. En este estudio se empleó este proceso y se manejó como herramienta una ficha de registro de observación. La recopilación de muestras fue explotada para la creación de una serie de unidades didácticas que fueron implementadas en aula de ELE. Tras su aplicación, los resultados confirmaron que la utilización de estos materiales es efectiva como recurso didáctico y desarrolló en los discentes la competencia comunicativa, y se corroboró que las películas

y canciones son fuentes válidas para la confección de corpus lingüísticos.

Referencias

Aijmer, K. (2009). *Corpora and language teaching*. Amsterdam, Philadelphia: John Benjamins.

Álvarez Montalbán, F. (2007). El uso de material auténtico en la enseñanza de ELE. En J. M. Izquierdo, F. Martos Eliche, A. Yagüe y otros (Eds.), *FIAPE: II Congreso Internacional: una lengua, muchas culturas* (pp. 26-29). Granada: FIAPE.

Atkins, S., Clear, J., y Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, *7*(1), 1-16.

Cancelas y Ouviña, L. P. (1998). Realia o Material Auténtico ¿Términos diferentes para un mismo concepto? En M. Cebrián de la Serna (Coord.), *Creación de materiales para la innovación educativa con nuevas tecnologías* (vol. 1, pp. 391-394). Málaga: Universidad de Málaga-Instituto de Ciencias de la Educación.

Corpas Pastor, G. (2002). Traducir con corpus: de la teoría a la práctica. En *Texto, terminología y traducción* (pp. 189-228).

Cruz Piñol, M. (2012). *Lingüística de corpus y enseñanza del español como 2/L*. Madrid: Arco Libros.

García Arreza, M., Segura Báez, J. J., y Zamora López, M. D. (1994). *La lengua inglesa en educación primaria.* Málaga: Aljibe.

García Mata, J. (2003). La autenticidad de los materiales de enseñanza-aprendizaje y el uso de los medios de comunicación audiovisuales y escritos en la clase de "ELE": una topología. En H. Perdiguero y A. Álvarez, A (coord.), *Medios de comunicación y enseñanza del español como lengua extranjera. (Actas del XIV Congreso Internacional de ASELE, Burgos, 2003)* (pp. 882-894). Burgos: Universidad de Burgos.

Herrero Moreno, G. (1989). El coloquio juvenil en los cómics marginales. En *Comunicación y lenguaje juvenil* (pp. 179-201). Instituto Alicantino de Cultura Juan Gil-Albert.

Márquez Martínez, J.R. (1993). *El tocho cheli. Diccionario de jergas, germanías y jerigonzas.* Madrid: El Papagayo.

Morant, R. (2001). Una asignatura pendiente: el lenguaje estudiantil. En *Aspectos didácticos de lengua y literatura, 11*, 55-73. Instituto de Ciencias de la Educación.

Pearson, J. (1998). Terms in Context, volume 1 of Studies in Corpus Linguistics. *John Benjamins, Amsterdam, The Netherlands*, *14*, 88.

Pozzobon, C. y Pérez, T. (2010). La autenticidad de los textos en la clase de lengua. *Acción Pedagógica*, *19*(1), 80-88.

Rodríguez González, F. (2002). *El lenguaje de los jóvenes*. Barcelona: Editorial Ariel, S.A.

Sinclair, J. (1996). Preliminary recommendations on corpus typology. EAGLES Document TCWG-CTYP/P.Recuperado de http://www. ilc. pi. cnr. it/EAGLES/corpustyp/corpustyp. html

Torruella, J., y Llisterri, J. (1999). Diseño de corpus textuales y orales. *Filología e informática. Nuevas tecnologías en los estudios filológicos*, 45-77.

Umbral, F. (1983). *Diccionario cheli*. Barcelona: Grijalbo.

Wallis, S., y Nelson, G. (2001). Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery*, *5*(4), 305-335.

The reframing of rape in extremist online rhetoric: Using corpus linguistics to compare identity construction in narrative and non- narrative discourses.

Kate Barber

Centre for Language and Communication Research, Cardiff University, Wales

***Keywords**: online extremism, sociocognitive discourse studies, narrative deconstruction, XML tagging, identity construction*

Abstract

The use of the internet by extreme right-wing groups is nothing new (Weimann & Van Knop 2008, Hale 2012) although its use by extremists has traditionally been in conjunction with activity organised on the ground (for example marches, meetings or organised events (Mudde 2000)). While some far-right groups still operate in this way (ibid.), the emergence of the Alternative Right (Alt-Right), which exists predominantly online, has redefined what it means to be a far-right extremist. The affordances of the online environment have similarly united a range of men's rights groups: from those campaigning against the repression of traditional gender roles and identities (in much the same way as feminist groups have), to extreme misogynistic groups which blame feminism for the perceived emasculation and humiliation of men (Kimmel 2015). Collectively, these men's rights groups are known as the *Manosphere*. Through a series of online campaigns and their shared misogynistic ideologies, the Alt-Right and the Manosphere are growing closer. It has been argued that anti-feminist rhetoric, particularly regarding the reframing of rape and sexual assault, is the 'gateway drug' (Romano 2016) that is now leading men from the Manosphere to the more extreme far right.

Research into the communication of extremist groups has found that narratives are one of the most prevalent types of discourse and are readily used to 'promote the adoption of beliefs, attitudes, and intentions that are consistent with the group's ideology' (Braddock 2015, p. 38). Intricately bound to the analysis of narrative is that of identity: how participants' identities are constructed and deconstructed within the developing structure of the narrative (Georgakopoulou 2006, p. 101). Key to this present study is whether identity construction varies in relation to the use of narratives; how the authors

construct and deconstruct their and other identities; and how this construction relates to how rape and sexual assault is reframed.

My paper outlines preliminary findings from the second year of my PhD research. Using corpus linguistics and an analytical framework based largely on van Dijk's (1984) sociocognitive approach to discourse studies, the paper outlines how identity construction promotes white male victimhood and portrays the mainstream concept of rape culture as a 'feminist-produced moral panic' (Gotell & Dutton 2016, p. 65). The presentation includes details of the network analysis I undertook in order to locate the online websites and blogs from which I selected my data; how I created my corpora using XML tagging;

and a comparative analysis of racist and misogynistic identity construction in narrative and non-narrative discourses. Finally, some of the ongoing challenges this research has presented will be discussed while I advocate the importance of applying linguistic analyses to this type of online extremism.

Bibliography

Braddock, K. (2015). The Utility of Narratives for Promoting Radicalization: The Case of the Animal Liberation Front. *Dynamics of Asymmetric Conflict*. 8(1). pp. 38- 59.

Georgakopoulou, A. (2006). Small and Large Identities in Narrative (Inter)action. In: De Fina, A., Schiffrin, D. & Bamberg, M. (Eds). *Discourse and Identity*. Cambridge: CUP. pp. 83-102

Gotell, L. & Dutton, E. (2016). Sexual Violence in the 'Manosphere': Antifeminist Men's Rights Discourses on Rape. *International Journal for Crime, Justice and Social Democracy*. 5(2). pp. 65-80.

Hale, W.C. (2012). Extremism on the World Wide Web: A Research Review. *Criminal Justice Studies*. 25(4), pp. 343-356.

Kimmel, M. (2015). *Angry White Men: American Masculinity at the End of an Era*. New York: Nation Books

Mudde, C. (2000). *The Ideology of the Extreme Right*. Manchester: Manchester University Press

Romano, A. (2016). '*How the alt-right's sexism lures men into white supremacy*.' Vox, 14th December 2016. Available online at: http://www.vox.com/culture/2016/12/14/13576192/alt-right-sexism-recruitment

van Dijk, T.A. (1984). *Prejudice and Discourse*. Amsterdam: John Benjamins Publishing Company.

Weimann, G. & Von Knop, K. (2008). Applying the Notion of Noise to Countering Online Terrorism. *Studies in Conflict & Terrorism*. 31(10), pp. 883-902.

La anotación de los grados de factualidad en el discurso escrito en español: el papel de las eventualidades

Leyre Barrios [a], Ana Fernández [b], Gloria Vázquez [a]

[a] *Universitat de Lleida;* [b] *Universitat Autònoma de Barcelona* [b]

Palabras clave: *factualidad, certeza, anotación de corpus, eventualidades, modalidad epistémica*

Abstract

El autor de un texto manifiesta su grado de compromiso respecto a cada una de las situaciones que describe: o bien expone dicha situación con un compromiso claro de que es cierta o de que no lo es, o bien lo hace sin un compromiso claro, presentándola como probable o dudosa. Al enunciarla el autor expresa a través de la ausencia o presencia de marcadores de tipo léxico, morfológico, sintáctico y discursivo este grado de compromiso. De esta forma, la certeza de dichas situaciones (eventualidades), llamada también factualidad, siempre está en relación al modo en que estas son presentadas. Además, hay que tener en cuenta que pueden ser narradas por una o más personas y en el mismo momento o en momentos distintos y que los grados de factualidad no tienen por qué coincidir.

En la última década, la anotación de textos con respecto a la certeza o factualidad ha experimentado un reciente interés en la lingüística de corpus y en el PLN. Uno de los trabajos pioneros en este campo es FactBank (Saurí & Pustejovsky, 2009), en el cual se presenta un anotador factual para el inglés que combina la modalidad de los eventos (CT: cierto; PR: probable; PS: posible; U: indefinido) con su polaridad (+: positiva; -: negativa; u: indefinida). Esto da como resultado ocho etiquetas: CT+, CT-, PR+, PR-, PS+, PS- CTu y Uu. Diversos autores se han basado en este modelo para la anotación de diferentes corpus, si bien algunos han modificado ciertos valores. Así, cabe mencionar a Lee et al. (2015), Soni et al. (2014), van Son et al. (2014) para el inglés; Matsuyoshi et al. (2010) y Narita et al. (2013) para el japonés; Minard, A. L., Speranza, M., & Caselli, T. (2016) para el italiano; o Wonsever et al. (2008) para el español, entre otros. Además de estos proyectos, destacan otros como el de Diab et al. (2009), el cual emplea otra serie de etiquetas para anotar la factualidad: CB (*committed belief*) y NCB (*non- committed belief*). En este proyecto se basan otros, como los de Prabhakaran et al. (2015) o Colomer et al. (2016), para el inglés y el chino y el español, respectivamente.

Nuestro trabajo se enmarca dentro del proyecto TAGFACT, cuyo el objetivo es la creación de una herramienta para la anotación automática de la factualidad para el español. En este momento estamos definiendo los valores factuales con los que etiquetar las eventualidades documentadas en el corpus recopilado.

Para nuestra propuesta hemos partido de otros trabajos del ámbito, especialmente, de FactBank (2009), adoptando algunas de sus premisas sobre el concepto de certeza, de Diab et al. (2009), inspirándonos en parte en la terminología que usan, y de Wonsever et al. (2008), clasificando de forma diferenciada las situaciones de futuro, ya que son, *per se*, no factuales. Ahora bien, nuestro modelo de anotación incluye algunas innovaciones importantes. La primera es que dicha propuesta se establece por niveles, de modo que el grado de refinamiento de las etiquetas puede adaptarse a los objetivos que se persigan según la aplicación. En segundo lugar, hemos diferenciado clases de eventualidades según la presencia o la ausencia de dinamismo, ya que solo las situaciones dinámicas

(procesos y eventos) pueden llegar a categorizarse como hechos. Para los estados o las acciones habituales hemos usado otra nomenclatura.

Referencias

Colomer, J. P., Lai, K., y Rambow, O. (2016). Detecting level of belief in Chinese and Spanish. En *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)* (pp. 22-30). Osaka.

Diab, M. T., Levin, L., Mitamura, T., Rambow, O., Prabhakaran, V., y Guo, W. (2009). Committed belief annotation and tagging. En *Proceedings of the Third Linguistic Annotation Workshop* (pp. 68-73). Suntec, Singapur: Association for Computational Linguistics.

Lee, K., Artzi, Y., Choi, Y., y Zettlemoyer, L (2015). Event detection and factuality assessment with non-expert supervision. En *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1643-1648). Lisboa.

Matsuyoshi, S., Eguchi, M., Sao, C., Murakami, K., Inui, K., y Matsumoto, Y. (2010). Annotating event mentions in text with modality, focus and source information. En *Proceedings of the International Conference on Language Resources and Evaluation*, LREC. Malta.

Minard, A. L., Speranza, M., y Caselli, T. (2016). The EVALITA 2016 Event Factuality Annotation Task (FactA). En *CLiC-it/EVALITA.* Nápoles.

Narita, K., Mizuno, J., y Inui, K. (2013). A lexicon-based investigation of research issues in Japanese factuality analysis. En *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 587-595). Nagoya.

Prabhakaran, V., By, T., Hirschberg, J., Rambow, O., Shaikh, S., Strzalkowski, T., ... y Dalton,

A. (2015). A new dataset and evaluation for belief/factuality. A new dataset and evaluation for belief/factuality. En *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics* (pp. 82-91). Denver, Colorado.

Saurí, R., y Pustejovsky, J. (2009). FactBank: a corpus annotated with event factuality. En

Language Resources and Evaluation, *43*(3) (pp. 227-268).

Soni, S., Mitra, T., Gilbert, E., y Eisenstein, J. (2014). Modeling factuality judgments in social media text. En *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*) (Vol. *2*, pp. 415-420). Baltimore, Maryland.

Tonelli, S., Sprugnoli, R., y Speranza, M. (2014). NewsReader Guidelines for Annotation at Document Level. En *Extension of Deliverable D3. Technical Report NWR-2014-2.* Trento.

Van Son, C., van Erp, M., Fokkens, A., y Vossen, P. (2014). Hope and fear: Interpreting perspectives by integrating sentiment and event factuality. En *Proceedings of the 9th Language Resources and Evaluation Conference* (*LREC2014*) (pp. 26-31). Reykjavik.

Wonsever, D., Malcuori, M. y Rosá, A. (2008). SIBILA: Esquema de anotación de eventos. En *Reportes Técnicos RT 08-11*. Universidad de la República, Uruguay.

The Coruña Corpus Tool: The Software Behind the Corpora

Anabella Barsaglini-Castro

English Department, University of A Coruña, Spain

Keywords: *corpus linguistics; software tools; programming; information retrieval; Coruña Corpus*

Abstract

Corpus-assisted studies have contributed to the development of Corpus Linguistics by providing new resources, methodologies and approaches that help researchers in the analysis of discourse and language (Biber *et al.*, 1998; McEnery *et al.*, 2006; Baker & McEnery, 2015). Thus, and as Laurence remarks, "corpora are often referred to as the 'tools' of corpus linguistics." (2013). However, it is important to point out as well that corpora are nothing more than linguistic data retrieved by specialised software tools in order to analyse them.

Since its inception, the Coruña Corpus Tool (CCT) has always been an indispensable companion for the study and use of the different subcorpora within the *Coruña Corpus of English Scientific Writing* (*CC*). The *CC* is a project on which the Research Group for Multidimensional Corpus-based Studies in English (MuStE) has been working since 2003 in the University of A Coruña (Spain). The main idea behind this project has always been the compilation of scientific texts of the late Modern English period into a specialised corpus sharing the same principles (Crespo and Moskowich, 2010; Moskowich, 2016). Conceived of as a 'tool' for the diachronic and synchronic study of language in English scientific writing in general as well as within the different scientific disciplines from several linguistic levels, the *CC* incorporates a tool that has been specially designed to carry out linguistic analyses with the samples contained in each corpus. This fact highlights the relevance of the compatibility that must exist between the corpora and the software to guarantee the best results in the analysis of linguistic data. Loosely speaking, the CCT is, therefore, a corpus management tool that facilitates the gathering of data from an indexed textual repository. It has been developed by the IRLab group in collaboration with the MuStE research group of the University of A Coruña (Parapar & Moskowich, 2007). Its main purpose is to retrieve information from a set of compiled documents that constitutes the *CC* in order to help linguists to extract specific data for their research. To this end, and from a technical point of view, all the texts in the *CC* are compiled, marked-up, coded and saved in XML format by following the Text Encoding Initiative conventions (see TEI Consortium 2018). Although it works like most concordance programs, the CCT offers some special features such as the possibility to search old-fashioned characters and tags in texts or in metadata files, among others.

Developed in 2009 as a beta version, released in 2012 with the publication of *Corpus of English Texts on Astronomy* (*CETA*), and taking advantage of the forthcoming publication of *Corpus of Historical English Texts* (*CHET*), the Coruña Corpus Tool returns with a series of improvements in efficiency and effectiveness that will undoubtedly make the difference in its use. This paper aims at presenting the main features of this new version of the CCT, focusing especially on its recent improvements.

To this end, it will offer an overview of the new characteristics implemented in contrast with previous versions of the tool by considering not only the importance in its design

and programming, but also providing some examples based on user experience. This examination of the CCT is expected to provide a clearer vision of how the tool works and how it can be used to take advantage of its full potential.

References

Baker, P., & McEnery, T. (2015). Introduction. In P. Baker and T. McEnery (Eds.), *Corpora and discourse studies: integrating discourse and corpora*. (pp. 1-19). London: Palgrave Macmillan.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: investigating language structure and use*. Cambridge; New York: Cambridge University Press.

Crespo, B. & Moskowich, I. (2010). CETA in the context of the Coruña Corpus.

Literary and linguistic computing, 25(2), 153-164.

Laurence, A. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30, 2, 141-161.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge Taylor & Francis.

Moskowich, I. (2016). Philosophers and scientists from the Modern Age. In I. Moskowich et al. (Eds.) *'The Conditioned and the Unconditioned': Late Modern English texts on philosophy*. (pp. 1-23). Amsterdam: John Benjamins.

Moskowich, I. & Crespo, B. (eds.) (2012). *Astronomy "playne and simple". The writing of Science between 1700 and 1900*. Amsterdam/Philadelphia: John Benjamins.

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

TEI Consortium (2018). TEI P5: *Guidelines for Electronic Text Encoding and Interchange*. Text Charlottesville: Encoding Initiative Consortium.

‘Nobody is guilty in football. That’s the first thing to understand’: A CADS analysis of the UK Press coverage of the Ched Evans case

Leanne Victoria Bartley

Departamento de Filologías Inglesa y Alemana, Universidad de Granada, Spain

Keywords: *CADS, Appraisal, Media Discourse, Sexual violence, Ched Evans*

Abstract

In 2012, Ched Evans, at one time known purely for his talents on the football pitch, was at the forefront of, potentially, one of the most controversial rape cases in recent decades. Nine years after the start of his football career, Evans was accused, alongside his former teammate, Clayton MacDonald, of raping a 19-year-old woman in a hotel room in Rhyl, North Wales. This case is particularly striking for several reasons. First and foremost, Clayton MacDonald was found innocent at trial in 2012, unlike Ched Evans, who was found guilty of the same crime against the same complainant. That said, Evans eventually served just 2 years of the 5-year sentence he was handed down, before being released in 2014; moreover, his conviction was later quashed at a subsequent retrial in 2016. Evidently, the truth about this case lays with Evans and the alleged victim, and this is not the concern in this paper. Rather, the aim here is to determine how the alleged victim and the accused are discursively represented in the British Press at separate stages of this case: (i) during the run up to the first trial in 2012 (corpus 1), (ii) following Evans’ release in 2014 (corpus 2) and (iii) during Evans’ 2016 retrial (corpus 3). With this in mind, the data for this study is taken from a corpus of newspaper articles that were retrieved from LexisNexis, an online database comprising legal and journalistic documents. Using the search terms *Ched Evans* and *Evans*, a selection of newspaper articles were retrieved across a range of broadsheet and tabloid newspapers and subsequently filtered to ensure that the news stories were primarily concerned with the Ched Evans rape case, as opposed to how he had performed on a particular occasion in a football match, for instance. Given the limited number of news stories available during the run up to Evans’ first trial (94 in total), all newspaper articles available for corpus 1 were retrieved from Lexis Nexis. Meanwhile, data for the other two stages has, so far, been collected from 8 different newspapers (*The Guardian*, *The Independent*, *The Telegraph*, *The Times*, *The Daily Mail*, *The Daily Star*, *The Mirror* and *The Sun*), giving a total corpus size of 524,918 words at present. These figures will inevitably increase as newspaper data concerned with Evans’ release and his subsequent retrial continues to be compiled for future research purposes.

To determine the discursive representation of the main parties involved in this case, an analysis is conducted of the appraisal patterns that emerge in the dataset, with a specific focus on Judgement, as proposed by Martin and White (2005), and Affect as put forward by Bednarek (2008), both pertaining to the Attitude system. Appraisal theory is useful in studies such as this one because, as a framework pertaining to the interpersonal metafunction, it serves to expose the linguistic choices that people make, and in turn reveal one’s stance towards a given proposition or entity (Martin & White, 2005, p. 7). Both a quantitative and qualitative analysis of the dataset is carried out here, through employing the UAM corpus software (O’Donnell, 2016).

References

Bednarek, M. (2008). *Emotion talk across corpora*. Basingstoke & New York, NY: Palgrave MacMillan.

Martin, J.R. & White, P.R.R. (2005). The language of evaluation: Appraisal in English. Basingstoke & New York, NY: Palgrave MacMillan.

O'Donnell, M. (2016). The UAM Corpus Tool (Version 3.3) [Computer Software] Madrid, Spain: Universidad Autonoma de Madrid. Available from http://www.corpustool.com.

'Rising number of homeless is the legacy of Tory failure': Discoursal changes and transitivity patterns in the representation of homeless people in UK press from 2000 to the present day

Leanne Victoria Bartley; Eva Gomez-Jimenez

Filologías Inglesa y Alemana, Universidad de Granada, España

***Keywords**: inequality, wealth, class, homeless, homelessness, newspaper discourse, CDA, corpus linguistics, transitivity, systemic functional grammar*

Abstract

This paper analyses the way in which homeless people have been represented, within the broader context of economic inequality, between 2000 and 2018 in the British press. According to experts, the UK has experienced a process of growing economic inequality since 1971 (cf. Cribb et al., 2012; Hills and Stewart, 2005; Lansley, 2012; Stiglitz, 2012; Marquand, 2013; Westergaard, 2012; Wilkinson and Pickett, 2009). Beyond economic, political, historical and social reasons, media outlets such as newspapers may have played a role in this, in that they do not just report on what happens in society, but also contribute to how this is perceived by readers by displaying and normalizing certain attitudes towards societal issues (Fowler, 1991; van Dijk, 1988; 1998; Fairclough, 1992; 1995). Therefore, the representation of homeless people in the UK press is explored here as one of the many factors that may have rendered the UK a less egalitarian society today.

Critical Discourse Analysis has recently looked at class as a form of social inequality, generally drawing on neoliberalism and the inequities of contemporary capitalism. Approaches regarding UK society have shown, for instance, that there is an overall public legitimization of these wealth inequalities, with generally negative attitudes towards benefit claimants (Baker, forth.; Baker and McEnery, 2015; Patterson et al., 2016; van der Bom et al., 2018) and an outpour of support for economic privileges for the upper-classes (Billing, 1992; Harvey et al., 2015). Related research has indicated a denial of class struggle in Britain today, and that class seems to have disappeared from the political and public agenda in this country (Bennett, 2013; Ortu, 2008; Toolan, 2016). All that said, few studies have focused on homelessness in particular (Huckin, 2002), with some research in the US (cf. Baum and Burnes, 1993) and Brazil (cf. De Melo Resende, 2016) but none, to our knowledge, in the UK. Additionally, there is a shortage of studies that examine the representation of homelessness from a Systemic Functional Linguistics perspective (Garcia da Silva, 2012), which serves to further

justify the rationale behind this research.

Our aim is to look for possible discoursal changes in the representation of homeless people in the UK press between 1971 and the present day. Our dataset here comprises a corpus of broadsheet and tabloid newspaper articles with different ideological orientations (*Daily Mail, the Guardian, the Sun, the Telegraph*) that were published from 2000 to 2018, given that these are the texts currently available online. In order to determine how homelessness and the homeless are represented in the British Press, an SFL transitivity analysis (Halliday and Matthiessen, 2014) is carried out on this dataset. Transitivity, as defined by Halliday (1973: 134), is "the set of options whereby the speaker encodes his [sic] experience of the processes of the external world, and of the internal world of his [sic] own consciousness, together with the participants in these processes and their attendant circumstances". Thus, through a look at the transitivity patterns that emerge in the newspaper articles under analysis, we expect to verify: (i) how homelessness and the homeless are represented in the British Press; (ii) whether homelessness and the homeless are represented in similar or different ways according to newspaper orientation (centre-right vs. centre-left); and (iii) whether homelessness and the homeless are represented in similar or different ways over time in the last two decades.

References

Baker, P. Forthcoming. 'Making the needy look greedy: Using corpus methods to examine The Sun's discourse around benefits', in J. Rahilly and V. Vander (eds.) *Crossing Boundaries: Interdisciplinarity in Language Studies*. Amsterdam: John Benjamins.

Baker, P. and T. McEnery. 2015. 'Who Benefits When Discourse Gets Democratised? Analysing a Twitter Corpus around the British Benefits Street Debate', in P. Baker and T. McEnery (eds.) *Corpora and Discourse Studies: Integrating Discourse and Corpora*: 244-265.

Baum, A. and D. W. Burnes. 1993. *A Nation in Denial: The Truth about Homelessness.* London: Routledge.

Bennett, J. 2013. 'Moralising class: A discourse analysis of the mainstream political response to Occupy and the August 2011 British riots'. *Discourse & Society* 24(1): 27-45.

Billig, M. 1992. *Talking of the Royal Family.* Abingdon: Routledge.

van der Bom, I., L. Paterson, D. Peplow and K. Grainger. ''It's not the fact they claim benefits but their useless, lazy, drug taking lifestyles we despise': Analysing audience responses to Benefits Street using live tweets'. *Context, Discourse & Media* 21: 36–45.

Cribb, J., P. Johnson, R. Joyce and Z. Oldfield. 2012. *Jubilees Compared: Incomes, Spending and Work in the Late 1970s And Early 2010s*. Institute of fiscal studies briefing note 128.

van Dijk, T. A. 1998. *Ideology: A Multidisciplinary Approach.* London: Sage. van Dijk, T. A. 1988. *News as Discourse*. New Jersey: Lawrence Erlbaum.

Fairclough, Norman. 1995. *Critical Discourse Analysis: The Critical Study of Language*. London: Longman.

Fairclough, N. 1992. *Discourse and Social Change.* Cambridge: Polity Press.

Fowler, R. 1991. *Language in the News.* Abington: Routledge.

Garcia da Silva, D. E. 2012. 'Social Representations and Experiential Metafunction: Poverty and Media Discourse'. *Proceedings of ISFC 35: Voices Around the World*. 100-105.

Halliday, M. A. K. 1973. *Explorations in the Functions of Language.* London: Edward Arnold.

Halliday, M. A. K. and C. M. I. M. Mathiessen. 2014. *Halliday's Introduction to Functional Grammar (4th Edition).* London: Routledge.

Hills, J., Stewart, K. (Eds.), 2005. *A More Equal Society? New Labour, Poverty, Inequality and Exclusion*. Bristol: Policy Press.

Huckin, T. 2002. 'Textual silence and the discourse of homelessness'. *Discourse & Society* 13(3): 347-372.

Harvey, L. K. Allen and H. Mendick. 2015. 'Extraordinary acts and ordinary pleasures: Rhetorics of inequality in young people's talk about celebrity'. *Discourse & Society* 26(4): 428-444.

Lansley, S. 2012. *The Cost of Inequality: Why Economic Equality is Essential for Recovery*. London: Gibson Square.

Marquand, D. 2013. *Mammon's Kingdom: An Essay on Britain Now*. London: Allen Lane.

de Melo Resende, V. 2016. 'Discursive representation and violation of homeless people's rights: Symbolic violence in Brazilian online journalism'. *Discourse & Communication* 10(6): 1-18.

Mount, F. 2012. *The new few: or a very British oligarchy.* London: Simon & Schuster.

Ortu, C. 2008. 'The denial of class struggle by British Governments in their anti-union discourse (1978–2007)'. *Critical Discourse Studies* 5(4): 289-301.

Paterson, L., L. Coffey-Glover and D. Peplow. 2016. 'Negotiating stance within discourses of class: Reactions to Benefits Street'. *Discourse & Society* 27(2): 195- 214.

Rowlingson, K. 2012. *Wealth inequality: key facts.* University of Birmingham: Policy Commission on the Distribution of Wealth.

Stiglitz, J., 2012. The Prize of Inequality: How Today's Divided Society Endangers Our Future. New York and London: Norton & Company.

Toolan, M. 2016. 'Peter Black, Christopher Stevens, class and inequality in the Daily Mail'. *Discourse & Society* 27(6): 642-660.

Westergaard, J., 2012. *Who Gets What? The Hardening of Class Inequality in the Late Twentieth Century*. London: Polity Press.

Wilkinson, R. and Pickett, K. 2009. *The Spirit Level: Why Inequality is Better for Everyone.* London: Penguin.

Insubordinate conditional constructions and (inter)subjectivity in spoken Brazilian Portuguese

Luciana Beatriz Bastos Ávila

Instituto de Humanidades, Artes e Ciências/CSC, Universidade Federal do Sul da Bahia, Brazil

Keywords*: insubordinate conditional constructions; intersubjectivity; Brazilian Portuguese*

Abstract

The present study aims at describing and analyzing the uses of insubordinate conditional constructions (ICCs) and their relation to (inter)subjectivity (Langacker, 1990; Nuyts). Insubordination refers to a grammatical and discursive phenomenon, defined as "the conventionalized main clause use of what, on *prima facie* grounds, appear to be formally subordinate clauses" (Evans, 2007, p. 367), as in "e se precisar de um esclarecimento / assim / no meio do negócio hhh // ("*and if you need an explanation / like this / in the middle of something*//", in which, despite of presenting a subordination marker (*"se"/"if"*), it appears independently, without being connected to a matrix clause. Conditional constructions, such the one presented in this example, have been traditionally neglected or treated as peripheral cases in grammar (cf. Quirk; Greenbaum; Leech; Svartvik, 1985, p. 841-842; Huddleston; Pullum, 2002, p. 945). We will take here a different conception of insubordination, synchronic and pragmatic, in which prosody is a central dimension for data analysis (Bossaglia *et al.*, 2017). This work is based on the framework of Language into Act Theory – L-AcT (Cresti, 2000), an extension of Speech Act Theory (Austin, 1962), and Cognitive Linguistics. Methodologically, we use a sample of 20 texts of C-ORAL-BRASIL (Raso; Mello, 2012), a corpus that represents spontaneous speech events of Brazilian Portuguese and has the same architecture design of C-ORAL-ROM corpora (Cresti; Moneglia, 2005), for Italian, Spanish, French and Portuguese. This corpus is prosodically segmented in utterances and tonal units, according to L-AcT (Cresti, 2000), that claims that discourse is organized into informational units, and utterance is taken as the minimal unit of pragmatic interpretation, the linguistic counterpart of an action, the illocution. In our sample, we found 109 conditional constructions, in different syntactic configurations and informational patterns. 23 of the total occurrences are those we consider to be insubordinate, carried out only with the protasis, 21 in the same utterance and 2 in different utterances. ICCs can be divided in two different groups: (a) protasis whose apodosis are not performed; and (b) protasis whose apodosis was performed before the protasis, or vice-versa, and can be recovered from context or performed by another speaker, as in examples in (i) and (ii), respectively:

(i) *ANE: eh /se cê nũ tiver um carrinho que [/1] que sobe aqui //

*ANE: *yeap* / *if you don't have a little car that [/1] that climbs here //*

(ii) *BRU: se ocê vai partir e <falar> cada letra //

*HEL: <uhn> //

*BRU: cê vai falar / “um” / que é a primeira letra da <palavra> //

**BRU: if you gonna break and say each letter //*

**HEL: <uhn> //*

**BRU: you gonna say / “one” / which is the first letter of the word //*

These constructions point out to different degrees of signaling from a subjective or intersubjective perspective of the participants in a communicative scene. Our data confirm the hypothesis that the prosodic configuration is an important dimension in the distinction of pairs of form and meaning in spoken language.

References

Austin, J. (1962). *How to do things with words*. Oxford: The Claredon Press.

Bossaglia, G.; Mello, H. & Raso, T. (2017). Prosody, syntax and pragmatics: insubordination in spoken Brazilian Portuguese. In: STIL 2017 XI Brazilian Symposium in Information and Human Language Technology/V Workshop of Portuguese Description, 2017. *STIL 2017 XI Brazilian Symposium in Information and Human Language Technology and Collocated Events Proceedings of the Conference*, p. 256-265.

Cortes, P. O.; Mello, H. R. (2013). Modalidade no Português Brasileiro: as estruturas condicionais. In: IX Brazilian Symposium in Information and Human Language Technology - STIL, 2013, Fortaleza. *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, p. 163-167.

Cresti, E. (2000). *Corpus di italiano parlato*. Firenze: Accademia della Crusca.

Evans, N. (2007). Insubordination and its uses. In: Nikolaeva, I. (Ed.). *Finiteness*. Theoretical and Empirical Foundations. Oxford: Oxford University Press, p. 366- 43.

Huddleston & Pullum (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Langacker, R. (1990). Subjectification. *Cognitive Linguistics*, 1-1, 5-38.

Nuyts, J. (2001a). Subjectivity as an evidential dimension in epistemic modal expressions. *Journal of Pragmatics*, 33: 383-400.

QUIRK. R.; GREENBAUM, S.; LEECH, G.; SVARTVIK, J. (1985). A comprehensive grammar of English Language. London/New York: Longman.

RASO, T.; MELLO, H. (Eds.). (2012). C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal e DVD multimedia. Belo Horizonte: Ed. UFMG, v. 1.

Of seagulls and unmarried men: Strata in the Old Irish Priscian Glosses

Bernhard Bauer

Department for Early Irish, Maynooth University

***Keywords**: Old Irish, Early Medieval Glosses, Celtic Linguistics*

Abstract

The corpus of the Old Irish glosses on Priscian's encyclopaedic Latin grammar *Institutiones Grammaticae* was first put together and translated by Stokes & Strachan (1903). Since then new manuscripts and glosses have been found. Currently, the following manuscripts are known to contain Old Irish material on Priscian's grammar:

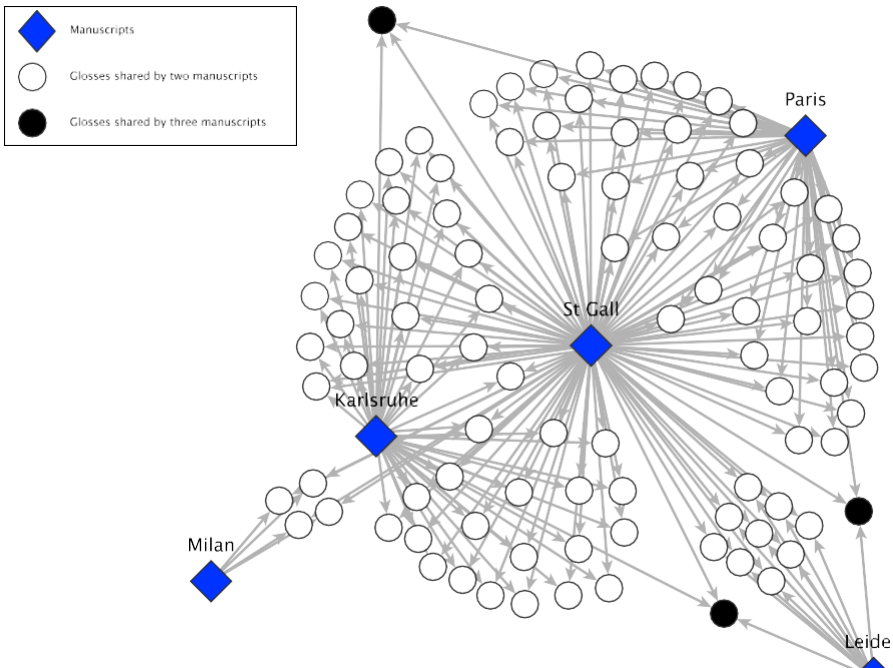
- St Gall, Stiftsbibliothek, ms 904
- Karlsruhe, Codex Augiensis (Reichenau) CXXXII
- Paris, BN ms lat. 10290
- Milan, Bibl. Ambr. Codex Ambrosianus A 138 sup.
- Leiden, Universiteitsbibliotheek, BPL 67

These manuscripts all date to roughly the same period, i.e. the middle of the ninth century. The glosses are now available online as part of the corpus dictionary compiled by Bauer (2015). The most extensively glossed manuscript – the so-called St Gall Priscian – is also subject of a database published by Bauer, Hofman & Moran (2017). The Old Irish Priscian glosses are not homogeneous. Strachan (1903) and more recently, e.g., Lambert (1996) and Roost (2013) have discussed different chronological layers in the Old Irish glosses. These studies, however, focus mainly on the glosses in the St Gall manuscript. The proposed paper will widen the scope and address two main questions:

(1) In what way are the glosses of the different manuscripts interconnected?

(2) Do these interconnections help to identify (chronological) layers?

To answer these questions the main focus of my talk will lie on the glosses in parallel transmission, i.e. glosses on the same lemma in the main text, which are transmitted in more than one manuscript. The following visualisation, created with the software *visone*, represents the network of parallel glosses:



There are only three instances – out of 97 examples – in which three manuscripts share the same gloss. This contrasts with a similar study carried out on five manuscripts with Latin and Old Breton glosses of Orosius' *Historiae adversus paganos*, which yielded a very different ratio of parallel glosses shared by three or more manuscripts, i.e. 16 out of 55 (a third).

The proposed paper will analyse the corpus and try to find reasons for this. Building on these results, I will discuss possible (chronological) layers visible in the corpus. Therefore selected features of the linguistic profiles of the glosses in the specific manuscripts will be compared using the historical comparative method. The occurrence or none- occurrence of certain soundchanges will help identifying the (relative) chronology of the glosses.

References

Bauer, B. (2015). *A dictionary of the Old Irish Priscian Glosses*. http://www.univie.ac.at/indogermanistik/priscian/.

Bauer, B., Hofman, R., & Moran, P. (2017). *St Gall Priscian glosses v2.0*. http://www.stgallpriscian.ie.

Lambert, P.-Y. (1996). Les différents strates de gloses dans le Ms. de Saint-Gall No. 904 (Priscien). In P. Ní Chatháin & M. Richter (Eds.), *Irland und Europa im früheren Mittelalter: Bildung und Literatur* (pp. 187-194). Stuttgart, Germany: Klett-Cotta.

Roost, J. (2013). *Stages of Language Development in the Old Irish Glosses on Priscian*. Unpublished BA Thesis, University College, Dublin.

Stokes, W. & Strachan, J. (1903). *Thesaurus Palaeohibernicus* (Vol. II). Cambridge, United Kingdom: University Press.

Strachan, John (1903). On the Language of the St. Gall Glosses. *Zeitschrift für celtische Philologie* 4.1, 470-492.

Corpus-based study of verbs in Tamazight: syntax and semantics

Nadia Belkacem

Institute of Applied Linguistics, Universitat Pompeu Fabra, Barcelona

Keywords: *Tamazight verbs, argument structure, syntax, semantics, corpus-based study, lexicography*

Abstract

Our research is concerned with the study of verbs in Tamazight with the main objective of discovering and formalizing information useful for lexicography in this language. We concentrate on the Northern-Central variant of Tamazight, otherwise also known as Kabyle (Mammeri, 1976), which we will refer to as Tamazight for short. Some of the syntactic characteristics of verbs in this variant have been studied previously (Basset, 1929; Chaker, 1973; Chaker, 1983) and some of their semantic information has been elaborated in Tamazight lexicography (Dallet, 1953; Dallet 1982). However, syntactico-semantic information on verbs in Tamazight remains far from complete and organized to answer the important demands of lexicography (Lorente, 1994; Osswald, 2015).

Methodologically, our study is based on an online digital corpus of written Tamazight available at [http://ugriw.net]. It's a corpus that contains a little over 1.2 million words from written texts of different genres (novels, essays, newspaper articles, etc.) and provides a powerful search interface (Belkacem, 2015). By using the Global Statistics function of this corpus, we have been able to obtain an ordered list of all the words with their frequency. In that list, it is the different inflexions of verbs that are ordered by frequency. We have thus, as a first step, grouped manually the different inflexions of each verb and summed their frequencies to obtain the total frequency of the verb. The result is the list of 100 verbs shown in table 1 below in order of frequency, from the most to the least frequent. Furthermore, for each verb, we have worked out the frequency of the inflexions by tense/aspect and person, as shown in the example of table 2. The general frequency of a verb is useful data to include in a lexicographic entry. It can, for example, help to ascertain which synonym is more frequent than another. Likewise, knowing which inflexions are more frequent can help to prioritize their inclusion in dictionaries where space is restricted.

Subsequently, using the sentence search function of the online corpus, we carried out a syntactico- semantic analysis on each of the 100 most frequent verbs in their various contexts. For each verb, we grouped the phrases with similar verbal argument structure and meaning and then selected one representative example. In this way, we were able to collect the syntactico-semantic information of each verb as shown in the example of table 3: sense, sentence example, number of arguments, argument structure, semantic roles and constraints, grammatical functions of arguments, syntactic characteristics, voice/diathesis and collocations. Using the same methodology, we extended our analysis to 50 more verbs chosen randomly. This has enabled us to discover and formalize verbal constructions in Tamazight not widely known previously. The final results are in the process of being organized and stored in a database with an interface accessible online in order to serve as a tool for lexicographers when compiling verb entries for Tamazight dictionaries.

References

Basset, A. (1929). La langue berbère (Morphologie). Le verbe: étude des thèmes. Paris: Leroux. Belkacem,

N. (2015). Objective-driven development of the first general language corpus of Tamazight. 8th International Corpus Linguistics Conference. Lancaster University, UK.

Chaker, S. (1973). Le système dérivationnel verbal berbère.Thèse de doctorat de 3ème cycle, Université de Paris V. Ecole Pratique des Hautes Etudes, IVème Section.

Chaker, S. (1983). Un parler berbère d'Algérie (Kabylie).Thèse de doctorat, Université de Paris V. Aix En Provence: Publications Université de Provence.

Dallet, J. M. (1953). Le verbe kabyle. Fort National: Fichier de Documentation Berbère. Dallet, J. M. (1982). Dictionnaire Kabyle-Français. SELAF: Paris.

Lorente, M. (1994). Aspectes de lexicografia: representació i interpretació gramaticals. Tesi doctoral. Universitat de Barcelona.

Mammeri, M. (1976). Tajerrumt n Tmaziɣt (Grammaire de Tamazight). Alger: Bouchene. Osswald, R. (2015). Syntax and Lexicography. In Syntax – Theory and Analysis. Volume 3. Berlin: De Gruyter.

Tables

1	ili be	ini say, tell	uɣal go back, come back, become	ẓeṛ see, know	af find
6	awi take	err give back	izmir can	ddu walk	efk give
11	wali see	ffeɣ go out	vɣu want	ṛuḥ leave	qqim sit down, stay
16	rnu add	aweḍ arrive	kcem go in	xdem do, work	eğğ leave, let
21	aɣ take, buy	issin know	kker get up	sel listen	ṭṭef hold
26	kkes remove	wwet hit	zzi turn around	sɛu have	idir live
31	as-d come	aru write	vdu start	mmet die	van appear, seem
36	ḍṛu happen	ečč eat	ɣli fall	eg make	ekk pass
41	ali go up	enɣ kill	ḥemmel love	ag^{w}ad fear	ezg keep doing sth.
46	veddel change	vedd stand up	mmeslay speak	lhu be good	nadi look for
51	fru resolve	muqel look at	ttu forget	fhem understand	ssiweḍ get sth. somewhere
56	ɣeṛ read	ɛeddi get through	ssiwel call	ṛğu wait	ger put
61	ččaṛ fill up	suffeɣ get sth. out	lḥu walk	mlil meet	rwel run away
66	ldi open	kfu finish	g^{w}ri remain	ddem take	ag^{w}i refuse
71	vnu build	lal be born	kemmel continue	imɣuṛ grow up	qvel accept
76	rfed carry	sers put down	amen believe	cfu remember	azzel run
81	ers come down	ak^{w}i wake up	zwir precede	ssusem keep quiet	steqsi ask
86	izdig be clean	zri pass	sew drink	ssuter request	ffer hide
91	zdeɣ reside	ssirem hope	ḥulfu feel	seddu include	xsi go off (light)
96	nnulfu be invented	rzu visit	gzem cut	ru cry	nnaɣ fight

Table 1: The 100 most frequent verbs in the corpus

verb	pos. 3	inflected forms	tense/aspect	person	frequency
uɣal (go back, come back, become)		yuɣal	preterite	3rd masc.	1612
		tuɣal	preterite	2nd fem.	740
		uɣalen	preterite	3rd masc.	456
		uɣaleɣ	preterite	1st singular	325
		nuɣal	preterite	1st plural	170
		yuɣalen	past participle	invariable	155
		tuɣaleḍ	preterite	2nd singular	113
		yettuɣal	intensive aorist	3rd masc.	112
		uɣalent	preterite	3rd fem.	59
		uɣal	imperative	3rd singular	56
		tettuɣal	intensive aorist	3rd fem.	51
		ttuɣalen	intensive aorist	3rd masc.	46
		yettuɣalen	intensive aorist participle	invariable	35
		tuɣalem	preterite	2nd masc.	21
		nettuɣal	intensive aorist	1st plural	18
		ttuɣalent	intensive aorist	3rd fem.	13
		tettuɣaleḍ	intensive aorist	2nd singular	12
		ttuɣaleɣ	intensive aorist	1st singular	11
		ttuɣal	intensive aorist imperative	2nd singular	09
total frequency					4029

Table 2: Corpus frequencies of the verb *uɣal* in its different inflected forms

Verb Pos. 2	Sense (translation)	Example in Tamazight and translation	Arg	Argument Structure	Semantic Roles	Grammar Function	Syntactic Characteristic	Voice	Comments
uɣal	1. come back	**Uɣal**en-d. *Come back(pret.)they-**d*** *They came back.*	**1**	X: n (*they*)	X: theme (animate)	X: SUB	intransitive	active	Particle **-d** is compulsory.
	2. go back	Y**uɣal** s axxam. *He go back(pret.) to home.* *He went back home.*	**1**	X: y (he)	X: theme (animate)	X: SUB	intransitive	active	Never with particle **-d**.
	3. become	T**uɣal d** taselmadt. *She become(pret.) teacher.* *She became a teacher.*	**2**	X: t (she) Y: taselmadt	X: theme Y: goal	X: SUB Y: COD	transitive direct	active	
	4. close	Tett**uɣal** tewwurt-a? *It close(intensive) door-this?* *Does this door close?*	**1**	X: tewwurt-a	X: patient (inanimate)	X: SUB	intransitive	active	Always in intensive aspect.
	5. be closed	T**uɣal** tewwurt. *It close(pret.) door.* *The door is closed.*	**1**	X: t (it)	X: patient (inanimate)	X: SUB	intransitive	medio passive	
	6. be back in place	Y**uɣal** uqejjiṛ-nni n ṭṭavla. *Go back(pret.)leg-the of table.* *The table leg is back in place.*	**1**	X: uqejjiṛ	X: patient (inanimate)	X: SUB	intransitive	active	
	7. be reconciled to one another	My**uɣal**en. ***My**+go back(pret.)they.* Ms**uɣal**en. ***Ms**+go back (prêt.)they.* *They are reconciled to one another.*	**1**	X: n (they)	X: theme (animate)	X: SUB	intransitive	reciprocal	Never with particle –**d**. Reciprocal can be expressed with morpheme **my** or **ms**.
Collocations and Idiomatic Expressions									

uɣal	1.walk backwards	Y**uɣal timendeffirt**. *He go back(pret.) backwards.* *He walked backwards.*	
	2. fail to keep one's words	Y**uɣal deg wawal**-is. *He go back(pret.) in word-his.* *He failed to keep his word.*	
	3.move back	Y**uɣal ɣer deffir**. *He go back (pret.) to behind.* *He moved back.*	
	4.attract attention	**Tuɣal tmuɣli ar** uqcic-a. *It go back(pret.) look towards boy-this.* *This boy attracts attention.*	
	5.get back in the straight and narrow	Y**uɣal-d s avrid**. *He go back(pret.) to path.* *He got back in the straight and narrow*	

Table 3: Example of syntactico-semantic information for the verb *uɣal*

The discursive construction of a serial killer: A corpus-based linguistic and affective analysis on agency

José Vicente Benavent Cháfer

Catalan Philology and General Linguistics Department, University of Barcelona (UB)- Faculty of Philology, Spain

***Keywords**: corpus functional stylistics analysis, idiolect, sociolect, agency, process types, affective components.*

Abstract

Serial killing and authors of serious criminal offences alike are often portrayed as deviant beings who do not fit into society, thus detaching them from the collective identity. While the traditional focus adopted in discursive psychology remains on the offender's nature

(Potter & Wetherell 1987, Auburn & Lea 2003), this paper explores how serial killers self-represent discursively through linguistic choices and affective components. Stylistics, or rather the notion of 'style' is adapted in this research as the way in which language is employed by a given speaker in a definite context for a certain end (Leech & Short 2007: 9). Since the serial killers examined (Ted Bundy, Jeffrey Dahmer, and Dennis Rader) share circumstantial facts, their idiolect reflects a unique language use motivated by the stylistic features inherent to their personality in a similar life phase, which differs from other individuals' usage (Dittmar 1996: 111). Therefore, this twofold analysis distinguishes between individual styles (idiolects) and collective styles (serial killers as social actors). The social actor's theory claims that the aforementioned individualised lexical choices conform to a 'collective identity as a mental model that comprises cognitive and affective components' (Koller 2012: 19), which enables comparison at the discourse level.

The main aim of this dissertation is to offer an insight into the idiolect of three serial killers by exploring their actions and emotions when referring to their victims, material objects, and crimes. A thorough analysis is offered by means of two theoretical frameworks from different fields accounting for processes and emotional states. Firstly, Halliday's transitivity system focuses on process types, namely material, mental, relational, verbal, behavioural, and existential processes (1970: 213), thus providing a linguistic angle to the serial killers' description of actions. Ben-Zeev's conceptual emotional realm, on the other hand, presents a taxonomy on emotional evaluations which classifies types of emotions depending on the positive/negative impact on the senser (reflecting upon oneself or others), as well as the praiseworthiness of agents, the desirability of the agents' fortune, and the appealingness as a whole of an agent (2000: 94).

As for the creation of the corpus, the subjects' (Dennis Rader, Ted Bundy, and Jeffrey Dahmer) speech is examined in six interviews from different contexts (three excerpts from a legal context, and three interviews obtained in a psychological context). For the quantitative analysis, the concordancing tool AntConc accounts for the frequencies of words and their collocations, thus undertaking a 'corpus functional stylistic analysis' (Mao et al. 2014: 70). In the qualitative section, quotations are made from the interview's transcripts so as to reveal the relevance of process types, affective components, and how/if they contribute to blame attribution (agency). Both analytical phases investigate each serial killer's idiolect individually, while also providing a collective account on the killers' sociolect thereafter, in both legal and psychological contexts.

Findings suggest that control and sexual desire are overriding themes in the serial killer's speech. By denying intention with cognitive cues or alluding to a controlling sexual force through existential processes, the subjects attempt to alleviate their sentencing by appealing on the grounds of insanity. However, this paper does not venture into determining the subjects' psychiatric condition.

References

Auburn, T., & Lea, S. (2003). Doing Cognitive Distortions: A Discursive Psychology Analysis of Sex Offender Treatment Talk. *British Journal of Social Psychology*, 42(1), 281-298.

Ben-Zeev, A. (2000). *The Subtlety of Emotions. Cambridge*. Cambridge, MA, US: MIT Press, 11-98.

Dittmar, N. (1996). Explorations in 'Idiolects'. In R. Sackmann & M. Budde (Eds.), *Theoretical Linguistics and Grammatical Description: Papers in Honour of Hans-Heinrich Lieb* (pp. 111-115). Amsterdam: Benjamins.

Halliday, M.A.K. (1970). Language Structure and Language Function. In J. Lyons (Ed.),

New Horizons in Linguistics (pp. 213-229). Harmondsworth: Penguin.

Koller, V. (2012). How to Analyse Collective Identity in Discourse: Textual and Contextual Parameters. *Critical Approaches to Discourse Analysis across Disciplines*, 5(2), 19-38.

Leech, G.N. & Short, M. (2007). *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. Edinburgh: Pearson Longman, 9-19.

Mao, Z., Li, N., & Xue, J. (2014). Corpus Functional Stylistic Analysis of Modal Verbs in Major Barbara and Its Chinese Versions. *Theory and Practice in Language Studies*, 4(1), 70-78.

Potter, J. & Wetherell, M. (1987). *Discourse and Social Psychology: Beyond Attitudes and Behaviour*. London: Sage, 102-134.

Revisiting ephemeral concessive adverbial subordinators: A study based on the *Penn Parsed Corpora of Historical English*

Cristina Blanco García

Department of English and German, University of Santiago de Compostela

***Keywords**: ephemerality; concessive subordinators; variation; Penn Parsed Corpora of Historical English*

Abstract

Adverbial subordinators have become a popular area of research over the last few decades both from a synchronic and a diachronic perspective (cf., among others, the monographs by Kortmann (1997), Pérez Quintero (2002) and Lenker (2010), and the collective volume edited by Lenker and Meurman-Solin (2007). The history of adverbial subordinators in English is intriguing, as the inventory of these connectives has been subject to great fluctuation over the different stages of the history of the language. The Middle English period was crucial in the development of adverbial subordinators, not only because of the bulk of connectives which arose at this stage, but also because many of these innovations have been preserved in Present Day English. In contrast, the Early Modern English period (sixteenth and seventeenth centuries) witnessed the specialization and establishment of those Middle English innovations. Besides, this period was also crucial in the development of what Kortmann (1997: 301) denominates 'ephemeral' adverbial subordinators, i.e. those that were added to the inventory of adverbial connectives in Late Middle English or, more commonly, Early Modern English, but did not have a lasting effect and died out eventually.

In this context, the present paper sets out to provide a comprehensive picture of three 'ephemeral' subordinators from the domain of concession, namely *albeit*, *howbeit* and *notwithstanding*, illustrated in (1)-(3).

(1) *I may well leese a worde on yow, or letter, **Albeit** I shal be never the better.* (c1386, Chaucer *Legend Good Women* (1359); OED s.v. *albeit*, conj. 1b)

(2) *I would fain have access and presence to The King even **howbeit** I should break up iron doors*. (1634, S. Rutherford Lett.; OED s.v. *howbeit* conj.)

(3) *Which synne worþily askide and disceruede a peyne þat schulde haue noon eende, **notwiþstondynge** þe deede of synne had eende.* (?a1425, *Orcherd of Syon* (Harl. 3432) (1996) 22; OED s.v. *notwithstanding* conj.)

Given that 'ephemeral' adverbial subordinators are a low-frequency phenomenon, this paper complements an earlier preliminary study of 'ephemeral' concessive subordinators in two small 'diagnostic' corpora, namely the *Helsinki Corpus* and ARCHER (Author 2017), by using data from three considerably larger corpora, the *Penn- Helsinki Parsed Corpus of Middle English*, the *Penn-Helsinki Parsed Corpus of Early Modern English* and the *Penn-Helsinki Parsed Corpus of Modern British English*. The *Penn Parsed Corpora of Historical English* will be used here as a base line, and will be

complemented with data from the *Oxford English Dictionary* and the *Middle English Dictionary*, together with their quotation databases.

The analysis will consider the following issues: (i) the overall quantitative distribution of the selected connectives in the material across time; (ii) their etymology and development; (iii) their combination with the 'pleonastic' *that* (cf. Beal 1988; Rissanen 1999); (iv) the use of correlative elements in the main clause as a reinforcement of the concessive relationship; (v) the verbal mood favoured by these subordinators in the sub-clause; and (vi) the position of concessive adverbial clauses introduced by these connectives in relation to their main clauses.

References

Blanco-García, C. (2017). Ephemerality in concessive subordinators. Evidence from the history of English. In Bemposta, S. *et al*. (Eds.), *New Trends and Methodologies in Applied English Language Research III: Synchronic and Diachronic Studies on Discourse, Lexis and Grammar Processing* (pp. 59-81) Linguistic Insights Series. Bern: Peter Lang.

Beal, J. (1988). Goodbye to All 'That'? The history and present behaviour of optional 'that'. In Nixon G. & J. Honey (Eds.), *An Historic Tongue. Studies in English Linguistics in Memory of Barbara Strang* (pp. 49-66). London and New York: Routledge.

Kortmann, B. (1997). *Adverbial Subordination: A Typology and History of Adverbial Subordinators Based on European Languages*. Berlin & New York: Mouton de Gruyter.

Lenker, U. (2010). *Argument and Rhetoric. Adverbial Connectors in the History of English*. Berlin & New York: Mouton de Gruyter.

Lenker, U. & Meurman-Solin A. (Eds.) (2007). *Connectives in the History of English*. Amsterdam & Philadelphia: John Benjamins.

Pérez Quintero, M. J. (2002). *Adverbial Subordination in English. A Functional Approach*. Amsterdam & New York: Rodopi.

Rissanen, M. (1999). Syntax. In Hogg, R. M. (Ed.) *The Cambridge History of the English Language*. Vol. 3. 1476-1776 (pp. 187-331) Cambridge: Cambridge University Press.

Newspaper reports of gender conflict in sports: A metapragmatic analysis

Patricia Bou Franch

Institut Universitari de Llengües Modernes Aplicades-IULMA. Universitat de València

***Keywords**: conflict, evaluation, impoliteness, metapragmatics, sexism*

Abstract

The aim of this paper is to explore evaluations of im/politeness and sexism in the digital press. The paper examines newspaper reports of the conflict involving tennis player Serena Williams during the finals of the 2018 US Open. In the course of some brief interactions that took place during and after the match, the athlete expressed great indignation at the Umpire's decisions to penalise her with three code violations, and accused him of lying and stealing points from her. She also claimed that sexism was behind these decisions. This event was widely reported in the news, which, in turn, triggered a sizeable number of comments in the Internet. Newspaper reports discussed participants' behaviour and the athlete's moral claim of sexism. Reports of this event, therefore, constitute a good point of departure for the study of evaluations of impoliteness and sexism in the press.

Research in pragmatics has recently focused on the interconnections between evaluation and im/politeness (Haugh, 2015; Kádár & Haugh, 2013; Kádár & Márquez Reiter, 2015). Although politeness and impoliteness are elusive terms, difficult to define (Culpeper, Haugh & Kádár, 2017, Garcés-Conejos Blitvich, 2018, O' Discroll, 2017), there seems to be general agreement among discursive scholars to conceptualize im/politeness in terms of assessments. Hence, there is a need to explore the (moral) grounds on which im/politeness assessments are made. Further, sexism is understood in connection to evaluations of impoliteness for, as Mills (2008) argues, sexism comes about when values and beliefs about gender are not shared by participants in an event (Bou-Franch, 2013; Bou-Franch & Garcés-Conejos Blitvich, 2014).

The present study adopts a corpus-based metapragmatic approach to the study of evaluations of impoliteness and sexism. The use of corpus linguistics methodologies in (im)politeness research started a decade ago and is currently gaining increasing scholarly attention (Culpeper 2009, 2011; Culpeper and Hardaker, 2016; Culpeper & Haugh, 2014; Haugh, 2018). These studies fall within what is known as corpus-based metapragmatics, which involves the study of conflict-related labels and comments through a combination of quantitative and qualitative methods (Haugh, 2018: 624).

In order to carry out this study, an ad hoc corpus of newspaper articles (n = 103 texts; +99,200 words) was compiled using *Nexis UK*. All newspaper texts in the corpus contained the search terms "Williams" and "US Open" and were published in American and British broadsheets the week following the Women's Singles final of the US Open, when the confrontation between the player and the umpire took place. The corpus was analysed quantitatively and qualitatively. The analysis involved the identification of collocations and concordances, using Wordsmith 5.0. All concordances containing metapragmatic comments were qualitatively analysed and classified into semantically similar groups to unveil the relevant discourse prosodies (Bou-Franch, 2017; Baker, 2005, 2010; Santaemilia & Maruenda, 2014) underlying evaluations of im/politeness and sexism.

References

Baker, Paul. (2005). *Public discourses of gay men*. London: Routledge.

Baker, Paul. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh: Edinburgh University Press.

Bou-Franch, P. (2013). Domestic violence and public participation in the media: The case of citizen journalism. *Gender & Language*, 7(3), 275-302. doi: 10.1558/genl.v7i3.275

Bou-Franch, P. (2017). Víctimas de la violencia de género en la prensa digital española. In J. A. Cutillas Espinosa, J. M. Hernández Campoy, R. M. Manchón Ruiz & F. Mena Martínez (Eds.), *Estudios de filología inglesa: Homenaje al profesor Rafael Monroy* (pp. 89-108). Murcia: Universidad de Murcia.

Bou-Franch, P., & Garcés-Conejos Blitvich, P. (2014). Gender ideology and social identity processes in online language aggression against women. *Journal of Language Aggression and Conflict*, 2(2), 226-248. doi: 10.1075/jlac.2.2.03bou

Culpeper, J. (2009). The metalanguage of impoliteness: Using Sketch Engine to explore the *Oxford English Corpus*. In P. Baker (Ed.), *Contemporary corpus linguistics* (pp. 64-86) London: Continuum. Culpeper, J. (2011). *Impoliteness: Using language to cause offence*. Cambridge: Cambridge University Press.

Culpeper, J., & Hardaker, C. (2016). Pragmatics and corpus linguistics. In P. Baker & J. Egbert (Eds.), *Triangulating methodological approaches in corpus linguistic research* (pp. 124-137). London: Routledge.

Culpeper, J., & Haugh, M. (2014). *Pragmatics and the English language*. Basingstoke: Palgrave Macmillan.

Culpeper, J., Haugh, M., & Kádár, D. (2017). Introduction. In J. Culpeper, M. Haugh &

D. Kádár (Eds.), *The Palgrave Handbook of Linguistic Impoliteness* (pp. 1-8). London: Palgrave-Macmillan.

Garcés-Conejos Blitvich, P. (2018). Globalization, transnational identities, and conflict talk: The superdiversity and complexity of the Latino identity. *Journal of Pragmatics, 134*, 120-133. doi: 10.1016/j.pragma.2018.02.001

Haugh, M. (2015). Im/politeness, social practice and the participation order. *Journal of Pragmatics, 58*: 52-72. doi: 10.1016/j.pragma.2013.07.003

Haugh, M. (2018). Corpus-based metapragmatics. In A. H. Jucker, K. P. Schneider & W. Bublitz (Eds.) *Methods in pragmatics* (pp. 619-643). Berlin: De Gruyter Mouton

Kádár, D., & Haugh, M. (2013). *Understanding Politeness*. New York: Cambridge University Press.

Kádár, D., & Márquez Reiter, R. (2015). (Im)politeness and (im)morality: Insights from intervention. *Journal of Politeness Research, 11*(2), 239–260

Mills, S. (2008). *Language and sexism*. Cambridge: Cambridge University Press.

Sobre el diseño de corpus orales discursivos construidos por y para lingüistas: el corpus Ameresco y el corpus Val.Es.Co. 2.0

Adrián Cabedo Nebot[a]; Andrea Carcelén Guerrero 2[b]

[a]*Departmento de Filología Española, Universitat de València, España;* [b]*Departmento*

de Filología Espaola, Universitat de València, España

***Keywords**: corpus orales, diseño, plataforma, búsqueda*

Abstract

En esta presentación tenemos como objetivo exponer las complicaciones inherentes al desarrollo de corpus textuales (Rojo Sánchez, 2010, 2015, 2017) y, en concreto, de corpus que recojan material hablado en forma de conversaciones, entrevistas, etc. (Love, Dembry, Hardie, Brezina, & McEnery, 2017). Los problemas que se encuentran son evidentes, dado que los corpus que actualmente existen sobre el español parten de variantes diatópicas concretas o recogen material distinto a la conversación, como es el caso de entrevistas (Cestero Mancera, 2012; Fernández-Ordóñez, 2005) o de material procedente de la radio o la televisión. Otros problemas se vinculan con los sistemas de transcripción y con el etiquetado de determinados fenómenos; también con el formato usado en la recolección de los datos (etiquetas TEI, JSON, etc.) o con los programas que se usan en la transcripción, como ELAN (Max Planck Institute, 2017), PRAAT (Boersma & Weenink, 2017), etc. Aun así, en la actualidad existen corpus de conversaciones que abren el camino a nuevas maneras de entender la recogida y la consulta de este tipo de material; como muestra, podría hablase del proyecto del Clarin Spokes, del corpus Cola (Jorgensen, 2007) o del corpus Eslora, entre algunos otros.

En ese marco de corpus conversacionales de español, presentaremos una plataforma

diseñada para facilitar el acceso y el uso de dos corpus orales: el corpus Ameresco (Albelda Marco & Estellés Arguedas, 2017), que recoge conversaciones de diferentes variantes geográficas de español hablado: Iquique (Chile), Habana (Cuba), Tucumán (Argentina), Valencia (España), México DF (México), Barranquilla (Colombia)..., y el corpus Val.Es.Co. 2.0, centrado en la variedad española de Valencia (España).

En general, los corpus Ameresco y Val.Es.Co. tienen una orientación marcadamente discursiva y, más en concreto, conversacional. Así mismo, la unidad de entrada, aquella que genera la primera entrada de transcripción, es una unidad fónica, el grupo entonativo, que se define como secuencia de habla entre pausas o entre inflexiones melódicas singulares. En el sistema de bases de datos relacionales, estos grupos se tokenizan en palabras y se compilan en unidades superiores, las intervenciones. De este modo, se puede acceder a información léxica (etiqueta morfosintácticamente mediante *Treetagger*), a información discursiva (intervenciones y duraciones) y a información sociolingüística (características de edad, sexo y nivel social de los hablantes), así como a características generales de la conversación (duración, número de hablantes, localización, etc.).

En definitiva, el interés principal de estos corpus (a través de la plataforma generada) es facilitar el acceso al material a los investigadores, dado que se entiende que los modos de consultar la información pueden ser variadas y responden a intereses de investigación diferenciados. En ese sentido, la plataforma construida para consultar los corpus Ameresco y Val.Es.Co. se basa en tecnología de bases de datos relacionales (PostgreSQL) y en la posibilidad de exportar el material consultado, tanto de los registros generados en la búsqueda como de todas las conversaciones (transcripciones y audios) presentes en ambos corpus.

Referencias

Albelda Marco, M., & Estellés Arguedas, M. (2017). *Corpus Ameresco*. www.esvaratenuacion.es

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer. Retrieved from http://www.praat.org/

Cestero Mancera, A. M. (2012). El Proyecto para el estudio sociolingüístico del español de España y América (PRESEEA). *Español Actual: Revista de Español Vivo*, (98), 227–236.

Fernández-Ordóñez, I. (2005*). El Corpus Oral y Sonoro del Español Rural*. Cantoblanco: Universidad Autónoma de Madrid, 2005.

Jorgensen, A. (2007). COLA. Un corpus oral de lenguaje adolescente. In *Discurso y oralidad: homenaje al profesor José Jesús de Bustos Tovar* (Vol. 1, pp. 225–234). Madrid: Arco Libros, [2007].

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.22.3.02lov

Max Planck Institute. (2017). ELAN (Version 5.2). Nijmegen: Max Planck Institute for Psycholinguistics. https://tla.mpi.nl/tools/tla-tools/elan/

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

Rojo Sánchez, G. (2015). Corpus textuales del español. In *Enciclopedia de Lingüística Hispánica* (Vol. 2, pp. 285–296).

Rojo Sánchez, G. (2017). Sobre la configuración estadística de los corpus textuales. *Lingüística*, 33(1), 121–134.

The construction of speaker stance and argumentation in persuasive political rhetoric

Ana Belen Cabrejas-Peñuelas & Mercedes Díez-Prados

Departament de Filologia Anglesa i Alemanya-IULMA. Universitat de València

Departamento de Filologái Moderna. Universidd de Alcalá de Henares

Keywords: *evaluation, argumentation, political discourse, stance, persuasion*

Abstract

For decades, the language of evaluation has been studied to find out how writers/speakers adopt attitudes towards their readers/listeners and the content they what to communicate (Eggins and Slade, 1997; Hood, 2004; Hunston, 2000; Martin, 2000; Martin and White, 2005). Both political discourse and political language seem well- suited to express evaluation, as politicians need to show an image of absolute certainty and intellectual power that serves to persuade their audience, while they discredit the opposing candidate (and the opposing party that the opponent represents). For that purpose, they make use of evaluative devices: positive attitude markers about themselves, negative ones for criticizing their adversaries, judgement for praising or criticizing a person's capacity or propriety and appreciation markers for evaluating things and states of affairs, following Martin and White's (2005) categorization of Affect. Pre-electoral debates, which are face-to-face confrontations before an audience and/or television viewers where political candidates spar in a dialogical duel, are ideal places for politicians to present themselves in a positive light, while they discredit their opponents and their actions. Thus, evaluative language plays a crucial role in an attempt to persuade the audience. The present study expands on our previous work on evaluation (Authors 2014, Authors 2018) in the Spanish 2011 pre-electoral debate, which applied Martin and White's (2005) categorization of Affect and Hunston's (2000) evaluation model. We now test the persuasive force of the evaluative language of the Spanish 2011 pre-electoral debate as used within an argumentative structure, following Toulmin's (1958, 2003) categories for arguments. For that purpose, we take account of the results obtained in our previous study on evaluation (Authors 2014) and add a study of the distribution of evaluative devices in the argumentative structure of the same pre-electoral debate. In order to shed some light into the politicians' words, we attempt to provide an answer to the following research questions:

RQ1. Is there a connection between the persuasive force of the politicians' words and the structure of their arguments?

RQ2. Are there any similarities and differences in the connection between the form of an argument and evaluative language between the Socialist and Conservative politicians?

Preliminary results indicate that both features (i.e. evaluation and argumentation) are closely involved in the persuasiveness of an argument. The present paper sheds some

light on the argumentative role played by evaluative items by examining their location (i.e. as part of the claim, data, warrant, qualifier or rebuttal) and function in the argumentative structure.

References

EGGINS, Suzanne and Diana Slade. 1997. *Analyzing Casual Conversation*. London: Cassell.

HOOD, Susan. 2004. *Appraising Research: Taking a Stance in Academic Writing*.

Doctoral dissertation, University of Sydney, NSW, Australia.

HUNSTON, Susan. 2000. "Evaluation and the Planes of Discourse: Status and Value in

Persuasive Texts". In *Evaluation in Text* edited by Susan Hunston and Geoff Thompson, 176-207. Oxford: OUP.

MARTIN, J.R. 2000. "Beyond Exchange: Appraisal Systems in English". In *Evaluation*

in Text: Authorial Stance and the Construction of Discourse edited by Susan Hunston and Geoff Thompson, 142-175. Oxford: OUP.

MARTIN, JR and P.R.R. White. 2005. *The Language of Evaluation: Appraisal in*

English. Basingstoke: Palgrave Macmillan.

TOULMIN, Stephen. [1958], 2003. *The Uses of Argument*. Cambridge: CUP.

Evaluation in political speeches:
A study of Theresa May's Brexit speeches

Ana Belén Cabrejas-Peñuelas; Rosana Dolón

Departament de Filologia Anglesa i Alemanya-IULMA. Universitat de València

Keywords: *evaluation; stance; political speech; Brexit; political discourse*

Abstract

Numerous studies (Hunston & Thompson [2000] 2003; Martin 2000; Martin & White 2005; Simon-Vandenbergen et al. 2007, 2008 and White 2002, among others) agree on the fact that speakers and writers use different linguistic mechanisms to convey their personal attitudes and assessments and that the journalistic, academic and political discourse may differ in terms of such mechanisms. Political language in general, and political speeches in particular, is a type of persuasive discourse that is especially suited for the expression of evaluation, since politicians need to create a distinct profile for themselves in order to persuade the audience of their standpoints. In doing so, they present their viewpoints and those of their political Others (typically, opposing candidates, but also political bodies such as the European Union) and evaluate these viewpoints and their opponents' actions (i.e. criticize their opponents and put themselves in a positive light). To this end, they use positive and negative attitude markers to praise themselves and their deeds and to criticize their opponents' negative qualities and actions. The final purpose of the discourse is to convince and appeal to the public, something that seems unlikely to happen without an expression of a particular stance. The evaluative uses of political discourse have been studied by a number of researchers (Simon-Vandenbergen et al.

2007, 2008; Author 2014) to find out how positive/negative evaluation contributes to persuasion in texts. The objective of the present study is to expand on our previous work on evaluation (Author 2014, Author in press), which applied Martin & White's (2005) categorization of Affect and Hunston's (2000, 2008, 2011) evaluation model in Spanish and American pre-electoral debates. We now test Hunston's (2000, 2008, 2011) evaluation model in three British political speeches, in which PM Theresa May deals with the UK's relationship with the European Union after Brexit. The characteristics of the speeches are equitable in that they all deal with the same topic, have similar word counts and are given within a little over one year. In order to gain insight into May's message, the research questions that we attempt to answer are the following:

RQ1. Does MP Theresa May's evaluative language contribute to the persuasive power of her political speeches? If so, to what extent?

RQ2. Are there any similarities and differences between her political speeches in terms of the expression of evaluation of status (i.e. degree of alignment between a proposition and the world, Hunston 2008: 65)?

The results indicate that in general there is a preference for statements that refer to the actual world (World-reflecting statements in Hunston's classification) rather than World-creating ones. This is not but May's strategy to be seen as taking a down-to-earth political position before the audience. There are however differences between the speeches, depending on whether the speaker wants to show a more rational stance or a more opinionated one.

References

AUTHOR. (2014)

AUTHOR. (In press)

HUNSTON, Susan. 2000. "Evaluation and the Planes of Discourse: Status and Value in Persuasive Texts." In *Evaluation in Text: Authorial Stance and the Construction of Discourse*, edited by Susan Hunston and Geoff Thompson, 176-207. Oxford: OUP.

—. 2008. "The Evaluation of Status in Multi-modal Texts." *Functions of Language* 15 (1): 64-83.

—. 2011. *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. New York, NY: Routledge.

HUNSTON, Susan and Geoff Thompson, ed. (2000) 2003. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: OUP.

MARTIN, J.R. (2000) 2003. "Beyond Exchange: Appraisal Systems in English." In *Evaluation in Text: Authorial Stance and the Construction of Discourse*, edited by Susan Hunston and Geoff Thompson, 142-175. Oxford: OUP.

MARTIN, J. R. and Peter R.R. White. 2005. *The Language of Evaluation: Appraisal in English*. Basingstoke: Palgrave Macmillan.

SIMON-VANDENBERGEN, Ann-M. 2008. "Hillary Clinton vs. Barack Obama: A Linguistic Study of Appraisal in Political Speeches." Master's thesis, University of Ghent.

SIMON-VANDENBERGEN, Ann-M., Peter R. R. White and Karin Aijmer. 2007. "Presupposition and 'Taken-for-granted' in Mass Communicated Political Argument: An Illustration from British, Flemish and Swedish Political Colloquy." In *Political Discourse in the Media: Cross-cultural Perspectives*, edited by Anita Fetzer and Gerda A. Lauerbach, 31-74. Amsterdam: Benjamins.

WHITE, Peter R.R. 2002. "Appraisal-the Language of Evaluation and Stance." In *Handbook of Pragmatics,* edited by Jef Verschueren, Jan-Ola Östman, Jan Blommaert, and Chris Bulcaen, 1-27. Amsterdam: Benjamins.

Boosters in Early Modern English: A Sociolinguistic approach

Javier Calle-Martín[a]; Laura Esteban-Segura[a]

[a]*Department of English, University of Málaga, Spain*

Keywords*: boosters; early modern English; sociolinguistics*

Abstract

Intensifiers or degree modifiers may be defined as "linguistic elements which convey the degree or the exact value of the quality expressed by the item they modify" (Huddleston and Pullum 2002: 585). Depending on their function, Quirk et al. (1985: 445-446) classify them into *amplifiers* (those scaling upwards from an assumed norm) and *downtoners* (those scaling downwards from an assumed norm). *Amplifiers* are divided into *maximizers* and *boosters*, according to whether the item expresses the absolute intensity (*absolutely true*, *completely depressed*) or simply a high degree of intensity (*very sad*, *awfully sorry*), respectively. *Downtoners*, in turn, present two functional variants: *approximators* or *moderators*, which slightly reduce the degree of the adjective that they modify (*quite expensive*, *rather unusual*); and *diminishers*, those that have a stronger attenuating force (*a little weird*, *somewhat damaged*).

The interest of scholars in intensifiers has been renewed during the last decades due to the advances in corpus and theoretical linguistics, semantics and grammaticalization in particular. As a consequence, multi-disciplinary approaches to the topic have been added to ground-breaking monographs in the field such as those by Stoffel (1901), Borst (1902) and Fettig (1934), offering comprehensive inventories of intensifying adverbs (Méndez-Naya 2008a: 213). These recent approaches have been conceived from different perspectives: (a) the distribution of intensifiers across social groups, varieties, registers and text-types (Macaulay 2002; Nevalainen 2008); (b) the competition of different intensifiers across time (Méndez-Naya and Pahta 2010); and (c) the study of individual items and the origin of their intensifying function (Méndez-Naya 2008b; Rissanen 2008).

All this considered, the present paper is conceived within the first perspective and proposes the study of degree words from a historical sociolinguistic perspective in order to determine whether there is any intrinsic preference for particular intensifiers across gender and, more importantly, social groups in Late Middle and early Modern English. The study therefore pursues the following objectives: (i) to analyze the use and distribution of boosters in the historical period 1350-1710; and (ii) to evaluate the socio- linguistic dimension of the phenomenon in the light of the informants' profiles. The source of evidence comes from the *Corpus of Early English Correspondence*.

References

Borst, E. (1902). *Die Gradadverbien im Englischen* [*Degree Adverbs in English*]. Heidelberg: Winter.

Fettig, A. 1934. *Die Gradadverbien im Mittelenglischen* [*Degree Adverbs in Middle English*]. Heidelberg: Carl Winter Universitätsbuchhandlung.

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

Macaulay, R. (2002). 'Extremely interesting, very interesting, or only quite interesting?' Adverbs and social class. *Journal of Sociolinguistics*, 6, 398-417.

Méndez-Naya, B. (2008a). Special issue on English intensifiers: introduction. *English Language and Linguistics*, 12, 213-219.

Méndez-Naya, B. (2008b). On the history of *downright*. *English Language and Linguistics*, 12, 267-287.

Méndez-Naya, B. & P. Pahta. (2010). Intensifiers in competition: the Picture from early English medical writing. In I. Taavitsainen & P. Pahta (Eds), *Early modern English medical texts* (pp. 191-213). Amsterdam: John Benjamins.

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

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

Rissanen, M. (2008). From 'quickly' to 'fairly': on the history of *rather*. *English Language and Linguistics*, 12, 345-359.

Stoffel, C. (1901). *Intensives and downtowners (Anglistische Forschungen 1).*

Heidelberg: Winter.

Auxiliary verbs in perfect tenses in Late Modern English fiction: conservative or influenced by motion situations?

Nuria Calvo Cortés

Universidad Complutense Madrid, Spain

***Keywords**: Late Modern English, perfect tenses auxiliaries, corpus of novels, motion situations*

Abstract

Studies on perfect tenses (Rydén & Brorström, 1987; Kytö, 1997) have confirmed that it was at the end of the 18th century when *have* started to become the preferred verb as opposed to *be*, which had been the predominant auxiliary until then. However, some verbs resisted the combination with *have* and continued to be combined with *be* (e.g. *gone* or *grown*). Most of these studies have concluded that men writers adopted the use of *have* much more quickly than women writers, and, therefore, the latter have been

considered conservative in this respect.

The present study aims at analysing the choice of auxiliary verb in perfect tenses that contain some verbs that resisted the use of *be* for a longer time, and which semantically encode the idea of motion, either physical or metaphorical (e.g. *become, come, gone, grown* and *passed*).

The two questions that will be dealt with in this analysis are, first, whether there were differences between the men and the women writers regarding the auxiliary verb used in the 18th and the 19th centuries; and second, whether the components of motion situations motivated the preference of the auxiliary verb. The components of motion situations include the Figure that moves, the Ground to which it moves, and the Path along which the Ground moves (Talmy, 2000). Two previous studies on women writers have suggested that both the different types of *Figure* and *Ground* may have conditioned the choice of auxiliary verb partly (Calvo Cortés, in press).

For the present study a corpus of novels was compiled specifically. It is divided into four subcorpora: women's novels of the 18th century, men's novels of the 18th century, women's novels of the 19th century and men's novels of the 19th century. The sizes are similar, roughly 2 million words in each of the 18th century corpora, and 2 and a half million words in the other two, but normalization of data was required despite these similarities. The software programme *AntConc* was used to extract the examples, which were later analysed manually.

The preliminary results show that, despite the differences among the authors and their own personal choices, in the 18th century there is a higher frequency of *be* than *have* in general, and it is in the 19th century when men seem to opt for *have* more often than women, which would confirm what had been stated in previous studies. However, the conclusions also point to a possible influence of the semantics of motion situations in the choice of auxiliary verbs. For instance, there seems to be a preference of a combination of *have* with non-human Figures more often than *be*, as well as the higher presence of metaphorical Grounds with *have* rather than with *be*. Both aspects indicate that *have* may be the preferred option in metaphorical contexts in general.

Further research is needed to rule out editors' manipulation and to observe if there were any differences in the evolution of these auxiliaries depending on different social classes, and different types of texts, not only in novels.

References

Calvo Cortés, Nuria. (in press). Conservatism or the influence of the semantics of motion situation in the choice of perfect auxiliaries in Jane Austen's letters and novels. In C. Claridge & B. Birte (Eds.), *Developments in English Historical Morpho-Syntax* (pp.175-198). Amsterdam: John Benjamins Publishing.

Kytö, M. (1997). Be/Have + past participle: the choice of auxiliary with intransitives from Late Middle to Modern English. In M. Rissanen, M. Kytö & K. Heikkonen (Eds.), *English in Transition: Corpus-based Studies in Linguistic Variation and Genre Styles* (pp. 19-85). Berlin: Mouton de Gruyter.

Rydén, M. & Brorstörm, S. (1987). *The Be/Have Variation with intransitives in English*. Stockholm: Almqvist & Wiksell International.

Talmy, L. (2000). *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.

"Too old to write love letters" and "too old for romantic leads": a contrastive corpus-based study of age collexemes in Late Modern Britain

Nuria Calvo Cortés; Marta Nadales Ruiz

Department of English Studies: Linguistics and Literature, Universidad Complutense de Madrid, Spain

***Keywords**: Corpus Linguistics, Late Modern Britain, Age Studies, Collostruction, Cultural Studies*

Abstract

Age is a culturally related concept that has evolved through time. This explains that in different cultures around the world somebody may be considered "too young" or "too old" for something, whereas this is not probably the case somewhere else. Similarly, as the age of life expectancy increases, a young and an old person nowadays may be seen differently when compared to some centuries ago.

Different genres reflect people's interest in age. Biographies represent a special interest both in the field of Age Studies and Discourse Analysis. They reflect individual identities, and shared cultural norms and conventions in a particular field and time.

Similarly, since the appearance of novels in the 18th century, they also have reflected the social concerns of the time they represented. Both types of texts, thus, constitute a valuable source of material to examine the concept of age from a linguistic perspective. The analysis of the collostructions "too old for/to" and "too young for/to" is the focus of attention of this corpus-based study. Stefanowitsch and Gries (2003) introduced the concept of collostruction, based on the construction grammar framework (Goldberg, 1995). The aim is to understand what people were considered, or considered themselves, to be "too old" or/and "too young" for in Late Modern Britain, and at what age in two contrastive corpora. One corpus includes the 18th and 19th centuries biographies included in the *Oxford Dictionary of National Biography*. The other comprises a selection of British novels written by male and female authors in the same period. Specifically, the interest lies in the analysis of the lexical items associated with such collostructions, according to the taxonomy proposed by Gries and Stefanowitsch (2010), and studies the contrasts between the structures in both corpora.

The preliminary results show that there are some differences in the lexical items used in the corpora. Whereas the biographies prefer collexemes related to profession and social function, in the novels social status and cognitive development are the predominant choices. Results also point to an awareness of the early stages of an established age for certain events. In fact, regarding social aspects, there was, for instance, a legal age for getting married, to join the army or to enter university, which is clearly manifested in the corpora. However, cognitive functions appear to be more subjective and not connected to the conventions of the time.

In conclusion, there is a predominance of social function associated with age over biological development in the biographies. In contrast, the novels reflect an interest in marriage and cognitive processes. Therefore, the use of contrastive corpora proves necessary to obtain a detailed picture of normative age in Late Modern Britain.

References

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago & London: University of Chicago Press.

Gries, S.T., & Stefanowitsch, A. (2010). Cluster analysis and the identification of collexeme classes. In S. Rice, & J. Newman (Eds.), *Empirical and experimental methods in cognitive/functional research*, (pp. 73-90). CSLI: Stanford.

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

Spring cleaning time: SFL-inspired CADS to overcome "dusty corners"

María Calzada Pérez

Departament of Translation and Communication, Universitat Jaume I, Castellón de la Plana, Spain

***Keywords**: CADS; Systemic Functional Linguistics; quantitative methods; qualitative methods; European Parliament; ECPC*

Abstract

Corpus-based work has been such a prolific contributor to discourse analysis that it has managed to build its own academic niche known as corpus-assisted discourse studies (or CADS; see Partington, Duguid, and Taylor 2013). CADS has particularly excelled inthe quantitative study of differences between (larger or smaller) compilations of texts along the synchronic dimension. However, it has its weak spots, or in Taylor and Marchi's (2018, 9) words, its "dusty corners"; that is, "neglected aspects of analysis and under- researched topics or text types".

As a way forward, the present paper aims to contribute to the discussion of these dusty corners and does so in at least two ways. On the one hand, it reinforces CADS' departing theoretical component by reviewing some basic concepts from systemic functional linguistics. Among them, the notions of realisation, instantiation and individuation (Martin 2011) are defined and explained as a gateway to prolific corpus-assisted work. On the other hand, and from this theoretical stance, the paper focuses on the combination of quantitative and qualitative methods and practical tools for the identification of linguistic absence and the examination of similarities between comparative (Xiao 2011), cross-lingual corpora along the synchronic and diachronic dimensions. The quantitative-

qualitative methods presented here draw inspiration from Baker (2010, 2018), and Baker and Egbert (2016). The practical tools resorted to in our study (see below) are both of a more traditional kind (e.g. word lists, keyword lists and concordances) and a more innovative flavour (e.g., detailed consistency relations, key keywords and associates, lockwords, c-collocates and s-collocates) (see Alan Partington et al 2013).

The corpora upon which theoretical concepts, methods and tools are implemented are part of the European Comparable and Parallel Corpus Archive of Parliamentary Speeches (ECPC). Compiled at the Universitat Jaume I (Spain), this archive (of almost 100 million tokens) covers the period 2004-2011 and consists of original and translated English and/or Spanish proceedings of the European Parliament, the Spanish Congreso de los Diputados and the British House of Commons. ECPC incorporates contextual (sociolinguistic and sociocultural) and metalinguistic information, the latter relating to the speakers' status, gender, constituency, party affiliation, birth-date, birth place, post, and institutional body and sub-body of representation, all encoded using XML annotation. The ECPC Archive has a modular structure that makes comparison along the spatial and temporal axes possible.

In short, this study illustrates that SFL-inspired CADS has effective means to perform the spring cleaning of some dusty corners. It comprises four main sections. Section 1 presents a theoretical framework deriving from the notions of realisation, instantiation and individuations. Section 2 reviews CADS' quantification and qualification synergies and selects specific (traditional and innovative) tools, notably keywords, detailed consistency and lockword. Section 3 performs an analysis of individuation/instantiation upon the ECPC corpora, moving from theory, through the generation of big quantitative data to the qualitative examination of specific linguistic levels such as text, sequence and clause, showing not just that the fusion of SFL and CADS is possible but highly desirable. Nevertheless, continuous self-reflection, self-criticism and self-awareness are a must to keep moving forward.

References

Baker, P. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, P., & J. Egbert, J. (Eds). (2016). *Triangulating Methodological Approaches in Corpus-Linguistic Research*. New York: Routledge.

Baker, P. (2018). Which Techniques of Down-Sampling Best Complement a Corpus-Assisted Discourse Analysis? A Case Studyon Press Representations of Obesity. Presented in *4th International Corpora and Discourse Conference*, Lancaster University, 22nd 24th June, 2018.

Martin, J.R. (2011). Systemic Functional Linguistics. In K. Hayland & B. Paltridge, *The Continuum Companion to Discourse Analysis*. London and New York: Continuum.

Partington, A., Duguid, A. & Taylor, C. (Eds.). (2013). *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins Publishing Company.

Taylor, C., & Marchi, A. (Eds.). (2018). *Corpus Approaches to Discourse: A Critical Review*. Milton Park, Abingdon, Oxon; New York: Routledge.

Xiao, R., (Ed.). (2010). *Using Corpora in Contrastive and Translation Studies*. Newcastle upon Tyne: Cambridge Scholars.

La recuperación de información en corpus orales del español: transcripción y etiquetado de corpus

Andrea Carcelén Guerrero

Departamento de Filología Española, Universitat de València, España

***Keywords**: lingüística de corpus, transcripción, etiquetado, discurso oral*

Abstract

El propósito de esta comunicación es comparar, por un lado, el diseño y la construcción de corpus orales del español con el objetivo de señalar sus diferencias y afinidades en cuanto al sistema de transcripción y codificación utilizados. Por otro, analizar y contrastar los motores de búsqueda de los que dispone cada corpus para determinar qué opciones de recuperación de la información ofrecen a los usuarios. Para este fin se han seleccionado los corpus siguientes: CORPES XXI, PRESEEA, COLA, COSER, ESLORA y AMERESCO. Todos ellos están disponibles en línea.

CORPES XXI (http://www.rae.es/recursos/banco-de-datos/corpes-xxi), elaborado por la Real Academia de la Lengua, constituye uno de los corpus de referencia del español. PRESEEA (http://preseea.linguas.net/) es uno de los proyectos de corpus más extensos ya que recoge entrevistas semidirigidas de más de 40 ciudades para estudios sociolingüísticos. El corpus COLA (http://www.colam.org/om_prosj-espannol.html), especializado en lenguaje adolescente, es el único que requiere registro previo para su consulta; COSER (http://www.corpusrural.es/) recoge muestras del español hablado en enclaves rurales. ESLORA (http://eslora.usc.es/) presenta 20 horas de grabaciones de discurso oral espontáneo del español hablado en Galicia. Por último, el corpus AMERESCO (http://esvaratenuacion.es/corpus/corpus_tablelist.php), proyecto que recoge conversaciones espontáneas obtenidas de forma secreta con petición de autorización y se encuentra aún en fase de desarrollo.

Para ello, en primer lugar, se ha recopilado y analizado la información que proporciona cada proyecto sobre los sistemas de transcripción y codificación que han seguido. Estos documentos pueden consultarse en línea, a excepción del sistema seguido por AMERESCO. En los casos de PRESEEA y CORPES XXI, también pueden descargarse en formato PDF. En segundo lugar, se ha elaborado un cuadro comparativo en el que se recogen los datos extraídos del análisis del motor de búsqueda de cada corpus.

Los primeros resultados apuntan a que los corpus que presentan un sistema de codificación que refleja mejor las particularidades de la oralidad son ESLORA y AMERESCO, no solo porque utilizan un sistema de etiquetas que incluyen marcas específicas para este tipo de discurso, sino que además trabajan con *software* que permite alinear el audio y el texto.

Respecto del análisis de los motores de búsqueda, los resultados indican que, a excepción de PRESEEA que solamente permite realizar búsquedas por texto y filtrar los resultados atendiendo a las características de los hablantes (sexo, grupo etario,

nivel de estudios y ciudad de procedencia), todos los corpus analizados presentan un nivel de filtrado y recuperación de información muy detallado. Se pueden hacer consultas simples y avanzadas, tanto por búsqueda exacta como por lema o etiqueta gramatical.

En cuanto a la posibilidad de descargar el audio y la transcripción completa, solamente lo permiten PRESEEA y AMERESCO. ESLORA permite la consulta, pero la descarga solo es posible previa solicitud al equipo. CORPES XXI, COSER y COLA recuperan la transcripción parcial, no completa.

Algunos autores han señalado la necesidad de seguir pautas estandarizadas e internacionales a la hora de diseñar corpus orales de manera que la representación sea lo más próxima a la oralidad, así como que permitan su reutilización (Torruella y Llisterri 1999, Hidalgo y Sanmartín 2005, Briz 2012).

Referencias

Albelda Marco, M. y Briz Gómez, A., (2009). Estado actual de los corpus de lengua española hablada y escrita: I+D. *El español en el mundo: Anuario del Instituto Cervantes*, pp. 165-226.

Briz Gómez, A. (2012), Los déficits de los corpus orales del español (y de algunos análisis). *Cum corde et in nova grammatica. Estudios ofrecidos a Guillermo Rojo*. Servizo de Publicacións e Intercambio Científico da Universidade de Santiago de Compostela, pp. 115-137.

Enghels, R., Vanderschueren, C. y Bouzouita, M. (2015), Panorama de los corpus y textos del español peninsular contemporáneo. *Manuel des anthologies, corpus et textes romans. In Manuals of Romance Linguistics* 7, pp. 147-170.

Hidalgo, A. y Sanmartín J. (2005), Los sistemas de transcripción de la lengua hablada, *Oralia* 8, pp. 13-36.

Llisterri, J., (1999), Transcripción, etiquetado y codificación de corpus orales, *Revista española de lingüística aplicada*, Vol. 1. Extra, pp. 53-82.

Payrató, L. (1995), Transcripción del discurso coloquial, en Cortés, L. (ed.). *El español coloquial. Actas del I Simposio sobre Análisis del Discurso Oral*. Almería: Universidad de Almería, pp. 43-70.

Recalde, M. y Vázquez, V. (2009), Problemas metodológicos en la formación de corpus orales, en *A survey of corpus-bases research*, Cantos Gómez, P. y Sánchez Pérez, A. (eds.), pp. 51-64.

Torruella, J. y Llisterri, J. (1999), Diseño de corpus textuales y orales, *Filología e informática. Nuevas tecnologías en los estudios filológicos*, Milenio: Barcelona, pp. 45-81.

A Contrastive Analysis of Multimodal Metadiscourse in Linguistics, Engineering and Medicine Academic Papers

Maria Luísa Carrió-Pastor

Departamento de Lingüística Aplicada, Universitat Politècnica de València, Spain

Keywords: metadiscourse, multimodality, academic English.

Abstract

Nowadays, due to technology and the free access to information, we have different ways of accessing to academic research, which include many visual elements that guide readers and illustrate ideas and research results. The elements that guide readers in textual academic discourse have been analysed by several researchers (Hyland and Tse, 2004; Hyland, 2005; Carrió-Pastor, 2014, 2016; Jiang and Hyland, 2016), but visual metadiscourse has received less attention (Kumpf, 2000). Additionally, the analysis of multimodal discourse in English for specific purposes has received till now little attention. To fill this gap, this paper focuses on the multimodal analysis of academic papers from the specific fields of linguistics, engineering and medicine. The paper pays attention to the identification of interactive metadiscourse devices (Hyland, 2005) and visual metadiscourse elements (Kumpf, 2000). The hypothesis of this paper is that writers with different academic backgrounds but the same linguistic background express in academic English differently in academic papers. Thus, the main objectives are, first, to study the interactive metadiscourse devices in the three specific fields of academic discourse; second, to identify the visual metadiscourse elements used in the three specific fields of knowledge, and finally, to classify the multimodal elements used to cohere the different parts of the paper with the aim of guiding the reader textually as well as visually. To this end, sixty academic papers were compiled from journals included in Journal Citation Reports (2016 edition). Twenty belonged to the specific field of linguistics, twenty to engineering and the last twenty to medicine. The corpus was analysed with the help of specific tools and also manually to check the data extracted. The results showed that there are in fact differences in the way academic writers use textual and visual metadiscourse, the study revealed that engineering writers prefer the use of visual metadiscourse meanwhile linguistic researchers prefer textual metadiscourse. Also, textual and visual patterns associated to the different specific fields of knowledge were identified. Finally, conclusions were drawn that answered the objectives of the study. This paper shows some of the results of the research project FFI2016-77941-P (Ministerio de Economía y Competitividad, Spain).

References

Carrió-Pastor, M. L. 2014. Cross-cultural variation in the use of modal verbs in academic English. *Sky, Journal of Linguistics*, 27, 153-166.

Carrió-Pastor, M. L. 2016. A contrastive study of interactive metadiscourse in academic papers written in English and in Spanish. In Francisco Alonso Almeida, Laura Cruz García & Víctor González Ruiz (Eds.), *Corpus-based studies on language varieties.* Bern: Peter Lang.

Hyland, K. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*, 7-2, 173-192.

Hyland, Ken and Tse, Polly. 2004. Metadiscourse in academic writing: a reappraisal. *Applied Linguistics*, 25, 156-177.

Jiang, F. K. and Hyland, K. 2016. Nouns and academic interactions: A neglected feature of metadiscourse. *Applied Linguistics*, published online: 1-25. doi:10.1093/applin/amw023. [Accessed 12/12/2016]

Kumpf, E. P. 2000. Visual metadiscourse: Designing the considerate text. *Technical Communication Quarterly*, 9 (4), 401-424.

Main syntactic features of English predicative prepositional phrases

Antonio Vicente Casas Pedrosa

Department of English Studies, University of Jaén, Spain

Keywords: *syntax, English, predicative prepositional phrase, subject complement, substitution, coordination, ellipsis, fronting, stranded preposition*

Abstract

This paper is aimed at analyzing the main syntactic features of predicative prepositional phrases (henceforth, PPPs) in English. Little attention has been paid to them so far since neither their relevance nor their frequency are supposed to be enough. Different examples illustrating those syntactic characteristics will be obtained from the *BNC* and the *ICE-GB* as well as from various grammars and monolingual dictionaries.

First of all, they can be described as those phrases which are headed by a preposition whose Complement (C) is a Noun Phrase (NP) and which perform the function of Subject Complement (C^s) at clause level. Such is the case of "of value" in [1]:

> [1] It can be concluded that Adomnan's life of Columba is *of value* for the structure of society which seems not to have changed greatly between Columba's time and Adomnan (ICE-GB:W1A-002 #115:1).

Thus, there is an intensive relationship between the Subject (S) and this unit since the latter predicates something of the S, thus conveying a condition, feature, quality, or state which is attached to the S.

Secondly, depending on the nature of the meaning they express, the examples in italics below can be replaced by the Adjective Phrases, Verb Phrases, and clauses included between square brackets:

> [2] If you do not listen your responses may be "*off beam*" [wrong] and you may appear to others to be rather stupid (BNC:EEB 603).
>
> [3] Mary then came to see me and was at this point *in tears* [crying] over the incident and her treatment by Mr Smith (ICE-GB:W1B-021 #72:5).
>
> [4] He was *a little out of breath* [breathing fast and with difficulty], having hurried to the town meeting (BNC:CFJ 311).

Moreover, it is possible to find examples of PPPs in coordination, what is an argument proving the similarities between them and Adjective Phrases:

[5] Doctors at the hospital say Mr Crowther is continuing to show signs of slight improvement although he's still *in intensive care* and *in a critical condition* (BNC:K22 1981).

[6] She is young and *in good health* (Quirk *et alii*, 1985, p. 732).

Many examples of PPPs are the result of a previous process of ellipsis (whereby some words have been omitted from a given structure; e.g. "an aspect" after the copulative verb "is" in [7]) or fronting (for instance, [8]):

[7] And uh o one particular aspect <,> o of the delays uh i is uh uh *of considerable relevance* uh on this occasion <,> (ICE-GB:S2A-063 #7:1:A).

[8] *At odds with the mayor* remain the rendents of the condos, who are willing to fight a long battle over the additional taxes (Huddleston & Pullum, 2002, p. 1389).

In addition, there are examples of stranded prepositions in exclamative, interrogative, and relative clauses, as shown in sentences [9]-[11]:

[9] *What a mess* she was *in* […] (ICE-GB:W2F-003 #107:1).

[10] *What stature* is she *of*? (Jespersen, 1909-49, vol. III, p. 398).

[11] But I think Heseltine's got what it takes to win the next election and get us out of

this mess that we are *in* (ICE-GB:S2B-003 #77:1:M).

Finally different conclusions will be drawn on the previous analysis. It can be highlighted, among others, that the syntactic behaviour of PPPs is very similar to that of Adjective Phrases.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, England: Longman.

British National Corpus. (1994). Retrieved from http://www.natcorp.ox.ac.uk/ ;http://corpus.byu.edu/bnc/.

Deuter, M., Bradbery, J., & Turnbull, J. (Eds.) (2015[9]). *Oxford advanced learner's dictionary of current English*. Oxford, England: Oxford University Press. Retrieved from https://www.oxfordlearnersdictionaries.com/.

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

Jespersen, O. (1909-49). *A Modern English grammar on historical principles* (7 vols.). London, England: Allen & Unwin.

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language. Working with the British component of the International Corpus of English*. Amsterdam, Netherlands: John Benjamins.

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

Rundell, M. (Ed.) (2019[3]). *Macmillan English dictionary for advanced learners*. Oxford, England: Macmillan. Retrieved from http://www.macmillandictionary.com/.

Simpson, J. (Ed.) (2018) (updated quarterly). *Oxford English dictionary online*. Oxford, England: Oxford University Press. Retrieved from http://www.oed.com/.

Survey of English Usage. (1998). *The International Corpus of English. The British component (ICE-GB). Release 1*. London, England: University College London.

The words that count: exploring theoretical questions in corpus linguistics

Richard Chapman

Dipartimento di Studi Umanistici, Università degli Studi di Ferrara, Italy

***Keywords**: theory; sampling; discourse analysis; context*

Abstract

This paper attempts to re-examine some fundamental theoretical issues concerned with the use of corpus linguistics to investigate language. What should be at the basis of the design of our corpora? Why are these elements important and how do they ensure significant and reliable results from interrogation of corpus data?

Using examples from small-scale corpora as points of comparison with the increasingly large general language corpora currently available or being constructed, a critical reading of corpus linguistic practice is proposed. The intention is not to question the fundamental value of corpus-based linguistic inquiry, but to remind us of the complexity of the tasks involved. Various issues are presented, including the tension produced by using 'hard' data to understand abstract ideas; the almost insoluble problem of truly effective and representative sampling; the task of dealing with 'natural' errors, mistakes and slips-of-the-tongue in counting frequencies; the role of Discourse Analysis and whether it can be based on 'the word' (as opposed to other linguistic indicators and aspects of context); whether smaller, more closely grounded and contextualised corpora are better suited to discourse analytical approaches, and if there can be a fruitful interaction between large-scale, mass-corpora and small, virtually hand-made computer-based archives.

This series of questions, coming directly from attempts to use corpus linguistics in 'real' case studies, prompts thinking about the most fundamental aspects of corpus linguistics methodology. What do we really think corpus linguistic analysis can tell us? Does it offer answers to questions like: What is language? How does a particular language work? What are the relations between form and sense? Can we perfect our descriptive grammars through corpus linguistics? Is it possible to describe how language is meaningful in certain contexts? Does corpus data analysis adequately take account of

problems of reception and interpretation in making judgements? How can we make corpus linguistics more objective and reliable?

It is not by chance that these issues present themselves again at this juncture. Big data and significantly increased computational capability renders corpus linguistics, on the one hand, not quite so unique as it may have seemed thirty or forty years ago, and on the other placed in a context where data is plentifully available. Corpus linguists have the highly-skewed, huge-but-imperfect archive, Google, at their fingertips: the relationship between 'serious' corpus linguistics and the amateur version needs to be theorised rather than ignored as irrelevant. Computing power and the internet also make the creation of numerous small-scale corpora an easy task. It is possible to compile a meaningful (in the sense of potentially genuinely useful) corpus in one day and begin to interrogate it the next. The present paper proposes a fruitful dynamic between these two opposite ends of the corpus linguistic spectrum: there are greater possibilities for iterative research; questions and puzzles might be thrown up more quickly and in more relief; pragmatics can be explored on a corpus level without losing details of context and synchronic and diachronic contrasts better made.

The paper concludes with the suggestion that corpus linguistics must always be highly cautions in its claims and unendingly vigilant in its sampling methodology, but that it is just about to enter a new phase which offers additional potential in the attempt to describe and understand language.

References

Bakhtin, M.M. (1986). *Speech Genres and Other Late Essays*. Austin: University of Texas Press.

Blommaert, J. (2005). *Discourse*. Cambridge: CUP.

Blommaert, J. (2010). *The Sociolinguistics of Globalisation*. Cambridge: CUP.

Brezina, V. (2018). *Statistics in Corpus Linguistics. A Practical Guide.* Cambridge: CUP.

Bruner, Jerome. S. (1990). *Acts of Meaning*. Cambridge: Harvard University Press.

Chapman, R. (2017). Naming or Shaming? Presentations of the Self in Specialised Weblog Discourse. In *Iperstoria X* Fall/Winter 2017

Chapman, R. (2018). Corpus Linguistic Analysis: How far can we go? In Costa, A.P., Reis, L.P., Souza, F.N. de, Moreira (Eds.) *Computer Supported Qualitative Research, 2nd International Symposium on Qualitative Research*, (pp. 42-47) ISQR 2017, A. Springer.

Dörnyei, Zoltan. (2007). *Research Methods in Applied Linguistics*. Oxford: OUP.

Duranti, A. (1997). *Linguistic Anthropology*. Cambridge: CUP.

Fairclough, N. (1995/2010). *Critical Discourse Analysis*. Harlow: Longman.

Hoey, M., Mahlberg, M., Stubbs, M., Treubert, W. (2007). *Text, Discourse and Corpora*. London: Continuum.

Hyland, K., Chau, M. H., Handford, M. (2012). *Corpus Applications in Applied Linguistics*. London: Bloomsbury.

McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-based Language Studies*. Abingdon: Routledge.

Milizia, D. *Phraseology in Political Discourse*. (2012). Milano: Edizioni Universitarie Lettere Economia Diritto.

Ricoeur, P. (1978/2003). *The Rule of Metaphor*. Abingdon: Routledge.

Taylor, D. and Marchi, A. (Eds.) (2018). *Corpus Approaches to Discourse. A Critical View*. Abingdon: Routledge.

The words that count: exploring theoretical questions in corpus

Linguistics

Richard Chapman

Dipartimento di Studi Umanistici, Università degli Studi di Ferrara, Italy

Keywords*: theory; token; type; frequency*

Abstract

The paper stems from direct experience in the creation of a number of small-scale language corpora for specific inquiry into various aspects of present-day linguistic behaviour. The suggestion is made that the very basic assumptions of corpus linguistic inquiry are in need of constant examination and informed readjustment if they are to provide any truly reliable help in discovering hidden or misunderstood aspects of language use. Even the ostensibly simplest aspect of language appearing in a corpus requires careful handling, informed interrogation and caution in analysis. The concept of the word, for example, requires greater theoretical examination.

Analysing a series of corpora compiled over some years leads to the emergence of apparently simple questions that are of surprising complexity: difficulties in counting and the accurate identification of types; questions concerning definitions; problems with complexity of usage; pragmatic issues affecting attribution of frequency (use of metaphor etc.). All of these represent challenges to confidence in the analytical conclusions we can associate with corpus linguistic data and its evaluation, and so, it is hoped, will prove of general interest to anyone involved in corpus-based approaches to linguistic investigation. Corpus linguistics is built upon the collection of bodies of language use, and attempts to generalise about language from the interrogation of instances of use. In other words, it is approaching *langue* through the reading of extensive examples of *parole*. This means, of course, that all relevant instances must be collected accurately, through transparent and effective sampling techniques, and the subsequent data interrogated with precision. Corpus linguistics is now mature enough to have debated this methodology, but from the current research doubts arise that suggest a re-evaluation of present corpus linguistic practice might be of some value.

After describing issues relating to simpler aspects of identifying frequencies in meaningful ways, the paper attempts an analysis of pragmatic issues underlying the

linguistic behaviour that will be encapsulated in any corpus. If we are dealing with *parole*, the paper suggests, we can never completely ignore the pragmatic element in any instance of language use, and assumptions about text type or genre are inadequate to ascribe purpose to each discrete example of usage. The question is whether rather naïve judgements of meaning (the pragmatics behind a particular utterance) are enough to allow researchers to identify and count linguistic elements with sufficient precision.

A greater layer of pragmatic complexity presents itself in later corpora where it is recognised that the pragmatics of language use is often dynamic rather than static and so resists categorisations that might currently be made. Observation of data also tempts us to propose the idea that pragmatic behaviour can be varied and manifold simultaneously: the same group of words is perhaps being employed both to inform and to humiliate, to reassure and to intimidate at one and the same time, leaving the corpus analyst with the task of sifting through the possibilities of assignment and making a subjective judgement that only reflects a part of the whole sociolinguistic world of the item. We are left to conclude that corpus linguistic study should, of course, be aware of pragmatic and discourse elements in play, but should also be cautious in its conclusions when data are deeply complex.

References

Bakhtin, M.M. (1986). *Speech Genres and Other Late Essays*. Austin: University of Texas Press.

Blommaert, J. (2005). *Discourse*. Cambridge: CUP.

Blommaert, J. (2010). *The Sociolinguistics of Globalisation*. Cambridge: CUP. Brezina, V. (2018). *Statistics in Corpus Linguistics. A Practical Guide.* Cambridge:

CUP.

Bruner, Jerome. S. (1990). *Acts of Meaning*. Cambridge: Harvard University Press.

Chapman, R. (2017). Naming or Shaming? Presentations of the Self in Specialised

Weblog Discourse. In *Iperstoria X* Fall/Winter 2017

Chapman, R. (2018). Corpus Linguistic Analysis: How far can we go? In Costa, A.P., Reis, L.P., Souza, F.N. de, Moreira (Eds.) *Computer Supported Qualitative Research, 2nd International Symposium on Qualitative Research,* (pp. 42-47) ISQR 2017, A. Springer.

Dörnyei, Zoltan. (2007). *Research Methods in Applied Linguistics*. Oxford: OUP. Duranti, A. (1997). *Linguistic Anthropology*. Cambridge: CUP.

Edwards, J. (2009). *Language and Identity*. Cambridge: CUP.

Fairclough, N. (1995/2010). *Critical Discourse Analysis*. Harlow: Longman.

Hoey, M., Mahlberg, M., Stubbs, M., Treubert, W. (2007). *Text, Discourse and Corpora*. London: Continuum.

Hyland, K., Chau, M. H., Handford, M. (2012). *Corpus Applications in Applied Linguistics*. London: Bloomsbury.

McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-based Language Studies*.

Abingdon: Routledge.

Milizia, D. *Phraseology in Political Discourse*. (2012). Milano: Edizioni Universitarie Lettere Economia Diritto.

Ricoeur, P. (1978/2003). *The Rule of Metaphor*. Abingdon: Routledge.

Taylor, D. and Marchi, A. (Eds.) (2018). *Corpus Approaches to Discourse. A Critical View*. Abingdon: Routledge.

Corpus-assisted discourse analysis with students: The use of individual do-it- yourself corpora in English for Academic Purposes

Maggie Charles

University of Oxford, United Kingdom.

Keywords: *academic writing; disciplinary differences; do-it-yourself corpora; student corpus use; confirmative claims; evaluative adjectives*

Abstract

Corpus-assisted discourse studies concerns the analysis of discourse types using corpus tools and techniques in combination with close reading of selected texts (Partington, et al., 2013). This combined quantitative and qualitative approach has been widely applied, especially to investigations of political and media discourse (e.g. Baker et al., 2013; Bednarek, 2018). The approach has also been used in the study of academic written discourse and has revealed significant discoursal differences between disciplines in such areas as citation practices and metadiscourse (e.g. Hyland, 2000, 2005). In this paper, I suggest that students of English for Academic Purposes (EAP) are themselves well- placed to carry out such investigations individually within the discourse of their own disciplines. Indeed, given the extent of the differences found not only between, but also within disciplines (Godnić Vićić & Jarc, 2015; Kwan, 2017; McGrath, 2016), such individual student investigations may offer one of the most effective ways of ensuring discipline-appropriate EAP input, especially for advanced-level graduate students. The present paper reports findings from a course in which students built their own do-it-yourself (DIY) corpora from research articles in their bibliographies and carried out their own corpus-assisted discourse analyses. The data come from courses given over the period 2013- 2017 and cover 104 students in total. Classes were aimed at graduates and were multidisciplinary (37% natural sciences, 33% social sciences, 30% humanities). Students used AntFileConverter (Anthony, 2017) for converting files to plain text format for corpus inclusion and AntConc (Anthony, 2016) for corpus searching. Although not all students kept up-to-date records of their corpora, those available range in size from approximately 34,000 – 1,800,000 words. The class procedure was as follows: students first studied examples of a target discourse function in short academic extracts, discussing how the function was carried out and identifying linguistic signals which could act as

probes for corpus investigations. They completed worksheets to record and comment on the results of their corpus searches and then compared their findings with those of students in different disciplines. For example, the terms *studies, study, research, work*

and *literature* were searched to shed light on discipline-specific ways of reviewing the literature (cf. Flowerdew and Forest, 2009). Working with a corpus of 314,979 words in medicine, Giovanna found that the most frequent term in her corpus was *study*, which was predominantly used with time adjectives *previous* (4 hits) and *recent* (5). However, she also noted the evaluative adjective *elegant* (2).

> ***This very elegant study*** *analyzes the gene expression profile of human circulating monocytes isolated from healthy volunteers exposed to LPS.*

By contrast, Brigitte's corpus of 327,143 words in history revealed *work* to be the most frequent term. Time adjectives were again prevalent: *early/er* (7) and *recent* (8), but Brigitte highlighted the evaluative adjectives *interesting* (3), *insightful* (2) and *pathbreaking* (2).

> ***In her pathbreaking work*** *on nationalism in colonial Tanzania, Susan Geiger focused on the role of nonliterate women.*

Students' comparisons of their findings not only raised their awareness of disciplinary differences, but also focused their attention on the ways in which disciplines use evaluative attributive adjectives to construct 'confirmative claims' (Kwan et al., 2012). This paper reports in more detail on the student data and argues that the use of corpus-assisted discourse analysis with DIY corpora provides a viable way of addressing the individual disciplinary needs of advanced EAP students.

References

Anthony, L. (2016). AntConc (Version 3.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Anthony, L. (2017). AntFileConverter (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Bednarek, M. (2018). *Language and television series: A linguistic approach to TV dialogue*.

Cambridge: Cambridge University Press.

Baker, P., Gabrielatos, C., & McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge; New York: Cambridge University Press.

Flowerdew, J., & Forest, R. (2009). Schematic structure and lexico-grammatical realization in corpus- based genre analysis: The case of *research* in the PhD literature review. In M. Charles, D. Pecorari, & S. Hunston (Eds.), *Academic writing: At the interface of corpus and discourse* (pp. 15–36). London: Continuum.

Godnić Vićić, S., & Jarc, M. (2015). Research articles in sociology: Variation within the discipline. In P. Thompson & G. Diani (Eds.), *English for academic purposes: Approaches and implications* (pp. 79–101). Newcastle upon Tyne: Cambridge Scholars.

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Harlow: Longman.

Hyland, K. (2005). *Metadiscourse*: *Exploring interaction in writing*. London: Continuum.

Kwan, B. (2017). A cross-paradigm macro-structure analysis of research articles in Information Systems. *English for Specific Purposes*, *45*, 14–30.

Kwan, B., Chan, H., & Lam, C. (2012). Evaluating prior scholarship in literature reviews of research articles: A comparative study of practices in two research paradigms. *English for Specific Purposes*, *31*(3), 188–201.

McGrath, L. (2016). Self-mentions in anthropology and history research articles: Variation between and within disciplines. *Journal of English for Academic Purposes*, *21*, 86–98.

Partington, A., Duguid, A., & Taylor, C. (Eds.). (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam: John Benjamins.

Personajes de serie. La construcción expresiva de personajes a través de los diálogos en *Águila Roja*

Luisa Chierichetti

Dipartimento di Lingue, Letterature e Culture, Università degli Studi di Bergamo, Italy

Keywords*: análisis del discurso; estilística de corpus; discurso telecinemático; series televisivas; caracterización; expresividad*

Abstract

Nuestra propuesta parte de la posibilidad de tener acceso a los 116 guiones originales de una afortunada serie de televisión contemporánea, *Águila Roja*, producida por Globomedia para Radio Televisión Española y emitida entre 2009 y 2016.

Nos planteamos contribuir a la exploración del uso de la lengua española en los medios de comunicación y, especialmente, en el ámbito de la ficción televisiva, uno de los espacios culturales de más éxito y desarrollo en la sociedad contemporánea. Además, la posibilidad de investigar de forma directa los guiones facilita nuestra comprensión de las creencias internalizadas de los autores, que se transmiten a través del diálogo a una comunidad de televidentes potencialmente globalizada. La creación de guiones de series es una tarea colectiva llevada a cabo por escritores que trabajan en equipo y se sitúa en específicas condiciones sociales de producción dentro de la industria televisiva (Richardson, 2010: 63-64). Los guionistas crean un discurso ficticio, mediado por las convenciones del género comunicativo, basándolo en sus intuiciones y su conocimiento implícito de cómo funciona la conversación, algo que comparten con su audiencia; de ahí que, finalmente, mediante su análisis, también podamos tener una idea de los modelos cognitivos de las personas, o de los "esquemas para la producción de conversación" (Lakoff & Tannen 1994: 139), especialmente a nivel pragmalingüístico (Bubel 2006: 61). En esta propuesta, acotamos nuestra investigación a la creación del personaje a través de su expresividad (Kozloff 2000: 33-34). Para ello, tenemos en cuenta que las

elecciones lingüísticas en el discurso telecinemático cobran significado a través de su asignación a determinados personajes y de su utilización en contrastes dialógicos, en el telón de fondo de las ideologías de los lenguajes dominantes (Androutsopoulos 2012: 147). Los rasgos expresivos pueden ser privativos de un personaje y formar su identidad individual, o bien pueden ser compartidos por un grupo de personajes, configurando una identidad social (Bednarek 2010, 2018). Pueden construir una identidad bien dinámicamente – por ejemplo en una determinada escena o episodio – o bien repetidamente, formando un patrón reconocible a lo largo de la serie.

Para nuestro objetivo, nos servimos de las técnicas de LC aplicándolas al corpus completo (1,299,646 tokens) y a distintos subcorpus que confeccionamos con los diálogos de los personajes principales y con los diálogos de conjuntos de personajes pertenecientes a determinados grupos sociales. Con la ayuda del programa SketchEngine nos proponemos identificar patrones lingüísticos frecuentes relacionados con una serie restringida de recursos expresivos que puedan definirse de manera formal, para luego relacionar su ocurrencia con la construcción de un determinado tipo de identidad expresiva (Bednarek 2010: 123) basada en la actitud explícita o implícita de los personajes (Culpeper 2001; Martin & White 2005; Kaplan 2004), construida bien con su discurso, bien con las acotaciones, las descripciones y las indicaciones de voz y de encabezado de escena que aparecen en los guiones. Para ello, nos basaremos en la búsqueda de palabras clave y de n-gramas, datos recabados mediante la comparación de los subcorpus con el corpus completo, utilizado como corpus de referencia, y luego desambiguados a través de la herramienta de Concordancias, también utilizada con la función CQL (Corpus Query Language) para realizar búsquedas complejas. La anotación mediante etiquetas de estructura nos permitirá incluir una reflexión sobre la estabilidad y evolución de los personajes a lo largo de los capítulos de la serie.

Referencias

Androutsopoulos, J. (2012). Introduction: Language and society in cinematic discourse, *Multilingua* 31, 139-154.

Bednarek, M. (2010). *The language of fictional television: Drama and identity*. London & New York: Continuum.

Bednarek, M. (2018). *Language and Television Series. A Linguistic Approach to TV Dialogue.* Cambridge: CUP.

Bubel, C. (2006). *The linguistic construction of character relations in TV drama: Doing friendship in* Sex and the City. Tesis doctoral, Universität des Saarlandes (Alemania). http://scidok.sulb.uni-saarland.de/volltexte/2006/598/ [Acceso 21/3/2017].

Culpeper, J. (2001). *Language and characterisation: People in plays and other texts*. London: Longman.

Fairclough, N. (2008). El análisis crítico del discurso y la mercantilización del discurso público: Las universidades. *Discurso & Sociedad*, 2 (1): 170-185.

Kaplan, N. (2004). Nuevos desarrollos en el estudio de la evaluación en el lenguaje: la Teoría de la Valoración. *Boletín de Lingüística*, 22: 52-78.

Kozloff, S. (2000). *Overhearing Film Dialogue*. Berkeley-Los Angeles: University of California Press.

Lakoff. R. & Tannen, D. (1994). Conversational strategy and metastrategy in a pragmatic theory: The example of *Scenes from a marriage*. En Tannen, D. (Ed.) *Gender and discourse.* Oxford: OUP: 137-173.

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

Richardson, K. (2010). *Television dramatic dialogue: A sociolinguistic study*. Oxford: OUP.

El problema de la anonimización y decisiones adoptadas en la transcripción y revisión de corpus orales: el caso del corpus Val.Es.Co 2.0

Tomás Córcoles Molina

Departamento de Filología Española, Universitat de València, España

***Keywords**: anonimización; corpus Val.Es.Co; transcripción; conversación coloquial; protocolo; sistematización*

Abstract

En el ámbito de la Lingüística, en el proceso de recogida y transcripción de un corpus oral, son varios los problemas que van surgiendo y a los que debe darse respuesta. Uno de estos aspectos es la confidencialidad de los hablantes que intervienen en los discursos recopilados. Desde los años sesenta, la creación de corpus orales ha experimentado un notable crecimiento. En concreto, en el campo de la lingüística hispánica destacan, entre otros, los siguientes: ESLORA, COLAm, COSER, PRESEEA (Briz y Albelda 2009).

El grupo Val.Es.Co., también recopiló, a lo largo de los años noventa, un corpus oral de conversaciones coloquiales (Briz y Grupo Val.Es.Co. 2002); en este, hallamos una sólida toma de conciencia en cuanto a la protección de datos, puesto que los investigadores responsables pedían un documento de autorización –previo o posterior– a los hablantes de las conversaciones para mostrar su conocimiento y conformidad con la grabación. No obstante, con respecto a la ocultación de los nombres propios de persona, no existen criterios sistemáticos en el corpus Val.Es.Co. o en el corpus digital Val.Es.Co. 2.0 (Cabedo y Pons 2013). Muestra de ellos son las diferentes actitudes que los transcriptores ofrecen: utilización del símbolo de asterisco, (P****, L****), sustitución por otro nombre diferente sin ningún criterio establecido (María en lugar de Pepa) o transcripción del nombre real.

Actualmente en este grupo y, concretamente, en el marco del Proyecto UDEMADIS (FF120216-77841-P), se están realizando diversos trabajos de revisión y ampliación de los corpus publicados y anteriormente citados. En dichas experiencias previas, pese a las autorizaciones solicitadas por cuestiones de moralidad y rigor profesionales, no existía una uniformidad en el proceso. En cambio, la mayor relevancia de las leyes de

protección de datos ha obligado a los investigadores a sistematizar la anonimización de nombres personales y de lugares. Debido a esto, el trabajo de revisión que se están llevando a cabo se centra en dos vertientes: por un lado, la revisión de las conversaciones antiguas de lso corpus y la introducción de etiquetas y nuevas normas permitirá futuras investigaciones sin perjuicio de la identidad de los hablantes; por otro, la consolidación de este protocolo de actuación de cara a nuevas conversaciones desechará el proceso de revisión. Esto último se enfocará de cara a la formación de nuevos investigadores en el ámbito docente universitario.

En esta comunicación presentaremos de forma sistemática los problemas a los que nos enfrentamos en la actual fase revisión del corpus Val.Es.Co. 2.0 con respecto a la anonimización, en dos planos: en el texto, hemos de considerar qué elementos deben ocultarse, en qué grado, de qué forma y por qué nombre deben cambiarse; en el audio, hemos de saber qué herramienta informática utilizar y qué pasos seguir. Planteados los

problemas, ofreceremos el protocolo de decisiones que hemos configurado: para el texto, los nombres han de cumplir algunos requisitos (misma letra, número de sílabas y estructura acentual) y hemos de añadir una etiqueta de anónimo; para el audio, tras marcar los nombres anonimizados, emplearemos el programa Audicity con el fin de silenciar los fragmentos de audio en los que aparecen dichos nombres. Los resultados de este proceso de decisiones aportarán datos relevantes en el campo de trabajo de creación y revisión de corpus en particular y arrojarán luz sobre los problemas metodológicos en general a la que este tipo de investigación se enfrenta.

References

Briz, A. y Grupo Val.Es.Co. (2002). *Corpus de español coloquial*, Madrid, Arco/Libros.

Briz, A. y Albelda M. (2009). «Estado actual de los corpus de lengua española hablada y escrita: I+D», El español en el mundo. Anuario del Instituto Cervantes 2009. Disponible en:

http://cvc.cervantes.es/lengua/anuario/anuario_09/briz_albeida/p01.htm

Cabedo, A. y Pons, S. (eds.) (2002). *Corpus Val.Es.Co 2.0*. Disponible en http://www.valesco.es.

PRESEEA (2014): Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. Alcalá de Henares: Universidad de Alcalá. Disponible en: <http://preseea.linguas.net>.

Recalde, M. y Vázquez, M.ª V. (2009). «Problemas metodológicos en la formación de corpus orales», en P. Cantos Gómez y Sánchez Pérez, A. (eds.), *A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus*. Murcia: AELINCO, 37-49. Disponible en: http://www.um.es/lacell/aelinco/contenido/titulos.html.

Vázquez, M.ª V. (2014). «ESLORA: Diseño, codificación y explotación de un corpus oral de español de Galicia», II Workshop de Procesamiento Automatizado de Texto y Corpus (WOPATEC-2014), Pontificia Universidad Católica de Valparaíso, Viña del Mar.

“Deber (de) + infinitivo” en el CAES, un corpus escrito de aprendices de español como L2

Cecilia Criado de Diego[a]; Alicia San Mateo Valdehíta[b]

[a b]*Departamento de Lengua Española y Lingüística General, Universidad Nacional de Educación a Distancia, Spain*

Keywords: *corpus de aprendices; perífrasis modales, deber (de) + infinitivo; modalidad; variación en español como L2*

Abstract

Es conocida la variación que las perífrasis modales con *deber* presentan en el español de los hablantes nativos. En la lengua culta de España y América encontramos desviaciones del uso recomendado por las Academias de la Lengua (RAE y ASALE 2005, 2009, 2013) para expresar conjetura y obligación. Como diversos estudios de corpus constatan (Blas Arroyo 2011, 2013; Blas Arroyo y Vellón Lahoz 2014; Gómez Molina 2013, entre otros) y las Academias reconocen (RAE 1973: 448; RAE y ASALE 2009: 2144), la variante preposicional se manifiesta para expresar obligación —modalidad deóntica—, al igual que *deber* + infinitivo sin preposición para indicar modalidad epistémica —es decir, conjetura— es común en la lengua de la comunidad hispánica.

Nuestro objetivo es conocer cómo experimentan este fenómeno de variación los aprendices de español como L2. Desde que Chomsky (1986) establece la teoría de los principios y parámetros, algunos estudiosos de la adquisición de la L2, como Haegeman (1988) o Lardiere (2008), han buscado explicaciones del error a partir de este marco generativo. Estos investigadores postulan que el error es una incapacidad para reajustar parámetros o reasignarles valores (Haegeman 1988), y que esta dificultad aumenta cuando los fenómenos objeto de aprendizaje dependen de variables semántico- pragmáticas (Lardiere 2008).

Para conocer en qué medida los alumnos se desvían de las recomendaciones académicas de un hecho lingüístico dependiente de variables semánticas, como es el uso de *deber* o *deber de* para indicar obligación o conjetura, acudimos al *Corpus de aprendices de español como lengua extranjera (CAES)* —compuesto por 575 000 elementos lingüísticos—, y analizamos las perífrasis con *deber* en relación con las siguientes variables: (1) variante preposicional o no preposicional; (2) modalidad epistémica o deóntica; (3) lengua materna (árabe, chino mandarín, francés, inglés, portugués, ruso); (4) nivel de dominio (A1-C1). En total, en el corpus de textos escritos de aprendices de español como L2 se registran 368 casos de la perífrasis con *deber*.

Los resultados muestran que la variante prepositiva es menos común entre los aprendices (3 %) que entre los nativos (11,7 %: Autor, en revisión); de igual modo, la modalidad epistémica entre estudiantes de español se manifiesta con la perífrasis con *deber* en pocas ocasiones (2,4 %) en comparación con el porcentaje hallado entre los hablantes nativos (28,1 %: Autor, en revisión). En cuanto a los usos no recomendados por las Academias, al igual que sucede en las producciones de nativos, en la modalidad epistémica la variante no canónica, es decir, la no prepositiva, es la más frecuente (88,9

%), mientras que en la modalidad deóntica la perífrasis con preposición, incorrecta según la normativa académica, es muy esporádica (2,8 %) en los textos de los aprendices.

Según estos datos, si tenemos en cuenta el factor de la lengua materna o el de nivel de dominio lingüístico, no existen diferencias significativas entre los distintos aprendices.

De cualquier manera, el número de ocurrencias obtenidas de la variante preposicional en la modalidad epistémica impide llegar a conclusiones extrapolables a la realidad lingüística de los hablantes de español como L2.

Referencias

Blas Arroyo, J. L. (2011). *Deber (de)* + infinitivo: ¿un caso de variación libre en español? Factores determinantes en un fenómeno de alternancia sintáctica. *Revista de Filología Española,* 91(1), 9-42.

Blas Arroyo, J. L. (2013). Norma y uso en un fenómeno de variación sintáctica. Nuevos datos a propósito de la oposición modal epistémico-deóntica y la variable *deber/deber de* + infinitivo. In *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas* (pp. 519-529). En E. Casanova & C. Calvo. Berlín, Germany: E. de Gruyter.

Blas Arroyo, J. L. & Vellón Lahoz, J. (2014). La supervivencia de *deber de* + infinitivo en el español moderno (ss. XVIII-XX). *Revista de Filología Española,* 94(1), 9-38.

Chomsky, Noam. (1986). *Knowledge of language: Its nature, origins, and use*. Nueva York: Praeger.

Criado de Diego, C., & San Mateo Valdehíta, A. (en prensa). Las perífrasis modales con *deber* en el español actual. Estudio de corpus. *Spanish in Context (SiC).*

Gómez Molina, J. R. (2013). Norma y usos de las perífrasis "deber + infinitivo" /"deber de + infinitivo. In J. R. Gómez Molina (Coord.), *El español de Valencia: estudio sociolingüístico* (pp. 71-108). Bern, Switzerland: Peter Lang.

Haegeman, L. (1988). The categorial status of modals and L2 acquisition. In S. Flynn & w. O´Neil (Coords.), *Linguistic theory in second language acquisition*. Boston, United States: Kluwer.

Lardiere, D. (2008). Feature-Assembly in Second Language Acquisition. In J. Liceras, H. Zobl & H. Goodluck (Coord.), *The role of formal features in second language acquisition*. New York, United States: Lawrence Earlbaum Associates.

Real Academia Española (RAE). Corpus del Español del Siglo XXI (CORPES). Retrieved from http://www.rae.es

Real Academia de la Lengua Española (RAE) (1973). *Esbozo de una nueva gramática de la lengua española*. Madrid, Spain: Espasa.

Real Academia de la Lengua Española (RAE) & Asociación de Academias de la Lengua Española (ASELE) (2005). *Diccionario panhispánico de dudas*. Madrid, Spain: Santillana.

Real Academia de la Lengua Española (RAE) & Asociación de Academias de la Lengua Española (ASELE) (2009). *Nueva Gramática de la lengua española.* Madrid, Spain: Espasa Calpe.

Real Academia de la Lengua Española (RAE) & Asociación de Academias de la Lengua Española (ASELE) (2013). *El buen uso del español*. Madrid, Spain: Espasa.

Rojo, G. & Palacios, I (Dir.). *Corpus de aprendices de español como lengua extranjera (CAES).* Versión: 1.2, agosto 2018. Instituto Cervantes.

<http://galvan.usc.es/caes>.

A synchronic study of tag questions in East African English

Miriam Criado-Peña

Department of English Philology, University of Málaga, Spain

***Keywords**: Tag questions, East African English, Kenyan English, Tanzanian English*

Abstract

Tag questions in standard British English (BrE) follow a standard pattern consisting of an operator and a subject. This operator generally coincides with the preceding statement and rules of polarity are usually obeyed: negative tags are attached to positive statements and vice versa (i.e. he likes you, doesn't he?) (Quirk et al. 1985: 810). The new varieties of English seem to play an essential role in the development of this construction insofar as apparently no standard rule is observed. Some of these varieties have received special attention in the literature, Asian Englishes, in particular (Wong 2007; Columbus 2009; Columbus 2010; Takahashi 2014; Criado-Peña 2016), whilst the use of the construction in the African varieties of English has been virtually neglected. Tag questions are still unexplored in East African English (EAfE) in spite of its linguistic distinctiveness and diversity. EAfE is composed of three varieties: Kenyan English (KenE), Tanzanian English (TznE), and Ugandan English (UgE), each of them having a different historical and sociolinguistic background, and thus making use of the language in different ways.

Given the linguistic richness of EAfE, the present paper investigates the use and distribution of tag questions in KenE and TznE. These varieties are studied due to the high degree of attestation and pervasiveness of the phenomenon according to *eWave* (Kortmann and Lunkenheimer 2013), labeled as "pervasive or obligatory" in KenE and "neither pervasive nor extremely rare" in TznE. Their analysis is also interesting insofar as they are sometimes placed at different stages in the evolutionary process of New Englishes. According to Kachru's Concentric Circles model (1986), the English of Kenya belongs to the "Outer circle" whereas the case of TznE is not clear-cut, mostly regarded as part of the "Outer circle", although sometimes considered to be part of the "Expanding Circle". On the other hand, if Schneider's Dynamic Model is considered, both varieties are placed at the same stage in the process of formation of the Postcolonial varieties of English: phase 3 (nativization), regarded as "the most vibrant one, the central phase of both cultural and linguistic transformation" (2007: 41). In the light of this, the present paper has been conceived with the following objectives: a) to analyze the distribution of

tag questions in EAfE and compare their use and features across the two varieties (i.e. KenE and TznE); b) to assess the frequency of the phenomenon across speech and writing, and across text types; and c) to evaluate the sociolinguistic variation in terms of age and gender. For the purpose, the East African English component of the *International Corpus of English* (ICE-EA) is used as source of analysis. Preliminary results point to a higher use of tag questions in KenE as well as a tendency towards the use of invariant tags in both varieties.

References

Columbus, G. (2009). A corpus-based analysis of invariant tags in five varieties of English. In A. Renouf & A. Kehoe (Eds.), *Corpus linguistics: Refinements and reassessments* (pp. 401-414). Amsterdam, Netherlands: Rodopi.

Columbus, G. (2010). A comparative analysis of invariant tags in three varieties of English. *English World-Wide*, 31(3), 288-310.

Criado-Peña, M. (2016). *'They are going tomorrow, isn't it?'* On the use of tag questions in Indian English and Hong Kong English. In A. Moreno-Ortiz & C. Pérez- Hernández (Eds.), *Proceedings of the 8th International Conference on Corpus Linguistics* (pp. 71-78). EPiC Series in Language and Linguistics.

Kachru, B. (1986). *The alchemy of English: The spread, functions and models of non-Native Englishes*. Oxford, United Kingdom: Pergamon Institute of English.

Kortmann, B. & Lunkenheimer, K. (2013). The electronic world atlas of varieties of English. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://ewave-atlas.org).

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

Schneider, E. (2007). *Postcolonial English. Varieties around the world*. Cambridge, United Kingdom: Cambridge University Press.

Takahashi, M. (2014). A comparative study of tag questions in four Asian Englishes from a corpus-based approach. *Asian Englishes*, 16(2). 101-124.

Wong, M. (2007). Tag questions in Hong Kong English: A corpus-based study. *Asian Englishes*, 10(1). 44-61.

Authorship Attribution in Medieval Private Correspondence: Theoretical and Methodological Implications for Historical Sociolinguistics and Forensic Linguistics

Juan A. Cutillas-Espinosa and Juan M. Hernández-Campoy

University of Murcia

Keywords: forensic linguistics, historical sociolinguistics, authorship, holograph/autograph letters, scribal/secretarial letters

Abstract

The development and diversification of archival data sources is allowing scholars to explore the role of new genres and text-types as adequate materials for historical sociolinguistic analysis: ego-documents, such as diaries, travel accounts, court records, recipes, and especially letters, are now seen as essential documents for research in this field at diastratic, diatopic and diaphasic levels (see Tieken-Boon van Ostade 2005, 2006; Nevala & Palander-Collin 2005; Schiegg 2016; or Voeste 2018, among many others). Some monographs have also confirmed the relevance of these documents to reconstruct the sociolinguistic contexts of language variation and change in the past (see Nevalinen & Tanskanen 2007; Dossena & Tieken-Boon van Ostade 2008; Sairio 2017; Dossena & Del Lungo Camiciotti 2012; or Auer, Schreier & Watts 2015). Unlike other surviving documents from the past, corpora of historical private and official correspondence and their substantial social metadata from the past is offering a very useful archival source to carry out studies in Historical Sociolinguistics, favouring the interest in linguistic variation given the different socio-demographic and geographical characteristics of both writers and recipients (see Nevalainen & Raumolin-Brunberg 1996; 2003; Hernández-Campoy & Conde-Silvestre 2012; Säily *et al* 2017).

However, authorship constitutes one of the most controversial issues when doing socio-historical research on the behaviour of linguistic forms. An important problem related to the authenticity and purity in the transmission of manuscripts is authorship and, hence, gender. Letters were often not autographed but rather written from dictation by a personal scribe of the author due to the widespread illiteracy that was characteristic of early historical periods. Illiteracy among female population in medieval times and the subsequent use of scribes make authorship and gender constitute some of the most controversial socio-demographic issues when doing sociohistorical research, since it might easily corrupt the representativeness and validity of any empirical research carried out ignoring this usual practice (Bergs 2015).

Forensic Linguistics appeared as a branch of Applied Linguistics to assist the law in legal processes, where authorship elucidation is often one of the most disputed questions. This interest in the discipline has also been reflected in the publication of introductory books and handbooks devoted to this subject (Turell 2005; Olsson 2004; or Coulthard & Johnson 2007). While admitting that it is possible to identify a particular speaker/writer by their linguistic choices, forensic linguists have disagreed about what particular methodology should be applied: Forensic Idiolectology (research-driven), Forensic Stylistics (data-driven) and Forensic Stylometrics (research-driven and statistics-based).

With this in mind, the object of this paper is twofold. First of all, we aim to present an example of historical epistolary documents where forensic linguistic techniques may help to elucidate authorship: letters authored Margery Paston, putting their theorical tenets and techniques to the test of time. Secondly, we will try to show how this set of texts offers a valuable perspective on the different approaches to forensic authorship identification. The data suggests that formal (spelling) features are less indicative of authorship than morpho-syntactic markers. Forensic Linguistics and

Historical Sociolinguistics can mutually benefit each other, by sharing their expertise in authorship research and its application to current and historical texts in their social context.

References

Auer, A.D., Schreier; D. & Watts, R. (eds.) (2015). *Letter Writing and Language Change*. Cambridge: Cambridge University Press.

Bergs, A. (2015). Linguistic fingerprints of authors and scribes. In: A. Auer, D. Schreier & D. Watts (eds.), 114-132.

Coulthard, M. & Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language in Evidence*. Oxon: Routledge.

Dossena, Marina & Del Lungo Camiciotti, G. (eds.) (2012). *Letter Writing in Late Modern Europe*. Amsterdam/Philadelphia: John Benjamins.

Dossena, Marina & Tieken-Boon van Ostade, I. (eds.) (2008). *Studies in Late Modern English correspondence*: *Methodology and Data*. Bern: Peter Lang.

Hernández-Campoy, J.M. & Conde-Silvestre, J.C. (eds.) (2012). *The Handbook of Historical Sociolinguistics*. Malden: Wiley-Blackwell.

Nevala, M. & Palander-Collin, M. (2005). Letters and letter writing: Introduction. *European Journal of English Studies* 9(1), 1-7.

Nevalainen, T. & Raumolin-Brunberg, H. (eds.) (1996). *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi.

Nevalainen, T. & Raumolin-Brunberg, H. (2003). *Historical Sociolinguistics. Language Change in Tudor and Stuart England*. London: Longman Pearson Education.

Nevalainen, T. & Tanskanen, S.K. (eds.) (2007). *Letter Writing*. Amsterdam/Philadelphia: John Benjamins.

Olsson, J. (2004). *Forensic Linguistics*. New York: Continuum.

Säily, T., Nurmi, A., Palander-Collin, M. & Auer, A. (eds.) (2017). *Exploring Future Paths for Historical Sociolinguistics*. Amsterdam/Philadelphia: John Benjamins.

Sairio, A. (2017). *Bluestocking Corpus: Letters of Elizabeth Montagu, 1730s-1780s*. Helsinki: University of Helsinki (http://bluestocking.ling.helsinki.fi/).

Schiegg, M. (2016). Code-Switching in Lower-Class Writing: Autobiographies by Patients from Southern German Psychiatric Hospitals (1852-1931). *Journal of Historical Sociolinguistics* 2(1), 47-81.

Tieken-Boon van Ostade, I. (2005). Of social networks and linguistic influence: the language of Robert Lowth and his correspondents. *International Journal of English Studies* 5(1), 135-157.

Tieken-Boon van Ostade, I. (2006). Eighteenth-century English letters: In search of the vernacular. *Linguistica e Filologia* 21, 113-146.

Turell, M.T. (ed.) (2005). *Lingüística Forense, Lengua y Derecho. Conceptos, Métodos y Aplicaciones*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Voeste, A. (2018). The self as a source. A peasant farmer's letters from prison (1848-1852). *Journal of Historical Sociolinguistics* 4(1), 97-118.

Billions of words for the price of… lack of diversity: Comparing web-crawled and traditional corpora

Václav Cvrček[a]; David Lukeš[a]

[a]*Institute of the Czech National Corpus, Charles University, Czech Republic*

***Keywords**: web-crawled corpora, representativeness, multi-dimensional analysis, register, Czech*

Abstract

Growing volumes of textual data freely available on the internet gave rise to web-crawled corpora (Baroni, Kilgarriff, Pomikálek, & Rychlý, 2006), which are often claimed to represent not only the language of the web but also the general linguistic patterns present in off-line language. While the advantages of web-as-corpus approach are obvious (low price, high volumes), their limitations (namely lack of metadata, opportunistic design) raise questions regarding their usability for general linguistic research – especially as to what extent a web-crawled corpus can represent offline language. While there are some promising results regarding the register classification of texts on the "searchable" web (Biber & Egbert, 2016; Sharoff, 2018), we can conclude that it is still a rather under-researched topic and the exact composition of web-crawled corpora remains unknown to a large extent.

In this paper, we compare a web-based corpus of Czech, Araneum Bohemicum (Benko, 2014, 2016), to Koditex (Zasina, Lukeš, Komrsková, Poukarová, & Řehořková, 2018, see also http://wiki.korpus.cz/doku.php/en:cnk:koditex), a traditional corpus representative of a wide range of Czech registers (written, spoken and web). While the former, with its opportunistic design, can be considered representative of the Czech "searchable" Internet, the latter is a mixture of (mainly offline language) corpora carefully designed with an emphasis on a highly diverse composition. Since it is impossible to directly estimate the relation between each of these corpora and the population they strive to represent (or the population of language as a whole), we should adjust the research question accordingly: our research question is not "which one is a better approximation of the population," but rather "what is their mutual overlap in terms of variation covered", and consequently "can web-crawled corpora fully replace traditional ones".

In order to assess the mutual replaceability of these corpora, we will use the multi-dimensional analysis framework (Biber, 1988, 1995), allowing us to discern which registers can be found in both sources and which are exclusive to one of them. As a highly representative and diverse corpus by design, Koditex was used to derive an 8-

dimensional model of Czech register variation (Cvrček et al., 2018) based on 122 linguistic features. The Araneum data (11 samples of randomly selected chunks of text) were then projected onto this model. The comparison of the space of variation covered by both corpora employs primarily the first two dimensions (1st dimension: dynamic versus static texts, 2nd dimension: spontaneous versus prepared texts), which explain the largest portion of variation.

In contrast to lexically driven methods of comparing corpora (Kilgarriff, 2001, 2012; Piperski, 2018), we present a comparison based on features from all linguistic levels. The comparison reveals that the parts of the multi-dimensional space of register variation which are covered by texts from both the Koditex and Araneum corpora are mainly newspapers and magazines, and non-fiction written texts in general. It is not surprising that these text classes are easily accessible on the web as well as in the offline domain, and that they are therefore comparably well represented in both corpora. On the other hand, areas covered by some parts of fiction and especially informal spoken conversation are exclusive to Koditex, which suggests that these text classes (or at least linguistically equivalent ones) cannot be easily obtained by web-crawling.

References

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: a web tool for instant corpora. In *Proceeding of the EuraLex Conference* (pp. 123–132).

Benko, V. (2014). Aranea: Yet another family of (comparable) web corpora. In *International Conference on Text, Speech, and Dialogue* (pp. 257–264). Springer.

Benko, V. (2016). Two Years of Aranea: Increasing Counts and Tuning the Pipeline. In *LREC* (pp. 4245–4248).

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge, England: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge, England: Cambridge University Press.

Biber, D., & Egbert, J. (2016). Register Variation on the Searchable Web: A Multi-Dimensional Analysis. *Journal of English Linguistics*, *44*(2), 95–137. https://doi.org/10.1177/0075424216628955

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2018). From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2018-0020

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, *6*(1), 97–133.

Kilgarriff, A. (2012). Getting to know your corpus. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech and Dialogue* (pp. 3–15). Springer.

Piperski, A. (2018). Corpus Size and the Robustness of Measures of Corpus Distance. In *Computational Linguistics and Intellectual Technologies* (p. 11). Moscow. Retrieved from http://www.dialog-21.ru/media/4327/piperskiach.pdf

Sharoff, S. (2018). Functional Text Dimensions for the annotation of web corpora. *Corpora*, *13*(1), 65–95. https://doi.org/10.3366/cor.2018.0136

Zasina, A. J., Lukeš, D., Komrsková, Z., Poukarová, P., & Řehořková, A. (2018). Koditex: corpus of diversified texts (Version 1). Prague: Institute of the Czech National Corpus. FF UK. Retrieved from www.korpus.cz

Migrants over time: a corpus-assisted newspaper discourse analysis of the metaphorical representation of migration. The case of two New York newspapers between 20th and 21st Century

Dario Del Fante[a]

[a]*Department of Literary and Linguistic Science, University of Padova, Italy; School of English, University of Sussex, United Kingdom*

***Keywords**: Metaphor, Migration, Newspaper discourse, Language change*

Abstract

The proposed research study is focused to two essential qualities of humanity: migratory movements and metaphors. Migratory movements can be undoubtedly defined as an enduring characteristic of our global history, an inherent feature of humanity which have showed ever-changing properties during history. Metaphor is a resource through which people "may enhance the expressiveness of their message through the most economical means available to them" (Charteris-Black 2004:17) and organize their experience of life (Lakoff and Johnson 1980). The majority of works concerning the metaphorical representation of migrants (Santa Ana 1999; El Rafaie 2001; Semino 2008; Dervinytė 2009; Hart 2011, 2014) have shown how newspaper discourse and public discourse in general have been giving a predominantly negative representation of immigrants\immigration through the selection of specific conventional metaphors. For instance, they have been represented as "ANIMALS", depicting them as non-humans, not just non-citizens. Other studies have reported the use of natural disaster metaphors, in particular "WATER" metaphors realised with expressions like "flood of" or "tides of". The only remarkable exception is represented by Salahshour (2016) who shows that in New Zealand immigrants are positive represented as "force" which is "gained" by the country. However, there is a lack of studies which take into account not only the contemporary migration discourse. With an understanding of how metaphors originate and are subsequently taken up and conventionalised over time, it would be easier to shed light on the cyclical process through which metaphors work and on how some metaphor "die" and become fossilized, losing its metaphorical effect (Lakoff 1987, Goatly 1997). Therefore, I

intend to address this problem by embarking a case study which would have a diachronic perspective based on empirical evidence provided by the corpus analysis. Newspapers are important influencer of people's perspective of reality and opinions (Partington et al., 2013), and by analysing the metaphorical representation of migrants in newspapers we might define, to some extent, how migrants are experienced within a specific group of people. This study, in line with SFL (Halliday and Matthiessen 2004), falls in the tradition of Corpus Assisted Discourse Study (Partington et al. 2013), which combines the quantitative analysis techniques of Corpus Linguistics, which are characterized by statistical reliability, with the qualitative approach of (Critical) Discourse Analysis. The data set consist of two newspaper corpora for two different time

periods both representing significant moments for the migratory movements history (Daniels 2002, Migration Policy Institute 2016[1]):

- "The New York Tribune" 1900-1905, downloaded from the database of the "Chronicling America" project at the Library of Congress.[2]
- "The New York Times" 2010-2015 downloaded from the database Lexis Nexis.

A corpus-driven approach has been adopted to retrieve all examples of metaphor surrounding discussion of migration (Steen et al. 2010), which have been subsequently validated through a qualitative analysis. The results from the two dataframe have been compared in terms of differences and similarities over time.

Preliminary results show that migration, for the period 1900-1905, was represented in terms of quantity through liquid metaphors such as "tide of", "stream of", "wave of", "flood of", which, nevertheless, show a low frequency of usage. It would be possible that this type of metaphor firstly appeared during this period and then has been conventionalized, losing its novelty as creative metaphor (Charteris-Black). In fact, the same expressions have a higher frequency of use in the contemporary corpus, which would correspond to its conventionalization. On the contrary, it seems that there are not particular metaphorical use in the past which have not anymore been used in the contemporary corpus. New metaphors have not emerged from the 2010-2015 data.

References

Charteris-Black, J. (2004). *Corpus approaches to critical metaphor analysis*. Berlin, Germany: Springer.

Lakoff, George, and Mark Johnson. (1980). *Metaphors we live by.* Chicago, United States: University of Chicago Press.

Ana, O. S. (1999). Like an animal I was treated': Anti-immigrant metaphor in US public discourse. *Discourse & society*, *10*(2), 191-224.

El Refaie, E. (2001). Metaphors we discriminate by: Naturalized themes in Austrian newspaper articles about asylum seekers. *Journal of Sociolinguistics*, *5*(3), 352-371.

Semino, E. (2008). *Metaphor in discourse*. Cambridge, United Kingdom: Cambridge University Press.

Dervinyte, I. (2009). Conceptual emigration and immigration metaphors in the language of the press: A contrastive analysis. *Studies about Languages*, *14*, 49-

55.

Hart, C. (2011). Force-interactive patterns in immigration discourse: A Cognitive Linguistic approach to CDA. *Discourse & Society*, *22*(3), 269-286.

Hart, C. (2014). *Discourse, grammar and ideology: Functional and cognitive perspectives*. London, United Kingdom: Bloomsbury Publishing.

Lakoff, G. (1987). The death of dead metaphor. *Metaphor and symbol*, *2*(2), 143-147.

Goatly, A. (1997). *The Language of Metaphors.* London, United Kingdom: Routledge.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)*(Vol. 55). Amsterdam, Netherlands: John Benjamins Publishing.

Halliday, M. A. K., & Matthiessen, C. (2004). *An introduction to functional grammar*.

London, United Kingdom: Arnold.

Daniels, R. (2002) *Coming to America. A History of immigration and ethnicity in American life.* New York, United States: Perennial.

1 Migration Policy Institute (2016). *Tabulation of data from U.S. Census Bureau, 2010-2016 American Community Surveys (ACS), and 1970, 1990, and 2000 Decennial Census*. Available at <https://www.migrationpolicy.org/programs/data- hub/charts/immigrant-population-over- time?width=1000&height=850&iframe=true> (Last visited at 08th December 2018)

2 Available at <http://chroniclingamerica.loc.gov/> . (Last access 24th November 2018)

Steen, G. (Ed.). (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam, Netherlands: John Benjamins Publishing.

Salahshour, Neda. (2016) "Liquid metaphors as positive evaluations: A corpus-assisted discourse analysis of the representation of migrants in a daily New Zealand newspaper." Discourse, Context & Media 13, 73-81.

[1] Migration Policy Institute (MPI) tabulation of data from U.S. Census Bureau, 2010-2016 American Community Surveys (ACS), and 1970, 1990, and 2000 Decennial Census.

Verb or copula? Sentence classification with neural network

Andrea Dömötör

Pázmány Péter Catholic University, Hungary

MTA-PPKE Natural Language Processing Research Group

Keywords: *sentence parsing, word sense disambiguation, neural network, word embedding, machine learning*

Abstract

This research is an experiment with the aim to find a machine learning based solution to a word sense disambiguation problem in Hungarian. The problem in the focus of the study is the multifunctionality of the word *"volt"* (*be/have*) which can either be a lexical verb used in locative and possessive sentences (Examples 1a, 1b) or a copula in case of nominal predicates (Example 1c). The main difference between these two functions is that the copula is ommitted in present tense 3rd person while the locative and possessive verb has its present form *"van"*. This means that in dependency parsing the latter should be considered the head of the sentence while the former is a complement of the predicative nominal, according to the annotation guidelines of Universal Dependencies (Nivre 2014).

(1) a, Feri otthon **volt** / van.
Fernando at_home be-PST-Sg3 be-Sg3

'Fernando was/is at home,'

b, Ferinek **volt** / van egy szalamandrája.
Fernando be-PST-Sg3 be-Sg3 a salamander-Poss.Sg3

'Fernando had/has a salamander.'

c, Feri nagyon szerény **volt** / 0.
Fernando very humble be-PST-Sg3

'Fernando was/is very humble.'

The study aims to solve this problem using word embedding and neural network based text classification. For word representation I used a Word2vec model (Mikolov et. al. 2013) which represents the words as vectors such that words in similar contexts in the training corpus are located close to each other in the vector space.

The classification of sentences is also based on a neural network and therefore needs a great amount of labelled data for training. As manual labelling is time-consuming, I used an English-Hungarian lemmatized, morphologically analyzed and disambiguated, word-aligned corpus (Novák et al., 2019) to annotate sentences automatically. First, I extracted the sentences that contained a form of the target word ("*volt"*) on the Hungarian side. These sentences were labelled according to the syntactic role (copula or lexical verb) of "*volt*" with a rule-based algorithm implemented in Python3. The labelling algorithm makes the decision based on the English words that are aligned to "*volt"* or that are following a form of "*be".* The programme's output was 791130 labelled sentences, with around 87% accuracy. As this performance is not high enough to use the retrieved data for training, the sentence classifier was tested with a manually revised sample of 1000 sentences.

The input layer of the classifier's network takes the sentence vectors (the concatenations of their word vectors) and the output is the corresponding label (verb or copula). The network was trained with 100000 iterations. The initial results demonstrated that with this small training data set the system still showed some progress in learning: in the first 1000 training steps the average accuracy was close to random (56,56%), while in the last 1000 iterations it reached an average of 69,38%.

In sum, the neural network based classification method can be useful for automatic sentence parsing. The experiments described in this proposal will be proceeded with the improvement of the quality of the training data.

References

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space.* In: Proceedings of Workshop at ICLR, Scottsdale, AZ, USA.

Nivre, J. (2014). Nonverbal Predication and Copulas in UD v2. http://universaldependencies.org/v2/copula.html. Accessed: 2019-03-27

Novák, A., Laki, L. J. and Novák, B. (2019). Mit hozott édesapám? Döntést – Idiomatikus és félig kompozicionális magyar igei szerkezetek azonosítása párhuzamos korpuszból [identification of Hungarian idiomatic and light verb constructions from a parallel corpus]. In *XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019)* [15th Hungarian Conference on Computational Linguistics] (pp. 63-71). Szeged. Szeged University.

Tokenizing and Part-of-Speech Tagging multilingual corpora

Irene Doval Reixa

Univerisidade de Santiago de Compostela, Spain

Keywords: *multilingual corpora; parallel corpora; tokenization; pos-tagging; tagsets*

Abstract

In the last decades, parallel and comparable corpora (bilingual and multilingual) have become a fundamental resource for cross linguistic studies and for a wide range of natural language processing tasks. Naturally, these potential applications depend, in practice, on the annotation layers of each corpus. Specifically lemmatization and part-of-speech tagging enable searching for specific phenomena that cannot be retrieved by the simple string search.

Although pos-tagging is a common task in the creation of a corpus, when dealing with multilingual data, in which multiple languages come into play, this task involves additional challenges. In this paper, we review the different challenges faced on tokenizing and POS tagging a bilingual parallel corpus, the PaGeS Corpus, consisting

of

25 million words of German and Spanish fictional texts. On the one hand, the performance of the taggers differs considerably from one language to another so that a different tagger has had to be used for each language: for Spanish Freeling (Padró 2012) and for German Treetagger (Schmidt 1995). On the other hand each language, with its specific grammatical features and its own grammatical terminology, uses tagsets, that diverge in the scope and detail. This made necessary to develop a common tagset, mapped to the Universal Tagset format (Petrov et al., 2012), in order to facilitate the interoperability and the corpus consultation.

References

Schmid, Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf (19 September, 2017).

Petrov, Slav, Dipanjan Das and Ryan McDonald. 2012. A universal part-of-speech tagset. In Eighth International Conference on Language Resources and Evaluation (LREC 2012): 2089-2096.

Padró, Lluís and Stanilovsky, Evgeny. 2012. FreeLing 3.0: Towards Wider Multilinguality. Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. May, 2012. http://nlp.lsi.upc.edu/publications/papers/padro12.pdf

Corpus-Based Research: Persian Pro-Conservative Concepts and their Implications for Translation

Mina Ebrahimi Erdi

Department of Translation and Interpretation & East Asian Studies, Autonomous University of Barcelona, Spain

***Keywords**: Persian; political news; pro-conservative news agency; marked terms*

Abstract

The aim of this paper is communicating some preliminary results of the study of sensitive terms and concepts included in a corpus of political news extracted from a Persian pro-conservative news agency. Frequency lists, concordances and n-grams will be analysed using SketchEngine and implications for translation will be drawn.

Baroni (2009) highlights the importance of word frequency and considers it as an integral part ‘in all branches of corpus linguistics’ and asserts that it is ‘what

distinguishes corpus-based methodologies from other approaches to language' (p. 803). He also distinguishes between 'rank/frequency profile' and 'frequency spectrum' which is related to the present paper. The former is specifically functional for investigating the characteristics of high frequency terms and the latter is specifically helpful to explore the 'properties of low frequency items' (p. 806).

This paper is based on a corpus which was designed meticulously to be representative so that the results would be used for quantitative analyses and be generalizable to some extent, an issue highlighted by Biber & Jones (2009) among others. The sampling method was 'stratified' in which, at first, the focused text categories (political texts) were selected and then, texts related to that category (political news of one news agency broadly and clearly known as pro-conservative) were collected (ibid, p. 1288). As the focus is presidential election, they were filtered according to their relevance to this topic and also from one month before the election date, May 19th of 2017, during which campaigns and hot political debates were dominating the news agencies. The size of corpus is another important issue, which in case of this paper, it is almost one million words. It is worth mentioning that Persian is among less studied languages but is, according to Ostler (2008), 'probably upwardly mobile' (p. 458).

Some main super-concepts previously selected on the basis of their relevance in the political context under investigation (e.g. power, enemy, election) were studied using SketchEngine. Here are some of the results revealed in this study:

a) There were numerous lexical gaps due to different political system of Iran and they would cause troubles during translating from Persian to English such as *estekbar* (countries which are bigoted and don't follow righteousness), *nezam-eh solteh* (domineering countries which don't respect the autonomy of the others), *basirat* (a special consciousness to recognize enemy, even the national one, from friend or wrong from right), *kaftar baazi-ye siyasi* (political playing with pigeons: a metaphor used to show the disapproval of political manoeuvring) and etc. .

b) There were very important political terms among words with low frequency (2 to 5 times of occurrence); for instance: *nofouz* (penetration), *fetneh* (sedition), *tote'eh gar* (revolt) and so on which proved that low frequency terms should not be ignored at all.

c) N-grams depicted to be functionally crucial because of their information-loaded aspects which can be ignored if just word list or concordances of higher frequency were investigated.

d) Last but not least important point was usefulness of corpus linguistics (CL) tools and techniques for the study of less studies languages, on the one hand, and bringing deep-structured and hidden semantic networks to surface.

References

Adel, A. & Reppen, R. (Eds.) (2008). *Corpora and discourse: The challenges of different settings*. Amsterdam/Philadelphia: John Benjamins.

Baroni, M. (2009). Distributions in text. In Ludeling, A. & Kyto, M. (Eds.), *Corpus linguistics: An international handbook* (Vol. 2, pp. 803 – 822). New York, Berlin: Walter de Gruyter.

Baroni, M. & Evert, S. (2009). Statistical methods for corpus exploitation. In Ludeling,

A. & Kyto, M. (Eds.), *Corpus linguistics: An international handbook* (Vol. 2, pp. 777 – 803). New York, Berlin: Walter de Gruyter.

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

Biber, D. & Jones, J. K. (2009). Quantitative methods in corpus linguistics. In Ludeling, A. & Kyto, M. (Eds.), *Corpus linguistics: An international handbook* (Vol. 2, pp. 1286 – 1304). New York, Berlin: Walter de Gruyter.

Ludeling, A. & Kyto, M. (Eds.) (2008). *Corpus linguistics: An international handbook* (Vols. 1-2). New York, Berlin: Walter de Gruyter.

Ostler, N. (2008). Corpora of less studied languages. In Ludeling, A. & Kyto, M. (Eds.), *Corpus linguistics: An international handbook* (Vol. 1, pp. 457 – 483). New York, Berlin: Walter de Gruyter.

Wynne, M. (2008). Searching and concordancing. In Ludeling, A. & Kyto, M. (Eds.), *Corpus linguistics: An international handbook* (Vol. 1, pp. 706 – 737). New York, Berlin: Walter de Gruyter.

Problemas en la elaboración de un corpus oral de conversaciones coloquiales: el paso de lo analógico a lo digital en el corpus Val.Es.Co.

Guadalupe Espinosa Guerri

Universitat de València

Palabras clave: *corpus orales, problemas metodológicos, digitalización, transcripción, etiquetado, revisión*

Abstract

El grupo Val.Es.Co., publicó su corpus de conversaciones coloquiales en 2002. Este corpus en papel recogía un total de diecinueve conversaciones de hablantes de distinto nivel sociocultural y empleaba un sistema de transcripción inspirado en el propuesto por Jefferson (1984, 2014). Este sistema posee una versión ancha (Briz y Grupo Val.Es.Co., 2002: 29) y otra más estrecha (Hidalgo, 1997) adaptable a los distintos objetivos de investigación. Posteriormente, Cabedo y Pons (2013) ampliaron dicho corpus con cuarenta y seis conversaciones más, pero esta vez, en formato electrónico (Corpus Val.Es.Co. 2.0), aunque manteniendo el mismo sistema de transcripción creado para el corpus analógico.

Actualmente, dicho corpus digital se encuentra en un proceso de revisión que consta de cuatro fases: 1) **detección** de posibles errores de transcripción; 2) **adaptación** a un nuevo sistema de transcripción con etiquetas, el cual se basa en desarrollado para el corpus Ameresco (Albelda y Estellés), inspirado a su vez en el sistema empleado por el corpus PRESEEA (2014); 3) **alineación** del texto y del audio mediante el programa

ELAN (Nijmegen, 2018); y 4) **anonimización** del audio con el programa Audacity (Audacity Team, 2018). Esta revisión del corpus se lleva a cabo con una doble finalidad. Por un lado, mejorar la adaptación de las conversaciones ya recopiladas al formato digital, de manera que los investigadores puedan hacer búsquedas más precisas y trabajar con las conversaciones y el audio ya alineados. Por otro, asegurar la accesibilidad del corpus y sistematizar todas las fases del proceso de revisión para la futura ampliación del mismo.

En dicho proceso surgen una serie de dificultades de distinta índole, las cuales se exponen en el presente trabajo, junto con algunas de las soluciones que se han adoptado. Los problemas encontrados se pueden clasificar en cuatro tipos. En primer lugar, aquellos problemas relacionados con la recopilación de las conversaciones, es decir, la obtención de muestras de habla espontánea, la autorización de los participantes, el deseo de minimizar o eliminar los efectos de la paradoja del observador y la anonimización de los nombres propios y todos aquellos elementos que puedan desvelar la identidad de los participantes. Se trata de dificultades compartidas por todos aquellos investigadores que se enfrentan a la confección de un corpus de conversaciones orales espontáneas (Recalde y Vázquez, 2009) y que, debido a su complejidad, serán expuestas en un trabajo independiente. En segundo lugar, otro tipo de dificultades surge en el proceso de transcripción, concretamente, en la adaptación del sistema clásico de transcripción a un sistema de etiquetas apto para el formato electrónico y los objetivos de búsqueda; es decir, en el paso de lo analógico a lo digital. En tercer lugar, también se dan problemas de tipo técnico relacionados con la calidad de los audios, los ruidos de fondo, la distinción de voces, etc., los cuales nos hacen plantearnos futuras líneas de mejora del corpus mediante material multimodal que permita minimizar algunos de estos problemas. En cuarto y último lugar, el proceso de revisión y los problemas hallados en su desarrollo han puesto en evidencia la importancia de sistematizar cada una de las partes que constituyen dicho proceso y elaborar un protocolo con el objetivo de conseguir un corpus homogéneo.

Referencias

Albelda, M. y Estellés, M. (coords.). *Corpus Ameresco*. Consultado en: <http://esvaratenuacion.es/corpus/corpussrch.php?cmd=reset#>.

Audacity Team (1999-2018). Audacity® [software informático]. Consultado en: <https://www.audacityteam.org/>.

Briz, A. y grupo Val.Es.Co. (2002). *Corpus de español coloquial*. Madrid, España: Arco/Libros.

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

Cabedo, A., y Pons, S. (Eds.). (2013). *Corpus Val.Es.Co 2.0.* Valencia: Val.Es.Co. (Valencia, Español Coloquial), Departamento de Filología Española, Universidad de Valencia. Consultado en <http://www.valesco.es>.

Nijmegen (2018): ELAN (Versión 5.2) [Software informático]: Max Planck Institute for Psycholinguistics. Consultado en: <https://tla.mpi.nl/tools/tla-tools/elan/>.

Jefferson, Gail (1984). On the organization of laughter in talk about troubles. En J. Maxwell Atkinson & John Heritage (Eds.), *Structures of Social Action: Studies in Conversation Analysis*, pp. 346-369, Cambridge: Cambridge University Press.

Jefferson, Gail (2004). Glossary of transcript symbols with an introduction. En Gene Lerner (Ed.), *Conversation analysis. Studies from the first generation* (pp. 13-31), Amsterdam: John Benjamin‘s Publishing Company. Consultado en:

<http://dx.doi.org/10.1075/pbns.125.02jef>.

PRESEEA (2014-): *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*. Alcalá de Henares: Universidad de Alcalá. Consultado en: <http://preseea.linguas.net>.

Hacia un corpus histórico anotado del euskera

Ainara Estarrona [a]; Manuel Padilla-Moyano [b]; Ander Soraluze [a]

[a]IKER – CNRS, Francia

[b] IKER – UPPA, Francia

***Palabras clave**: corpus histórico anotado; anotación morfosintáctica; vasco; euskera*

Abstract

Presentamos el proyecto *Basque in the Making: A Historical Look at a European Language Isolate*, recientemente financiado por la *Agence Nationale de la Recherche* francesa (ANR). Uno de los objetivos principales de este proyecto es la creación de una base de datos anotada del vasco histórico. Esta herramienta permitirá búsquedas sistemáticas, estableciendo un modelo para futuros trabajos de desarrollo de bases de datos históricas de la lengua vasca.

La base de datos se sustenta en un corpus de referencia de aproximadamente 750.000 palabras, tamaño aceptable teniendo en cuenta las habituales limitaciones que conciernen los materiales históricos, especialmente acusadas en lenguas como el vasco (cf. Claridge 2009). Este corpus de referencia abarca la producción escrita en euskera entre los siglos XV y XVIII, es decir, el período de tiempo en el que están representados todos los dialectos históricos vascos. Los textos han sido seleccionados atendiendo a

1) representatividad dialectal e histórica; 2) contexto social; y 3) existencia de ediciones fiables. En la estructura de metadatos, cada texto se clasifica de acuerdo al período, dialecto, autor, género, tipo de texto e información sociolingüística.

El motor de búsqueda de la base de datos permitirá explorar el corpus por morfemas, por categorías gramaticales, por secuencias de categorías gramaticales y por configuraciones estructurales específicas que también anotamos. Asimismo, el desarrollo del proyecto implica la creación de tres corpus paralelos: 1) el de los textos originales —ediciones facsímile—; 2) el de las transcripciones; y 3) el de los textos anotados. No existe una herramienta así para la lengua vasca, y recursos similares

tampoco son fáciles de encontrar en otras lenguas minoritarias de Europa —cabe mencionar el caso del galés (Willis 2004)—.

Nuestro objetivo es dar cuenta de algunos de los problemas que estamos tratando durante el desarrollo del corpus. Por una parte, trabajamos con elementos morfológicos que bien han desaparecido del vasco actual, bien resultan marginales en términos de distribución dialectal. Por otra parte, tenemos en cuenta la evolución de los rasgos sintácticos del euskera histórico, factor que añade complejidad al proyecto. Dadas las diversas cuestiones que se están formulando durante constitución del corpus, los problemas prácticos derivados del trabajo con un material gramatical altamente diversificado, así como la falta de un sistema de anotación establecido para una gran parte de las formas históricas del euskera, este proyecto también es técnicamente pionero.

Referencias

Claridge, C. (2009). Historical corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An International Handbook* (pp. 242-259). Berlin: Mouton de Gruyter.

Willis, D. (2004). *A Historical Corpus of the Welsh Language 1500-1850*, University of Cambridge. Online: http://people.pwf.cam.ac.uk/dwew2/hcwl/menu.htm

Combinaciones de palabras en alemán en páginas electrónicas de alojamientos de Alemania, Austria y Suiza

Maricel Esteban-Fonollosa

Department of English and German Studies, University of Valencia, Spain.

***Palabras clave**: combinaciones de palabras, lingüística de corpus, fraseología, alemán, turismo*

Abstract

Sobre la base de un estudio comparativo esta comunicación explora las combinaciones de palabras recurrentes en alemán de textos turísticos en páginas electrónicas de alojamientos de los países de lengua alemana: Alemania, Austria y Suiza. Con ello se persigue, por una parte, determinar las combinaciones que tipifican este tipo de textos y, por otra, determinar cuáles son aquellas que dan cuenta de la posible variación diatópica entre los mismos. Para ello se ha elaborado un corpus específico realizado para esta investigación.

En primer lugar, se presenta brevemente el corpus sobre el que se fundamenta el trabajo, así como los criterios de selección de los textos que conforman la base de datos (Biber, 1993; Lemnitzer & Zinsmeister, 2006; McEnery & Wilson, 2012). Se trata de un corpus monolingüe, compilado entre marzo y septiembre de 2018, que contiene alrededor de

2.400.000 tokens y que está dividido en tres subcorpus, correspondientes a cada uno de los tres países mencionados. Dentro del sector turístico de alojamientos se ha decidido acotar el campo a los alojamientos hoteleros urbanos, asumiendo la diversidad arquitectónica de las ciudades y el diferente enclave de las mismas en los tres países seleccionados.

En segundo lugar, se muestra la metodología para la extracción de combinaciones de palabras recurrentes en cada uno de los subcorpus, empleando técnicas de corpus y estableciendo umbrales de fiabilidad (Fuster-Márquez, 2014; Steyer, 2013). Además, se muestran los criterios de clasificación y filtrado de dichas combinaciones que van a ser objeto de estudio. A fin de determinar la especificidad de las mismas en nuestro corpus especializado frente a un corpus general de la lengua alemana, las frecuencias obtenidas son puestas en relación con las obtenidas en un corpus de referencia del alemán, en este caso en el Deutsche Referenzkorpus (DeReKo, 2016) del IDS.

En tercer lugar, se muestra el procedimiento de análisis de las secuencias filtradas a partir de su grado de tipicidad, de su perfil de contextualización y de su constelación paradigmática (Bubenhofer, 2009). Este procedimiento se llevará a cabo para cada uno de los subcorpus con el objetivo de realizar un contraste entre los resultados de cada uno de ellos y analizar las posibles diferencias geolectales.

Finalmente, la discusión de resultados planteará la idoneidad del empleo de metodología de corpus y su apoyo en cálculos estadísticos, por una parte, para determinar si las secuencias recurrentes halladas perfilan el carácter idiosincrático de los textos turísticos estudiados y, por otra parte, para hallar rasgos de variedad dialectal diatópica en textos específicos. Las conclusiones obtenidas serán de interés desde el punto de vista pedagógico, lexicográfico y traductológico.

Referencias

Biber, D. (1993). *Representativeness in corpus design. Literary and Linguistic Computing*, 8(4), 243-257. DOI: 10.1093/llc/8.4.243.

Bubenhofer, Noah (2009). *Sprachgebrauchsmuster: Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin u.a.: de Gruyter.

Burger, H. (2015). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Erich Schidt Verlag

DeReKo, (2016). *Deutsches Referenzkorpus*. Mannheim: Institut für Deutsche Sprache. http://www.ids-mannheim.de/DeReKo.

Fuster-Márquez, M. (2014). *Lexical bundles and phrase frames in the language of hotel websites*. English Text Construction, 7(1), 84–121. https://doi.org/10.1075/etc.7.1.04fus

Lemnitzer, L., & Zinsmeister 1967, H. (2006). *Korpuslinguistik: eine Einführung*.Tübingen: Narr.

McEnery, T., & Wilson, A. G. (Arthur G.). (2012). *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press.

Steyer, K. (2013). *Usuelle Wortverbindungen: zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Studien zur deutschen Sprache (Vol. Bd. 65). Tübingen: Narr.

Atribución semitautomática de la tipología de textos un corpus del patrimonio cultural (Corpus Lessico dei Beni Culturali)

Annick Farina

Università di Firenze, Italia

***Keywords**: Lexicografía, Lingüística de Corpus, Análisis de la traducción*

Abstract

Este trabajo presenta el marco metodológico desarrollado de control de los corpus textuales del grupo de investigación Lessico dei Beni Culturali, ya consultables en el sito del grupo (www.lessicobeniculturali.net), para mejorar tanto el modelo de búsqueda como los tipos y las proporciones de los textos que incrementarán los corpus en el futuro. Queremos verificar, en particular, la adecuación de la atribución de tipología según el grado de especialidad de los textos introducidos y la posibilidad de usar esta atribución de tipología con fines lexicográficos para proponer de modo automático o semi-automático marcas lexicográficas (en particular diatécnicas y diacrónicas).

Hemos pensado evaluar el corpus francés LBC (actualmente de tres millones de palabras) trabajando sobre combinaciones léxicas especializadas del ámbito artístico de forma contrastiva entre los diferentes tipos de textos a través de un análisis de concordancias.

Como Cabré (2002: 12) pensamos que los términos "se diferencian, no de las palabras, sino de las unidades que no son términos porque adquieren un valor de contenido especializado en cuanto se 'sitúan' y son utilizados 'situadamente'". Este "situadamente" se puede medir en parte con la tipología que los especialistas que han escogido los textos de nuestros corpus han usado, basándose en las condiciones de producción de los mismos textos (tipo de público, biografia del autor, ...), pero también analizando el modo en que estos textos usan el material léxico.

Este segundo tipo de análisis es indispensable debido a la dificultad que tenemos para categorizar algunos textos que presentan características híbridas, una mezcla de lenguaje de especialidad y general no fácilmente reconocible. Se trata sobre todo de las guías turísticas y de algunas traducciones de textos especializados.

Nos basamos en análisis que no consideran la lengua de traducción como un "tercer código" (Frawley, 1984), sino que creemos que los textos traducidos tienen un lugar propio entre los textos con la misma base textual, están junto a los textos "de la lengua natural". Incluir textos traducidos ha sido considerado necesario para tener un corpus extenso en particular por lo que respecta la descripción del arte italino, y para que sea un corpus comparable con las otras lenguas del proyecto. Esto nos ha obligado a plantear medidas de control de calidad de las traducciones (cfr. Farina & al., 2018 y Uzkoreit & al., 2013). El análisis que plantearemos nos permitirá hacer un control de calidad, poniendo de relieve los problemas de «nivelación» y de «simplificación» (Baker, 1993) en algunos textos traducidos, ya que algunas traducciones no respetan las condiciones de producción del texto original, utilizan un léxico que no corresponde al que sería previsible o adecuado según su tipología, según se trate de un texto especializado o de un texto de de divulgación, por ejemplo. Estas diferencias con respecto a lo que se

supone que estos textos contienen hay que preveerlas porque podría falsear, el análisis automático y la extracción de concordancias que se hará para la redacción del diccionario.

Para crear la macroestructura de nuestro futuro diccionario de la lengua del arte (cfr. Farina, 2016), de hecho, habíamos seleccionado las palabras de nuestro corpus que no se encontraban en la misma proporción en corpus de lengua general (cfr. Billero & al., 2017). Nos proponemos ahora analizar las combinaciones notablemente usuales en contexto especializado (verificable como las otras con criterios de frecuencia) y diferenciar las coocurrencias frecuentes de las restricciones combinatorias típicas del ámbito artístico (cfr. L'Homme, M.-C., 2007)

References

Baker, M. (1993) Corpus Linguistics and Translation studies. Implications and Applications. In M. Baker & al. (Ed), *Text and Technology* (pp. 233–250), Amsterdam/Philadelphie: Benjamins.

Billero, R. & Nicolas Martinez, M. C. (2017). Nuove risorse per la ricerca del lessico del patrimonio culturale corpora multilingue LBC. *Chimera*, 4, 203-216,ISSN:2386-2629

Cabré M. T. (2002). Textos especializados y unidades de conocimiento: metodología y tipologización. In J. García Palacios & M. T. Fuentes, (Eds), *Texto, terminología y traducción* (pp.15-36). Salamanca: Ediciones Almar.

Farina, A. (2016). Le portail lexicographique du Lessico plurilingue dei Beni Culturali, outil pour le professionnel, instrument de divulgation du savoir patrimonial et atelier didactique. *Publif@rum*, 24, ISSN:1824-7482

Farina, A. & Billero, R. (2018). Comparaison de corpus de langue « naturelle » et de langue « de traduction » : les bases de données textuelles LBC, un outil essentiel pour la création de fiches lexicographiques bilingues. In *JADT 2018 - International Conference on Statistical Analysis of Textual Data, Roma, 12-15 giugno 2018* (pp. 108-116), Roma: UniversItalia.

Frawley, W. (1984) Prolegomenon to a theory of translation. In W. Frawley (Ed), *Translation: Literary, Linguistic and Philosophical Perspectives* (pp. 159-175), Newark: Univ. of Delaware Press.

L'Homme, M.-C. (2007). Using Explanatory and Combinatorial Lexicology to Describe Terms. In L. Wanner (Ed.), *Selected Lexical and Grammatical Topics in the Meaning-Text Theory. In Honour of Igor Mel'cuk* (pp-11-40). Amsterdam: John Benjamins.

Uzkoreit, H., Burchardt, A. & Lommel, A. (2013), A New Model of Translation Quality Assessment. *Tralogy - Session 5 - Assessing Quality in MT / Mesure de la qualité en TA* http://lodel.irevues.inist.fr/tralogy/index.php?id=188

Analizando sentimiento en Twitter: el fenómeno de Operación Triunfo

Carla Fernández Melendres; Jorge Galván Berdonces; Antonio Moreno-Ortiz

Universidad de Málaga, Spain

***Palabras clave**: análisis del sentimiento; corpus; trending topic; hashtag*

Resumen

Operación Triunfo (OT) es un *talent show* musical en donde se valora a los concursantes no solo por sus dotes artísticas sino también por su personalidad. OT se reestrenó en 2017 después de un largo periodo de inactividad tras el desgaste del formato, pero se ha conseguido reimpulsar trabajando desde las redes sociales en conjunción con los canales tradicionales, generando un éxito musical sin precedentes que traspasa fronteras (el fenómeno se expande hoy también a Latinoamérica). Su gran presencia en redes sociales, principalmente Twitter, YouTube e Instagram, hace que OT consiga mantenerse diariamente como *trending topic* en España gracias al uso de un hashtag diario, llegando a ser *trending topic* mundial en los días de gala.

En OT el publico tiene un papel importante, ya que decide a través de una aplicación móvil (totalmente gratuita pero limitada a un voto diario), llamadas o sms, cual es su concursante favorito/a y a qué concursante salvan cada semana. Los usuarios de Twitter comparten y comentan sus votos durante la semana, así como todos los acontecimientos de la academia. Por ello, Twitter se convierte en una encuesta diaria donde encontrar las intenciones de voto.

El objetivo de este estudio es analizar los tuits de OT2018 haciendo un seguimiento de los hashtags diarios y de las galas para intentar predecir, por un lado, quién va a ser el favorito/a del publico y, por otro, quien va a ser el siguiente expulsado/a de la academia. Para ello hemos aplicado las herramientas de análisis basado en el corpus y análisis de sentimiento.

Metodológicamente, empezamos descargando los tuits diarios, el hashtag de los cuales se compone por #OTDirecto + la fecha (por ej. "#OTDirecto20SEP"), y los tuits de las galas cuyo hashtag es #OT18Gala + el numero de la gala (por ej. "#OT18Gala0"), y así sucesivamente hasta acabar el concurso. Una vez descargados, hemos creado varios subcorpus clasificando los tuits por semanas y concursantes nombrados usando Lingmotif (Moreno-Ortiz 2017, 2018) y su riqueza léxica. Esta herramienta analiza textos desde la perspectiva del Análisis de Sentimiento y es capaz de determinar la orientación semántica (si es positivo, negativo o neutro y en qué grado) de un texto o conjunto de textos, mediante la detección de expresiones lingüísticas que indican una determinada polaridad.

A continuación empleamos un método de aprendizaje no supervisado basado en Pointwise Mutual Information (PMI) (Turney 2002), que nos permite establecer una correlación biunívoca entre cada palabra clasificada por Lingmotif y cada palabra asignada al concursante. Como no siempre se nombra al concursante, se ha creado un diccionario de referentes para cada uno de ellos que incluye otros nombres, motes o canciones por las que se han dado a conocer durante el concurso.

Los resultados obtenidos semanalmente de nuestro análisis muestran una correlación entre la polaridad del vocabulario de los seguidores de OT con los eventos producidos en el *talent show*. Como futura investigación y en base a otras predicciones, pretendemos ampliar este método para aplicarlo al análisis de sentimiento basado en el aspecto para política en Twitter.

Bibliografía

Hasan, Ali, Sana Moin, Ahmad Karim, Shahaboddin Shamshirband, Ali Hasan, Sana Moin,

Ahmad Karim, and Shahaboddin Shamshirband. 'Machine Learning-Based Sentiment

Analysis for Twitter Accounts'. *Mathematical and Computational Applications* 23, no. 1

(27 February 2018): 11.

Liu, Bing. 'Sentiment Analysis and Opinion Mining'. *Synthesis Lectures on Human Language*

Technologies 5, no. 1 (23 May 2012): 1–167.

Moreno-Ortiz, Antonio. *Lingmotif* (version 1.0). Málaga, Spain: Universidad de Málaga, 2018.

http://tecnolengua.uma.es/lingmotif.

Moreno-Ortiz, Antonio. 'Lingmotif: A User-Focused Sentiment Analysis Tool'. *Procesamiento*

Del Lenguaje Natural 58 (2017): 133–40.

O'Keeffe, Anne, and Michael McCarthy, eds. *The Routledge Handbook of Corpus Linguistics*.

Routledge Handbooks in Applied Linguistics. London; New York, NY: Routledge, 2010.

Pang, Bo, and Lillian Lee. 'Opinion Mining and Sentiment Analysis'. *Foundations and Trends*

in Information Retrieval 2, no. 1–2 (2008): 1–135.

Turney, Peter B. 'Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised

Classification of Reviews'. *Proceedings of the 40th Annual Meeting of the ACL*, 2002,

417–24.

El léxico de prácticas de cura místico religiosas en un corpus de procesos criminales brasileños de los siglos XIX y XX

Renata Ferreira Costa

Departamento de Letras Vernáculas, Universidade Federal de Sergipe, Brasil ***Keywords:*** *Crítica Textual; Lexicología; Discurso Civilizador; Curanderismo; Represión.*

Abstract

El período monárquico brasileño termina el 15 de noviembre de 1889, cuando se promulga la Primera República, también conocida como República Vieja (1889-1930), cuyos ideales se fundamentan en los presupuestos del Positivismo, según los cuales el progreso de una nación está directamente relacionado al respeto al orden. En ese contexto se aprueba el Código Penal de 1890, que, en sus artículos 156, 157 y 158, como forma de manutención del orden público y social, criminaliza las prácticas de cura místico religiosas, especialmente las provenientes de creencias minoritarias – especialmente las de matriz afrobrasileña, contra la salud pública, una vez que, para el Estado, representaban conductas antisociales, en desacuerdo al proyecto republicano de modernización, higienización y saneamiento urbano. Así, a pesar de haber libertad religiosa desde la Constitución de 1891, el discurso de las clases dominantes y de la legislación de Brasil se apoya en la idea de cura por medio de saberes mágicos, denominados en portugués como sinónimos de "feitiçaria", "bruxaria", "magia negra", "curandeirismo", "macumba", "ocultismo" y "espiritismo", como algo no civilizado, que remite al recuerdo de un pasado no distante de esclavitud y de una herencia de religiones originarias de África, lo cual se deseaba olvidar. En 1940 un nuevo Código Penal entra en vigor, aún en vigencia, que sigue criminalizando las prácticas de cura populares como ejercicio ilegal de la medicina – Art. 284, siguiendo con la persecución, represión y condenación penal de hombres y mujeres que, en contrapunto al saber médico científico, ofrecían la cura de enfermedades por intermedio de sustancias de la naturaleza, objetos, gestos o palabras. La documentación criminal depositada en archivos públicos que se insertan en ese contexto es fuente primaria necesaria e indispensable que atesta el universo sociocultural, jurídico, religioso y lingüístico de una sociedad. Así que la investigación del universo lexical presente en ese tipo de documento sintetiza el discurso, o los discursos, de dos grupos – los practicantes del curanderismo y principalmente de los agentes represores de esa práctica, que en las entrelíneas de los textos diseminan prejuicios y racismo disimulados en la defensa de la moral y las buenas costumbres, según los patrones socioculturales europeos. Este trabajo busca contribuir al tema con un estudio de las características lexicales de prácticas de cura místico religiosas, en perspectiva discursiva apoyada en la constitución de un corpus de procesos criminales producidos en Sergipe – Brasil, entre 1897 y 1939. Esta investigación, que se dio en el ámbito del proyecto *Edición Filológica y Estudio Léxico Semántico de Procesos Criminales de Curanderismo,* en desarrollo en la Universidad Federal de Sergipe – UFS, se fundamenta en los principios teórico metodológicos de la Crítica Textual (Spina, 1994; Spaggiari y Peruggi, 2004; Cambraia, 2005), de la Lexicología, con énfasis en la Teoría del Campo Lexical (Coseriu, 1977), del Análisis del Discurso (Charaudeau, 1982, 1999, 2008a, 2008b) y de los Estudios Socio-históricos sobre el curanderismo en Brasil (Souza, 2009; Santos, 2009). La constitución de un corpus documental mínimo de investigación permitió establecer el estudio de una variedad lexical de la lengua portuguesa de Brasil, con sistematización y descripción del vocabulario de las prácticas de curanderismo en el período delimitado. Ese trabajo fue fundamental para la delimitación de los campos léxico-semánticos de esa temática y para la elaboración de definiciones según los usos especiales de ese universo discursivo.

Referencias

Cambraia, C.N. (2005). *Introdução à Crítica Textual*. São Paulo: Martins Fontes.

Charaudeau, P. (1982). Eléments de sémiolinquistique d'une théorie du langage à une

analyse du discours. *Connexions*, 38, 7-30.

Charaudeau, P. (1995). Une analyse sémiolinguistique du discours. *Langages*, 117, 96-111.

Charaudeau, P. (1999). Análise do discurso: controvérsias e perspectivas. In H. MARI, S. PIRES, A. CRUZ & I. MACHADO (Orgs.), *Fundamentos e dimensões da análise do discurso*. Belo Horizonte: Carol Borges.

Charaudeau, P. (2008a). *Linguagem e discurso: modos de organização*. São Paulo: Contexto.

Charaudeau, P. (2008b). *Discurso Político*. São Paulo: Contexto.

Coseriu, E. (1987). *Gramática, semántica, universales estudios de la lingüística funcional*. 2. ed. rev. Madrid: Gredos.

Santos, E.F. (2009). *O Poder dos Candomblés: perseguição e resistência no Recôncavo da Bahia*. Salvador: EDUFBA.

Souza, L.M. (2009). *O diabo e a terra de Santa Cruz: feitiçaria e religiosidade no Brasil Colonial*. 2. ed. São Paulo: Companhia das Letras.

Spaggiari, B., & Peruggi, M. (2004). *Fundamentos da Crítica Textual*. Rio de Janeiro: Lucerna.

Spina, S. (1994). *Introdução à Edótica: Crítica Textual*. 2. ed. rev. atual. São Paulo: Ars Poética; Edusp.

A corpus-based discourse and lexical analysis of WhatsApp conversations

Imma Ferri-Miralles; Keith Stuart

Departament de Lingüística Aplicada, Universitat Politècnica de València, España

***Keywords**: chat corpus analysis; social interaction; creative language and lexis*

Abstract

WhatsApp is a software application that allows the exchange of text messages, videos, audios, images and photo between mobile phone users. The multi-modality function of WhatsApp represents an interesting mode of casual conversation that can be examined simultaneously from a sociological and linguistic point of view. This study is a discourse and lexical analysis of WhatsApp conversations from two distinct sociological groups. Aspects of discourse that can be studied with sociolinguistic corpora like ours are discourse units and phenomena such as turns and turn taking, speech acts, and interactions. Lexical aspects such as pinpointing which are the most frequently used lexical words among the users, whether these words contain more positive or negative evaluation, how they relate to and realize chat topic, the use of orthotypographical elements and conventions.

The conversations were collected from the two researchers' phones, exporting chat histories via e-mail and saving the chat histories in an Excel spreadsheet. The chat history is anonymized and statistically evaluated, and the anonymized chat and the statistics are stored in a separate Excel sheet. When we say that the chat history is statistically evaluated, we mean that initial data about the chat history such as the number of messages sent, length of the messages, the shortest and the longest text messages are counted and the users who are most and least loquacious are identified. The language of the text messages is then analysed separately using various applications and software tools.

A sociological focus through the analysis of user profiles and the analysis of semantic and thematic content alongside a linguistic analysis of conversational text was carried out. Furthermore, the study explores how a technological environment such as the WhatsApp application impacts on the linguistic structure and sequence of message exchanges in casual conversation. The impact of given contexts is revealed and realised linguistically in messages, in speaker's meaning and intention. The power of emoji and smileys which are the digital version of body language in reducing ambiguity and adding emotional context to what is said is analysed in our corpus

Initial findings from our chat corpus confirm that age, being the group's manager, and group's subject or topic play an important role in the participation level. Interaction tends to be structured and formulaic in the forms of address, opening and closing formulae (lots of repetition of lexical chains). Greetings are preferred over deferential forms of address that could indicate a way of showing solidarity between interactants. Findings obtained show that people use conventionally indirect strategies and a great deal of syntactic modification. Above all else, the chats illustrate highly creative use of language for social purposes and meaning creation in the form of vocal spelling, lexical surrogates (textual representations of vocal sounds that are not words e.g. haha), initialism, clipping, shortening, reduplication amongst other linguistic realisations.

References

Anglano, C. (2014). Forensic analysis of WhatsApp Messenger on Android smartphones.

Digital Investigation, 11 (3), 201-213. doi: 10.1016/j.diin.2014.04.003.

Maíz-Arévalo, C. (2018). Emotional Self-presentation on WhatsApp: Analysis of the Profile Status. *Russian Journal of Linguistics*, 22(1), 144-160.

Pérez-Sabater, C. (2015). Discovering language variation in WhatsApp text interactions.

Onomázein, 31 (1), 113-126.

Sánchez-Moya, A., & Cruz-Moya, O. (2015). "Hey there! I am using WhatsApp": a preliminary study of recurrent discursive realisations in a corpus of WhatsApp statuses. *Procedia-Social & Behavioral Sciences*, 212, 52-60.

Schwind, A., & Seufert, M. (2018). WhatsAnalyzer: a Tool for Collecting and Analyzing WhatsApp Mobile Messaging Communication Data. In *2018 30th International Teletraffic Congress* (ITC 30) (Vol. 1, pp. 85-88). IEEE.

Seufert, M., Hoßfeld, T., Schwind, A., Burger, V., & Tran-Gia, P. (2016). Group-based communication in WhatsApp. *2016 IFIP Networking Conference (IFIP Networking) and Workshops*, 536-541.

Grammaticalization and pragmaticalization of second person plural forms in World Englishes: a corpus-based study

Liviana Galiano

English Language and Linguistics, Lancaster University, UK

Keywords: *second person pronouns, quantifiers, grammaticalization, pragmaticalization, intersubjectivity, World Englishes*

Abstract

English used to distinguish between singular and plural second person pronouns. This distinction was lost during its transition into Modern English which saw the establishment of the single form *you* for both singular and plural reference (Wales 1996).

Yet, many dialects of English have always continued to explicitly mark number on second person pronouns by resorting to different linguistic strategies, both morphological and analytic. Among the morphological strategies, suffixation of the second person pronoun *you* with the regular pluralising mark -s, i.e. *yous*, represents the most common form alongside a host of different spelling variants such as *youse, yiz, yez*, etc. As far as the analytic strategies are concerned, pluralisation of *you* is mainly carried out by making it follow by other linguistic elements (quantifiers, plural nouns etc.) as happens for *you guys*, but also *you all, y(')all, you two, you three, you lot, y'uns*, and so on.

The proposed research is a synchronic corpus-based investigation of second person plural forms (2PL forms) in twenty varieties of English. The corpus (GloWbe) contains 1.9 billion words collected on the web in 2012.

The analysis of corpus data has uncovered features and uses of 2PL forms that are not accounted for in the literature: first, there are geographically-specific preferences as to which 2PL forms are going to be used as plural(s) of *you*. Second, 2PL forms have developed new functions besides the one of marking plurality: they can work as markers of politeness and empathy, singular-reference emphatic markers, attention-getting devices, and possessive determiners. All these functions are seen as clues to on-going processes of grammaticalization and pragmaticalization, which, together with other clues such as syntactic position (right- periphery) and the speech acts 2PL forms tend to be involved in (mainly expressing politeness and emphasis), contributed to the interpretation of 2PL forms as markers of intersubjectivity (Traugott 2014).

Therefore, the analysis of 2PL forms in English supports both the theory of intersubjectification outlined by Traugott (2014), whereby the right periphery of a sentence or utterance tends to be occupied by pragmatically marked items (e.g. attention-getting 2PLs), and the theory of mor- phopragmatics (Dressler and Merlini-Barbaresi 2015, 2017), according to which morphology is capable of a direct interface with

pragmatics without being mediated through semantics, as happens with singular-reference emphatic *yous(e).*

References

Dressler, W.U. and Merlini-Barbaresi, L. 2015 (online)/2017 (print). „Pragmatics and Morphology: Morphopragmatics "in Huang, Y. (ed.). *The Oxford Handbook of Pragmatics*. Pp. 493-510.

Traugott, E. 2014. "Intersubjectification and clause periphery" in Brems, L., Ghesquiére, L., and Van de Velde, F. (eds.) *Intersubjectivity and Intersubjectification in Grammar and Discourse*. Amsterdam/Philadelphia: John Benjamins. 7-28

Schneider, E.W. 2007. *Postcolonial English*: *Varieties around the World*. Cambridge: Cam- bridge University Press.

Schriffin, D. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

Wales, K. 1996. *Personal Pronouns in Present-Day English.* Cambridge: Cambridge University Press.

Wales, K. 2003. „Second Person Pronouns in Contemporary English: The End of a Story or Just the Beginning? "in *Franco-British Studies,* 33-34: 172-

Tratamiento de VOX en la prensa española: un estudio del corpus PRODISNET de discursopolítico

Beatriz Gallardo Paúls; Salvador Enguix Oliver

Universitat de València

Palabras Clave: *Discurso político. Corpus periodístico. Pragmática. Ultraderecha*

Abstract

En esta comunicación presentaremos los resultados de un análisis de corpus sobre el tratamiento léxico y temático de VOX en la prensa española. Pretendemos un objetivo doble: el primero, dar cuenta de cómo la prensa escrita española aborda la cobertura de este partido de extrema derecha; el segundo, comprobar si el mitin de Vistalegre del 07/10/2018 supuso un cambio en el tratamiento que los distintos medios dan a Vox.

Para ello, asumimos un análisis del discurso centrado en la noción de encuadre, que consideramos desde el punto de vista lingüístico (desde la propuesta pre-cognitivista de Fillmore 1976 a las de Lakoff 2004, 2006), y según las versiones más utilizadas en la teoría de la comunicación mediática (por ejemplo, Entman 1991, 1993); en concreto, seguiremos la metodología desarrollada por Gallardo (2014, 2017). Además, incorporaremos algunas nociones de la Escuela de Essex (Howarth 2005), en la medida en que el nuevo protagonismo de Vox puede suponer una *dislocación* discursiva (respecto a la idea de que en España no hay extrema derecha), y tendremos en cuenta

algunos análisis recientes sobre el tratamiento de estas posturas políticas en los medios de comunicación y en los *social media* de países democráticos (Wodaj 2015; Hochschild 2016; Marwick y Lewis 2017; Starbird 2017).

Se presentará brevemente, en primer lugar, el corpus PRODISNET de discurso político, que nace en 2015 y que está integrado por tuits, por vídeos electorales y por datos de prensa escrita; detallaremos los procesos de elaboración, diseño y análisis del corpus en nuestro grupo de investigación. En segundo lugar, presentaremos un ejemplo concreto de análisis, referido al tratamiento que ha hecho la prensa española del partido VOX, siguiendo esta estructura:

a) Elaboración y preparación del corpus: el corpus ha sido elaborado a partir de datos obtenidos con la plataforma FACTIVA, tomando como eje cronológico el mitin de Vistalegre del 7 de octubre de 2018; se extiende desde el 7 de agosto de 2018 hasta el 7 de noviembre de 2018, es decir, 2 meses antes y 1 mes después de dicho evento; descargamos todas las menciones del partido en El País, El Mundo, La Vanguardia y ABC, los cuatro diarios líderes del EGM. Estos datos se criban después por procedimientos manuales (eliminando duplicados y falsos positivos) que suponen ya una primera visión de los datos.

b) Criterios de análisis: para cada texto del corpus se tienen en cuenta elementos formales de identificación: fecha, medio, autoría.

c) Análisis posterior: se utilizará el software QDA Miner para analizar los encuadres del término "Vox" en diferentes niveles pragmáticos, según nuestro modelo de análisis pragmático del encuadre dispuesto en estrategias discursivas; en concreto, se analizará la estrategia léxica (adjetivación y sinonimia aplicada a "Vox"), la estrategia temática (cuándo Vox es tema central o tema periférico de la noticia), la estrategia predicativa (qué acciones atribuye la prensa a Vox y a sus líderes), y la estrategia afiliativa (según los medios se alineen o no con la ideología de Vox).

Referencias

Entman, Robert M. (1991): "Framing US coverage of international news: contrast in narratives of the KAL and Iran air incidents", *Journal of Communication* 41, pp. 6-27.

Entman, Robert M. (1993): "Framing: Toward Clarification of a Fractured Paradigm", *Journal of Communication* 43 (4), pp. 51-58.

Fillmore, Charles (1976): "Frame semantics and the nature of language", *Annals of the New York Academy Sciences*, 280, 20-32.

Gallardo Paúls, Beatriz (2014): *Usos políticos del lenguaje. Un discurso paradójico*, Barcelona: Anthropos/Siglo XXI.

Gallardo Paúls, Beatriz (2017): «Pseudopolítica en la red: indicadores discursivos de desideologización en Twitter», *Pragmalingüística* 25, pp. 189-210

Gallardo Paúls, Beatriz (2018): *Tiempos de hipérbole*, Valencia: Tirant Lo Blanch.

Hochschild, Arlie R. (2016): *Extraños en su propia tierra. Réquiem por la derecha estadounidense*, Madrid: Capitán Swing, 2018. Trad. de Amelia Pérez de Villar.

Howarth, David (2005): "Aplicando la teoría del discurso: el método de la articulación", *Studia Politicae* 5, pp. 37-88.

Lakoff, George (2004): *No pienses en un elefante,* Madrid: UCM, 2007. Trad. de Magdalena Mora.

Lakoff, George (2006): *Puntos de reflexión. Manual del progresista,* Barcelona: Eds. Península, 2008. Trad. de Judith Wells.

Marwick, Alice, and Rebecca Lewis (2017): *Media manipulation and disinformation online,* New York: Data & Society Research Institute.

Starbird, Kate (2017): "Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter", International AAAI Conference on Web and Social Media (*ICWSM*) 2017, http://faculty.washington.edu/kstarbi/Alt_Narratives_ICWSM17-CameraReady.pdf

Wodak, Ruth (2015). *The politics of fear: What right-wing populist discourses mean*. Sage Pubs.

Problems documenting a Philippine Negrito language: Corpus and grammar

Alexandro Garcia-Laguia

University of Barcelona, Spain.

Keywords: *Language Documentation, Corpus-based grammar, Austronesian languages, Negrito populations*

Abstract

The first part of this talk presents the Language Documentation corpus (LD) of an Austronesian language spoken by less than 300 Philippine Negritos on Luzon Island. The NALDoC is an LD-type of corpus, as described in Mosel (2012, p.236), that was developed over a five-year span, and which contains a collection of 19 hours of video and audio recordings of the language. The corpus and collection are accessible online and include a series of monologues and dialogues that focus on two textual genres: explanatory and oral history texts.

The first component of the NALDoC is the collection of annotated texts, which comprises 13 hours of time-aligned transcription and translation using the software ELAN. Part of collection of annotated texts was used to create and develop the second component of the NALDoC: the lexical database, which was developed with the software FLEx. The lexical database contains 100 minutes of texts that were imported from the ELAN annotated texts, and that were enriched with morpheme by morpheme glosses. The lexical database also includes a lexicon of the language that contains 2400 entries.

The collection of annotated texts and the lexical database are linked in several ways to the third component, the grammar of the language, some of which are discussed in this presentation. First, the generalizations appearing on the grammar are made by running searches in the collection of annotations using regular expressions (regex). In addition,

each example in the grammar carries a reference number and hence can be retrieved in its natural context. Another relation between the components is the transcription system which appears in the collection of annotated texts, in the lexical database, and in the grammar. Finally, the classification of parts of speech appearing in the grammar is also reflected in the lexicon of the language, in which each entry carries an abbreviation of the word class it belongs to.

The last component of the corpus is the metadata structure, created with the software Arbil and including all of the metadata related with the documentation project and corpus. For each recording session, the metadata structure includes all the recorded and annotated files, and provides information such as the title of the recording, the content and the setting, the date and location, the participants and their roles, as well as the textual genre. The purpose of this metadata structure is to make the archived materials "understandable, findable, preservable and usable", in the same way as stated in Austin (2016, p.151).

The second part of this presentation explores with more detail the methods used to exploit the collection of annotated texts when carrying out the grammatical analysis of the language, and discusses some of the problems encountered. Finally, some areas of future research are presented.

References

Austin, P. (2016). Language documentation 20 years on. In M. Pütz & L. Filipović (Eds.), *Endangerment of languages across the planet* (pp. 147–170). Amsterdam: John Benjamins.

Bellwood, P. (2011). Holocene Population History in the Pacific Region as a Model for Worldwide Food Producer Dispersals. *Current Anthropology*, *52*(S4), 363-S378.

Bouda, P., & Helmbrecht, J. (2012). From corpus to grammar: how DOBES corpora can be exploited for descriptive linguistics. In S. Nordhoff (Ed.), *Electronic Grammaticography. LD&C Special Publication 04* (pp. 129–159). Honolulu: University of Hawai'i Press.

Dwyer, A. M., & Mosel, U. (2001). *Metadata Description Recommendations: General Contents*.

Himmelmann, N. P. (2012). Linguistic data types and the interface between language documentation and description. *Language Documentation and Conservation*, *6*, 187–207.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based Language Studies: an Advanced Resource Book*. (C. N. Candlin & R. Carter, Eds.). London; New York: Routledge.

Mosel, U. (2012). Advances in the accountability of grammatical analysis and description by using regular expressions. In S. Nordhoff (Ed.), *Electronic Grammaticography. LD&C Special Publication 04* (pp. 235–250). Honolulu: University of Hawai'i Press.

Mosel, U. (2014). Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. In *The art and practice of grammar writing. LLD&C Special Publication 08* (pp. 135–157). Honolulu: University of Hawai'i Press.

Reid, L. A. (1991). The Alta Languages of the Philippines. In Ray Harlow (Ed.), *VICAL 2, Western Austronesian and contact languages: Papers from the Fifth International Conference on Austronesian Linguistics* (pp. 265–297). Auckland: Linguistic Society of New Zealand.

Reid, L. A. (2013). Who Are the Philippine Negritos? Evidence from Language. *Human Biology*, *85*(1), 328–358.

Enunciados averbales ilocutivos: entre la sintaxis y la pragmática

Óscar García Marchena

EILA, Université Paris 3 Nouvelle Sorbonne, Francia

Palabras clave: *Enunciados averbales, fragmentos, sintaxis, pragmática, actos de habla*

Abstract

(1) Yo, encantado.

(2) ¡Suerte en esta nueva aventura!

(3) ¿Una copa?

(4) El día 26, a votar.

(5) ¡Castigado!

Los enunciados sin verbo representados por los ejemplos (1-5) han sido analizados como elípticos en varios estudios, como Merchant (2004, 2006). Sin embargo, el análisis de un gran número de ejemplos de enunciados sin verbo en un corpus del español oral de cierta envergadura permite observar ciertas propiedades que distinguen estos casos del resto de los enunciados analizados como elípticos.

De este modo, la mayor particularidad de estos enunciados es que no recuperan contenido semántico de su contexto sintáctico, sino directamente del contexto situacional, como el acto de habla que realizan, como una orden o una petición. Así, estos enunciados pueden clasificarse en cinco tipos según los actos de habla de la clasificación de Searle (1975): *presentativos* (1), si presentan alguna propiedad de un participante; *expresivos* (2), si expresan un deseo; *promisivos* (3), si constituyen una propuesta o invitación; *directivos* (4), si dan una orden o hacen una petición; y *performativos* (5), si al pronunciarlos, el hablante produce un cambio en la realidad. A la luz de estas propiedades, parece evidente que el contenido elíptico de estas unidades tiene un origen pragmático y no sintáctico.

En consecuencia, si estos enunciados no corresponden a frases elípticas, deben constituir entonces unidades diferentes de la oración. Y si es así, ¿Son unidades de orden sintáctico, con la misma diversidad de configuraciones sintácticas que caracteriza a las

oraciones (tipos sintácticos, valor ilocutivo, estructuras determinadas,etc.), o de estructuras fijas?

Para dilucidar la naturaleza de estas unidades, hemos extraído los 1578 casos que se encuentran en el corpus oral CORLEC (Marcos-Marín 1992), los hemos agrupado según la tipología de Searle (1975), y hemos clasificado su variedad sintáctica según su estructura, tipo sintáctico y categoría gramatical del núcleo. A continuación, hemos analizado su estructura semántica, distinguiendo entre el contenido proveniente de la sintaxis y el que es de naturaleza pragmática, descubriendo patrones en los modos de composición de su contenido semántico.

El análisis del conjunto de ejemplos del corpus ha permitido extraer varias conclusiones: primero, que estas unidades son diferentes de las frases elípticas, ya que no toman su contenido únicamente del contexto sintáctico; segundo, que no se trata de expresiones fijas, ya que muestran diversidad composicional; tercero, que se integran sintácticamente, ya que pueden aparecer en relaciones de coordinación y subordinación; cuarto, que el acto de habla que realizan es predecible por sus propiedades sintácticas; y quinto, que comparten las propiedades sintácticas de las oraciones verbales, pero están restringidas semánticamente por las condiciones de felicidad del acto de habla que expresan. Las originales propiedades de estos enunciados no elípticos sugieren que éstas son unidades con propiedades sintácticas y pragmáticas que deben ser integradas en el estudio de ambas subdisciplinas. Igualmente, esta dualidad permite la identificación automática del acto de habla que realizan, con las consecuencias que ello puede tener para el etiquetado automático de corpus.

References

Fernández, R. (2006). Non-Sentential Utterances in Dialogue: Classification, Resolution and Use. Ph. D. thesis, King's College London.

Garcia-Marchena, Oscar. (2015). Phrases Averbales et Fragments de l'Espagnol Oral. Étude de corpus. Linguistique. Université Paris Diderot (Paris 7).

Lefeuvre, F. (1999). *La phrase averbale en français*. L'Harmatan.

Marcos-Marín, F. (1992). Corpus de referencia de la lengua española contemporánea: Corpus oral peninsular.

Merchant, J. (2004). Fragments and ellipsis. *Linguistics and Philosophy* 6, 661–738. Merchant, J. (2006). Small structures: A sententialist perspective, in Provogac et al., *The syntax of nonsententials: Multidisciplinary perspectives.* pp. 73–91. John Benjamins.

Schlangen, D. (2003). A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue. Ph. D. thesis, School of Informatics, University of Edinburgh.

Searle, J. (1975). *Les actes de langage*. Hermann.

Weir, A. (2014). "Fragments and Clausal Ellipsis". Doctoral Dissertations May 2014

Researching teacher identity construction in in-service teachers' discourses: Combining corpus linguistics and critical discourse analysis

María Dolores García-Pastor[a]; Montserrat Pérez Giménez[b]

[a]*Department of Teaching Languages and Literature, Universitat de València, Spain;*

[b]*Department of Teaching Languages and Literature, Universitat de València, Spain*

Keywords*: identity; teacher identity; corpus linguistics; critical discourse analysis; in- service teachers*

Abstract

This Research on identity in language education and related fields such as applied linguistics has grown exponentially in the past fifteen years (Norton & Toohey, 2011). Scholars in these fields have commonly explored this construct from poststructuralist and narrative perspectives to the neglect of alternative approaches like corpus linguistics. Except for certain studies within applied linguistics (e.g., Hyland, 2010), we are not aware of investigations that use corpus-based methods to examine identity issues in language education. The present study thus intends to show how a corpus-based analysis of teacher interviews that comprises a computerised linguistic analysis and combines quantitative and qualitative analyses can shed light on the construction of teacher identity by Primary and Secondary Education teachers from public Catalan schools in Spain. To this end, we have followed Richards' (2015), and Pennington and Richards' (2016) notion of teacher identity as a dynamic construct that a) is shaped by the context in which the teacher works; b) may have different features at different times;

reflects the teacher's self-image and self-awareness as captured in the stories s/he tells about her/himself, and as this is projected to and understood by others; and d) is modified in interaction with new individuals, social groups, and contexts of action. The data for this study comprises 20 semi-structured interviews that were longitudinally collected before, during, and after the teachers received training on L1 writing and grammar, and implemented a specific instructional treatment on these aspects of language in the classroom. These data belong to a larger funded project, which includes a qualitative study of teacher beliefs about their own teaching and their students' learning with regard to the writing of school genres, and a quantitative study on different teacher, learner, and textual variables. Teacher identity construction in the interviews has been analysed mainly in terms of the teachers' view of their institutional role and their projection of their unique individual identity through the autobiographical information they offered (Pennington & Richards, 2016; Richards, 2015). In particular, our analysis consists of a combination of a corpus-based linguistic approach and critical discourse analysis (Cheng, 2013). We thus analysed the corpus using the qualitative data analysis computer programme Atlas.ti and the freeware corpus analysis toolkit AntConc 3.5.7. This programme facilitates the rapid calculation of frequency counts and the identification of keywords in context in the data. Our preliminary results of the in- depth analysis of the three interviews to one of the teachers indicate an increasing presence of the existential "I" and verbal forms in the first person singular over time. This teacher therefore seemed to reinforce her teacher identity whilst indirectly resisting the identity she was ascribed in the training she received and the instructional treatment she had to implement (Pérez-Giménez & García-Pastor, 2018). The present study aims to enquire whether this

pattern is recurrent in the rest of the corpus through the examination of personal pronouns, i.e., number of personal pronouns per one hundred words, and their concordances with other linguistic elements. Overall, this study intends to highlight the potential of corpus linguistics for researching issues in language education that go beyond the pedagogical uses of this approach.

References

Cheng, W. (2013). Corpus-based linguistic approaches to critical discourse analysis. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. London: Blackwell.

Hyland, K. (2010). Community and individuality: Performing identity in applied linguistics. *Written Communication*, 27(2), 159–188. doi: 10.1177/0741088309357846

Norton, B., & Toohey, K. (2011). Identity, language learning, and social change. *Language Teaching*, 44(4), 412–446. doi: 10.1017/S0261444811000309

Pennington, M. C., & Richards, J. C. (2016). Teacher identity in language teaching: Integrating personal, contextual, and professional factors. *RELC Journal*, 47(1), 5–23. doi: 10.1177/0033688216631219

Richards, J. C. (2015). *Key issues in language teaching*. New York: Cambridge University Press.

Pérez-Giménez, M., & García-Pastor, M. D. (2018, Nobember 28). *La construcción de la identidad docente en la Educación Primaria: estudio de caso.* Paper presented at the XIX Congreso Internacional de la SEDLL, Almería

Patterns of Stylistic Variation in Historical Sociolinguistics: Evidence from Corpora of Private Correspondence

Tamara García Vidal; Belén Zapata Barrero

Department of English Philology, University of Murcia, Spain

Keywords: *sociolinguistics, style-shifting, audience design, speaker design, historical corpora, private correspondence*

Abstract

Style constitutes an essential component for the social meaning of speakers' sociolinguistic behaviour in interpersonal communication. It enjoys a central position in the correlation of *social*, *linguistic* and *stylistic* variation, legitimating the distinction between *inter*-speaker (social) and *intra*-speaker (stylistic) variation (Bell 1984). However, despite the importance of style in Present-day Sociolinguistics, the study of style-shifting within the variationist tradition has been secondary until very recently. Fortunately, now it is becoming a major focus in the research field (see Eckert &

Rickford 2001; Coupland 2007; Hernández-Campoy & Cutillas-Espinosa 2012; or Hernández Campoy 2016). Similarly, Historical Sociolinguistics has been applying the tenets and findings of contemporary sociolinguistic research to the interpretation of linguistic material from the past (see Nevalainen & Raumolin-Brunberg 1996; 2003; Hernández-Campoy & Conde-Silvestre 2012; Säily *et al* 2017), but unfortunately neglecting the relevance of intra-speaker variation.

Research based on corpora of historical correspondence has confirmed the relevance of letters as sources to reconstruct the sociolinguistic contexts of language variation and change in the past (see Nevalinen & Tanskanen 2007; Dossena & Tieken-Boon van Ostade 2008; Sairio 2017; Dossena & Del Lungo Camiciotti 2012; or Auer, Schreier & Watts 2015). But the proved validity of private correspondence in Historical Sociolinguistics now becomes of paramount importance not just to detect the nature and direction of language change longitudinally along a group of homogeneous speakers at a macro-level, but also to find out how a change in progress acts at a micro-level (Pallander-Collin, Nevala, & Nurmi 2009; Hernández-Campoy & Conde-Silvestre 2012; or Hernández-Campoy & García-Vidal 2018a; 2018b). Epistolary correspondence is intended as a dialogic exchange where addressivity, recipiency and relationality are key elements. With the study of their interactional communication in letters, informants are put under the lens of the microscope and their sociolinguistic behaviour is observed on the basis of their addressees –rather than addressers– and context types, approaching them atomistically, cross-sectionally, and microscopically.

The aim of this paper is to show results and interpretation from the investigation of the linguistic mechanisms and personal motivations for the patterns of style-shifting in late Medieval England by applying current theoretical models of intra-speaker variation to historical corpora of private correspondence: audience design (Bell 1984) and speaker design (Coupland 2007). The study is carried out through the analysis of the use of the orthographic variable (TH) by male members of the Paston family from the *Paston Letters* corpus when addressing recipients from different social ranks. The data show addressee and referee design patterns in the communicative practice of Medieval individuals. In addition to tracing language variation and change in speech communities, private letters may also shed light onto the motivations and mechanisms for intra-speaker variation in individuals and their stylistic choices in past societies such as those of the late Middle and early Modern English periods. This study also contributes to account for the social meaning of *inter-* and *intra-* speaker variation in the sociolinguistic behaviour of speakers at the individual level as a linguistic resource for identity construction, representation, and even social positioning in interpersonal communication. Ultimately, the extension and extrapolation of conclusions obtained from sociolinguistic studies on patterns of stylistic variation of current English situations to Late Middle English and Early Modern English communities also allow us to test the validity of current theoretical models of diaphasic variation.

References

Auer, A.D., Schreier; D. & Watts; R. (eds.) 2015. *Letter Writing and Language Change*. Cambridge: Cbambridge University Press.

Bell, Allan. 1984. Language Style as Audience Design. *Language in Society* 13. 145-204.

Coupland, Nikolas. 2007. *Style: Language Variation, and Identity*. Cambridge: Cambridge University Press.

Dossena, Marina & Del Lungo Camiciotti, G. (eds.) 2012. *Letter Writing in Late Modern Europe*. Amsterdam/Philadelphia: John Benjamins.

Dossena, Marina & Tieken-Boon van Ostade, I. (eds.) 2008. *Studies in Late Modern English correspondence*: *Methodology and Data*. Bern: Peter Lang.

Eckert, Penelope & Rickford, John. (eds.) 2001. *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Hernández-Campoy, Juan Manuel. 2016. *Sociolinguistic Styles*. Malden: Wiley-Blackwell.

Hernández-Campoy, J.M. & Conde-Silvestre, J.C. (eds.) 2012. *The Handbook of Historical Sociolinguistics*. Malden: Wiley-Blackwell.

Hernández Campoy, J.M. & Cutillas-Espinosa, J.A. (eds.) 2012. *Style-Shifting in Public: New Perspectives on Stylistic Variation*. Amsterdam/Philadelphia: John Benjamins.

Hernández-Campoy, J.M. & García-Vidal, T. 2018a. Persona management and identity projection in English medieval society: Evidence from John Paston II. *Journal of Historical Sociolinguistics* 4(1): 1-31.

Hernández-Campoy, J.M. & García-Vidal, T. 2018b. Style-Shifting and Accommodative Competence in Late Middle English Written Correspondence: Putting Audience Design to the Test of Time. *Folia Linguistica Historica* 39(2): 383-420.

Nevala, M. & Palander-Collin, M. 2005. Letters and letter writing: Introduction. *European Journal of English Studies* 9(1): 1-7.

Nevalainen, T. & Raumolin-Brunberg, H. (eds.) 1996. *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi.

Nevalainen, T. & Raumolin-Brunberg, H. 2003. *Historical Sociolinguistics. Language Change in Tudor and Stuart England*. London: Longman Pearson Education.

Nevalainen, T. & Tanskanen, S.K. (eds.) 2007. *Letter Writing*. Amsterdam/Philadelphia: John Benjamins.

Pallander-Collin, Minna, Nevala, Minna, & Nurmi, Arja. (2009). The Language of Daily Life in the History of English. Studying how Macro Meets Micro. In: Arja Nurmi, Minna Nevala and Minna Pallander-Collin (eds.), *The Language of Daily Life in England (1400-1800)*. Amsterdam and Philadelphia: John Benjamins, 1-23.

Säily, T., Nurmi, A., Palander-Collin, M. & Auer, A. (eds.) 2017. *Exploring Future Paths for Historical Sociolinguistics*. Amsterdam/Philadelphia: John Benjamins.

Sairio, A. 2017. *Bluestocking Corpus: Letters of Elizabeth Montagu, 1730s-1780s*. Helsinki: University of Helsinki (http://bluestocking.ling.helsinki.fi/).

Recursos de expansión dialógica en las sentencias del Tribunal Supremo de España y del Tribunal de Justicia de la UE: el caso de los Conectores Condicionales Complejos

Giovanni Garofalo

Dipartimento di Lingue, Letterature e Culture Straniere, Università degli Studi di Bergamo

***Palabras clave**: Sentencia judicial, español judicial de España, eurolecto judicial español, Conectores Condicionales Complejos, Teoría de la Valoración*

Abstract

La investigación que se propone apunta a analizar de manera contrastiva el uso de los conectores condicionales complejos (CCC) en las sentencias en español de dos altos tribunales, uno nacional, el *Tribunal Supemo* (TSE) y otro europeo, el *Tribunal de Justicia de la Unión Europea* (TJUE). En concreto, se pretende destacar la asimetría en la distribución de los CCC en los dos subcorpus que conforman el corpus monolingüe comparable de estudio (803.886 palabras el subcorpus del TSE y 611.488 palabras el del TJUE).

A partir de la clasificación de los CCC ofrecida por Montolío (1999: 3701-3710) –a saber, su diferenciación en *CCC afirmativos* (*a condición de (que), con tal (de) (que), siempre que, siempre y cuando*), *CCC negativos* (*a no ser que, a menos que, salvo que, excepto que*), *CCC con sustantivos creadores de mundos* (*en el caso/en la suposición/en el supuesto/en la hipótesis/en la eventualidad de que*) y *Prótasis de verbos creadores de mundos* (*suponiendo, imaginando, admitiendo que*)– se subraya la función tetxualizadora y el valor lógico-pragmático de estos nexos, que expresan condiciones excepcionales, mucho más circunstanciales y específicas con respecto al conector *si* (Montolío 1999: 3704; Montolío 2000, Visconti 2000; Garofalo 2006: 75-97).

Acudiendo a la herramienta *Sketch Engine*, se observará el índice de frecuencia normalizada de los CCC en ambos subcorpus, comparando los datos obtenidos con la distribución de tales conectores en un corpus de cotejo del español estándar. Los datos así recabados evidencian una frecuencia normalizada sensiblemente mayor de CCC en el español judicial y, más específicamente, su acusada prevalencia en el eurolecto del TJUE. Por otra parte, pese a la menor frecuencia normalizada de CCC en el subcorpus del TSE, en las sentencias de este órgano aparece un mayor número de tipos de CCC (16 tipos en las sentencias del TSE, frente a 13 tipos en las del TJUE). Considerando que las resoluciones del TJUE son textos traducidos del francés, la gama más reducida de CCC en el eurolecto puede relacionarse con la tendencia a la *estandarización*, ya observada en estudios previos (Garofalo 2018), y con la conocida *Hipótesis del ítem único (Unique Item Hypotesis*: "target language specific elements, which do not have equivalents in the source language, tend to be under-represented in translated texts, since 'they do not readily suggest themselves as translation equivalents" Laviosa 2009: 309, Tirkkonen-Condit 2004: 177-8).

Asimismo, las pertinentes relaciones de colocación y coligación (Sinclair 1996) revelan que, en la argumentación judicial, la mayoría de los enunciados regidos por CCC manifiestan el punto de vista del juez y expresan mandatos, prohibiciones, recomendaciones o condiciones necesarias y suficientes para otorgar permisos (Mazzi 2013).

Por último, a la luz de la Teoría de la Valoración (Martin y White 2005: 110-111), se recurre a la semántica del *compromiso* (*engagement*) para explicar la preponderancia de los valores modales (Montolío 1999: 3701) expresados por CCC en los textos del TJUE. De hecho, desde una perspectiva heteroglósica, tales valores son *recursos de expansión* del discurso y, en concreto, representan marcas de *consideración* (*entertain*), estrategia mediante la cual el juez europeo, al introducir condiciones, se muestra más proclive que su homólogo español a sopesar argumentos encontrados y a barajar múltiples escenarios, 'abriendo su texto' a posibles alternativas dialógicas.

Referencias

Garofalo, G., (2006): "I connettivi condizionali complessi nei testi normativi spagnoli", *RITT*, 9: 75- 97.

Garofalo, G., (2018): "La contraargumentación del Tribunal Supremo de España y del Tribunal de Justicia de la UE: hacia una interpretación funcionalista", *Cuadernos AISPI*, 12: 77-94.

Laviosa, S. (2009): "Universals", in Baker, M., Saldanha, G. (eds.), *Routledge Encyclopedia of Translation Studies*, London/New York: Routledge, pp. 306-310.

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

Mazzi, D. (2013): "If it be the case that the appellants are under such an obligation…": A comparative study of conditionals in English legal discourse, in Ondelli, S. (a cura di), *Realizzazioni testuali ibride in contest europeo. Lingue dell'UE e lingue nazionali a confront*, Trieste: Edizioni Università di Trieste, pp. 27-40.

Montolío, E. (1999): "Las construcciones condicionales", en Bosque, I., Demonte, V. (eds.),

Gramática descriptiva de la lengua española, vol. 3, Madrid: Espasa Calpe, pp. 3643-3737.

Montolío, E. (2000) "On affirmative and negative complex conditionatr connectives", in Elizabeth Couper-Kuhlen, E., Kortmann, B. (Eds.), *Cause - Condition - Concession – Contrast. Cognitive and Discourse Perspectives*, Berlin/ New York, Mouton de Gruyter, pp. 143-171

Sinclair, J. (1996): "The Search for Units of Meaning", *Textus*, 9, 1: 75-106.

Tirkkonen-Condit, S. (2004): "Unique Items – Over or Under- represented in Translated Language?", in Mauranen, A., Kujamäki, P. (eds.), *Translation Universals. Do they Exist?*, Amsterdam/ Philadelphia: John Benjamins, pp. 177-186.

Visconti, J. (2000): *I connettivi condizionali complessi in italiano e in inglese*, Alessandria: Edizioni dell'Orso.

Patterns of evaluation in the blurb genre: A corpus-linguistic approach

Maria-Lluïsa Gea-Valor

Department of English and German Philology, University of Valencia, Spain

***Keywords**: book blurb; evaluation; keywords; collocational patterns*

Abstract

Book blurbs have been defined as a mixed genre which fulfils both a descriptive and a promotional function, being widely used by publishing firms to market their products (Kathpalia 1997, Bhatia 2004, Gea-Valor 2005). In order to achieve their rhetorical goals, blurbs make abundant use of evaluative language aimed at praising the book and highlighting its positive features (Shaw 2006, Gesuato 2007, Cacchiani 2007, Bednarek 2014).

With the assistance of the concordancing software MonoConc Pro 2.2 (Barlow 2004) and the qualitative analysis software NVivo 10 (QSR 2012), the present paper explores the top keywords and most salient collocational patterns of evaluation in a corpus of 100 blurbs of fiction books from the New York Times Bestsellers List. A second objective is to analyse the distribution of the most common collocational profiles along the schematic structure of the genre, which comprises the following moves and submoves: *Move 1 Catchphrase > Move 2 Description > Move 3 Appraisal (Submove: Endorsement or Testimonial) > Move 2 & 3 Merger (Description + Evaluation) > Move 4 Author's credentials* (Gea-Valor & Inigo-Ros 2009).

The results indicate that the most key evaluative word in the corpus is the adjective beautiful, followed by moving and funny. These evaluative items tend to occur in predicate-intensifier collocations (Cacchiani 2004) consisting of an upgrading adverbial premodifying the adjective (e.g., *truly beautiful, brilliantly moving, hysterically funny*). Similarly, accumulation, that is, the sequencing of terms referring to various appraisal dimensions, especially in the form of adjective triads (e.g., *beautiful, gripping and deeply moving; fresh, funny and accomplished*), is used as a powerful rhetorical device to increase and intensify praise. With regard to move distribution, most of these evaluation-laden resources are found in the endorsement submove, followed by the move 2 & 3 merger.

All in all, it may be claimed that the lexico-grammatical choices displayed in blurbs of fiction books contribute to fulfilling the promotional function of the genre by making the book appealing to the public and by providing assessment along various semantic dimensions. Interestingly, these collocational profiles reveal the assumptions made by the publishing industry about the interests, tastes and expectations of the prospective audience.

References

Barlow, M. (2004). *MonoConc Pro 2.2*. Houston, Texas: Athelstan.

Bednarek, M. (2014). 'An astonishing season of destiny!': Evaluation in blurbs used for advertising TV series. In G. Thompson & L. Alba-Juez (Eds.), *Evaluation in context* (pp. 197-220). Amsterdam/Philadelphia: John Benjamins.

Bhatia, V. K. (2004). *Worlds of written discourse: A genre-based view*. London & New York: Continuum.

Cacchiani, S. (2004). Towards a model for investigating predicate-intensifier collocations. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX international congress* (pp. 943-947). Université de Bretagne Sud.

Cacchiani, S. (2007). From narratives to intensification and hyperbole: Promotional uses of book blurbs. In M. Davies, P. Rayson, S. Hunston, & P. Danielsson (Eds.), *Proceedings of the corpus linguistics conference*. University of Birmingham.

Gea-Valor, M. L. (2005). Advertising books: A linguistic analysis of blurbs', *Ibérica*, 10, 41-62.

Gea-Valor, M. L., & Inigo-Ros, M. (2009). On the dynamic nature of genre: A diachronic study of blurbs. In K. Hyland & G. Diani (Eds.), *Academic evaluation: Review genres in university settings* (pp. 199-216). Basingstoke: Palgrave Macmillan.

Gesuato, S. (2007). Evaluation in back-cover blurbs. *Textus*, XX (1), 83-102.

Kathpalia, S. S. (1997). Cross-cultural variation in professional genres: A comparative study of book blurbs. *World Englishes*, 16 (3), 417-426.

NVivo 10 (2012). http://www.qsrinternational.com/products_nvivo.aspx

Shaw, P. (2006). Evaluative language in evaluative and promotional genres. In G. Del Lungo Camiciotti, M. Dossena, & B. Crawford Camiciottoli (Eds.), *Variation in business and economics discourse: Diachronic and genre perspectives* (pp. 152- 165). Rome: Officina Edizioni.

La adaptación de extranjerismos en la lengua española a través de obras y corpus académicos actuales

David Giménez-Folqués

Departamento Filología Española, Universitat de València, España.

Palabras clave: *Adaptación de extranjerismos, corpus académicos, diccionarios académicos, lexicología, lexicografía.*

Abstract

En el año 2005, el *Diccionario panhispánico de dudas* propuso una gran cantidad de adaptaciones de extranjerismos, principalmente anglicismos, en un intento de gestionar y asimilar al sistema lingüístico español este caudal léxico. En este punto, la Real Academia Española, en colaboración con la Asociación de Academias de la Lengua Española, adaptó, según señalaron en sus obras, aquellos extranjerismos necesarios, es decir, aquellos que no disponían de un equivalente patrimonial en la lengua española. Esta posición parte de la clasificación que realizan estos organismos académicos en

extranjerismos necesarios y extranjerismos innecesarios. De este modo, *hobby* formaría parte de los extranjerismos innecesarios, ya que la lengua española dispone del equivalente "pasatiempo" y *jacuzzi*, por otro lado, aparecería como extranjerismo necesario, ya que no cuenta con ningún equivalente en la lengua meta, por lo que, en este caso, se propone la adaptación "yacusi".

Más de diez años después de la aparición de esta obra, tenemos la posibilidad de observar qué ha sucedido con estas propuestas de adaptación. Las últimas obras académicas, entiéndase la *Ortografía de la lengua española*, la *Nueva gramática de la lengua española* y la 23.ª edición del *Diccionario de la lengua española*, han realizado modificaciones sobre esta primera propuesta. Estas modificaciones se deben a varios factores, entre ellos destacan el uso que han empleado los hablantes de la lengua en la esfera oral y escrita y la posición de los medios de comunicación. Hay que tener en cuenta que las obras académicas se basan en corpus que recogen textos de todos los ámbitos, entre ellos entidades influyentes en la evolución de la lengua como los mencionados medios de comunicación, las obras literarias o internet, entre otros. Como consecuencia, observaremos una gran diferencia en la propuesta de adaptaciones que realizó el DPD con respecto a la inclusión de este tipo de voces que se han llevado a cabo en la última versión del *Diccionario*.

De esta manera, nuestro propósito es el de investigar la situación actual de estos extranjerismos y sus respectivas adaptaciones en las últimas obras académicas y, conjuntamente, los factores que han podido influir en cualquier modificación que se haya llevado a cabo. Concretamente, nos interesa investigar el mantenimiento de adaptaciones, su modificación y su exclusión entre las propuestas realizadas por la Real Academia en 2005 y su situación actual en las últimas obras académicas. Asimismo, el segundo foco tiene que ver con la aproximación a una posible relación entre la posición de entidades como los medios de comunicación e internet con las decisiones que han tomado los organismos académicos. Para analizar esta relación nos basaremos en el vaciado realizado en el corpus de referencia *Corpus del español del siglo XXI*, en el cual se recogen textos de diferente índole dentro del mundo panhispánico, entre ellos los mencionados en esta propuesta.

Referencias

REAL ACADEMIA ESPAÑOLA Y ASOCIACIÓN DE ACADEMIAS DE LA LENGUA ESPAÑOLA (2005): *Diccionario panhispánico de dudas*. Madrid: Santillana.

REAL ACADEMIA ESPAÑOLA Y ASOCIACIÓN DE ACADEMIAS DE LA LENGUA ESPAÑOLA (2010): *Nueva gramática de la lengua española*. Madrid: Espasa.

REAL ACADEMIA ESPAÑOLA Y ASOCIACIÓN DE ACADEMIAS DE LA LENGUA ESPAÑOLA (2010): *Ortografía de la lengua española*. Madrid: Espasa.

REAL ACADEMIA ESPAÑOLA Y ASOCIACIÓN DE ACADEMIAS DE LA LENGUA ESPAÑOLA (2014): *Diccionario de la lengua española*. Madrid: Espasa.

REAL ACADEMIA ESPAÑOLA: Banco de datos (CORPES XXI) [en línea]. *Corpus del español del siglo XXI.* <http://www.rae.es> [04/12/2018]

Vocabulary selection for didactic purposes: report on a machine learning approach

Patrick Goethals

Department of Translation, Interpreting and Communication, Ghent University, Belgium.

***Keywords**: vocabulary learning; difficulty levels; vocabulary selection; machine learning; Spanish*

Abstract

Research Question

We will contribute to one of the main discussions in the field of vocabulary learning (Nation 2016), namely the selection and grading of vocabulary items, considered as crucial but also complex steps in the design of didactic materials (Nation 2016, Vincze & Alonso Ramos 2015). It has become common practice to complement or substitute purely introspective methods by empirical and mainly corpus- based methodologies. Within a corpus-based methodology the most obvious parameter is frequency, the assumption being that the most frequent words in a corpus are also the most interesting or useful ones (see Davies 2006 for Spanish). Yet, it has also become clear that raw corpus frequencies should be handled carefully, and corrected, for example, by:

- improving the representativeness of the corpora, e.g. by building corpora that include a sufficiently wide variety of text types (Davies 2006), or by validating the representativeness of different corpora (Duchon et al. 2016);
- complementing the overall frequencies with data on the distribution or dispersion of words throughout the corpus (Nation 1991, Davies 2006, Gries 2008);
- taking into account cognate effects between words in the target language and the mother tongue of the students (Izquierdo Gil 2005);
- critically evaluating the outcome of the empirical selection procedures by taking into account the intuition of experienced teachers or didactic authors (Instituto Cervantes).

We will take these insights as a starting point to explore the question that inevitably follows the identification of these possibly relevant factors, namely how these factors can be calibrated and combined into one single selection procedure. In other words, we will try to evaluate on an empirical basis which (combinations of) criteria have the highest predictive power.

Methodology

We propose an experimental machine learning approach, in which we will (a) define the dependent variable (the difficulty level of the lexical items), by collecting students' evaluations of this difficulty level, (b) gather data representing frequency in different corpora, dispersion between corpora, cognateness and independently assigned difficulty levels, and, finally, (c) evaluate the accuracy of different machine learning systems that are trained on these data. We apply the SFFS feature selection method

(Sequential Forward Floating Selection) to let the algorithms find a minimal subset of features that also maximize prediction performance. Also, we compare the results of a linear regression model and a decision tree model.

Results

The SFFS method yields promising results. The best model obtains a relatively low Mean Average Error (MAE) and a relatively high Pearson Correlation Score, which means that the automatically assigned difficulty level highly correlates with the students' evaluations. With respect to the difference between linear regression and decision tree, the conclusion is that decision tree works better than linear regression.

Discussion

In the discussion of the results, we will also show how the logarithms that were developed can be integrated in an on-line vocabulary learning method. These logarithms can guide both teachers and autonomous learners in their fascinating journey through the infinite world of lexis, avoiding that they feel as "amateur fishermen in the middle of the ocean" (Santos Palmou 2016: 166, our translation).

Bibliography

Davies, M. (2006). *A frequency dictionary of Spanish: Core vocabulary for learners*. New York: Routledge.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. , 2013, vol. 45, no 4, p. 1246-1258. *Behavior research methods*, *45*(4), 1246-1258.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*, 403-437.

Izquierdo Gil, M. d. C. (2005). *La selección de léxico en la enseñanza del español como lengua extranjera. Su aplicación al nivel elemental en estudiantes francófonos*. Málaga: ASELE Colección Monografías.

Nation, P. (1991). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, P. (2016). *Making and Using Word Lists for Language Learning and Testing*. John Benjamins. Santos Palmou, X. (2016). La selección del vocabulario en ELE: estado de la cuestión y nuevas metodologías. In El español como lengua extranjera en Portugal II: retos de la enseñanza de lenguas cercanas. In *El español como lengua extranjera en Portugal II: retos de la enseñanza de lenguas cercanas* (pp. 164-178). Subdirección General de Documentación y Publicaciones.

Vincze, O., & Alonso Ramos, M. (2013). Incorporating frequency information in a collocation dictionary: Establishing a methodology. *Procedia-Social and Behavioral Sciences, 95*, 241-248.

Python from the Tree of Knowledge: How to Study a Speech Corpus Effectively

Daria Gorbunova

Saint Petersburg State University, Russia

***Keywords**: Python; speech corpus; colloquial speech; ORD*

Abstract

The report describes the technology of processing and accounting based on the «One Day of Speech» Corpus (ORD) material. ORD aims to study the speech behavior of a native speaker during the day (using the technique of 24-hour recording[1]) and nowadays it is the largest linguistic resource for studying Russian oral discourse [1]. However, a lot of work in corpus are entirely processed by hand. The report presents an attempt to summarize many years of experience with corpus material and describes a minimal set of tools to start automatic processing based on the specific properties and related materials of the speech corpus. For such a large amount of structured data (speech corpus decryptions), we decided to use Python scripts to get ready-to-analyze linguistic material. Processing materials includes:

- count the total number of words of informants and communicants, excluding all marks and decoding signs;
- divide informants into extroverts and introverts according to psychological test results;
- search by lexemes;
- provide complete list of replicas for a certain informant / communicant;
- search by word combinations and provide context.

To count, systematize and automate all the available corpus material, we created a special program in Python with the following basic requirements:

1) calculate the total number of word usage in the speech of introverts and extroverts;
2) provide a convenient and simple expression search which provides an extended context;
3) use the following data:
   - several files with decryptions with all replicas of the informants and communicants, their code, and decryption signs;
   - a file with data about informants` psych type;
4) exclude:
   - words of communicants;
   - words of informants for whom we have no information about psychotype;
   - decoding marks;
   - time sheets;
5) provide sorted processed data.

Processed data allows to evaluate the quantitative component in the speech of ORD informants and to identify the group of the most representative speakers. The results

and methods of automatic processing open many new opportunities for studying Russian spontaneous oral speech in various aspects. The developed tool can be used in UNIX without knowledge of any programming languages and it can provide full clean text data ready for analysis. Further manual processing of clean text with psych type provided per replica is fast and convenient as opposed to Word or Excel, which cannot deal with huge data sets.

References

Bogdanova-Beglarian, N., Martynenko, G., Sherstinova, T. The «One Day of Speech» Corpus: Phonetic and Syntactic Studies of Everyday Spoken Russian // SPECOM 2015, Lecture Notes in Artificial Intelligence, LNAI, vol. 9319. – Springer International Publishing Switzerland, 2015. – Pp. 429-437.

Shibata, S. Japan's Bilingualism and Social Issues // The Journal of Popular Culture XVII,1983. – Pp 111-119.

To not let it happen or not to let it happen? Corpora-based analysis of negative infinitive alternation in discourse

Marina Gorlach

English Department, Metropolitan State University of Denver, USA

Keywords: *negative infinitive, discourse role, literary analysis, corpus-based*

Abstract

The cognitive complexity of the negative meaning underlies the various lexical, syntactic, and pragmatic ways of expressing it in English. This paper discusses the structural expressions of the negative infinitive in various types of discourse. The focus is on the alternation between the two forms of the negative infinitive, not to VERB vs. to not VERB, as they are distributed across written and spoken texts of different genres. Only one of them, not to VERB, is recognized as existing in English by grammarians. The corpus data, however, demonstrate that the so-called split negative infinitive is actively and more frequently used in certain types of discourse.

This is a corpus-based study making the connection between the forms and meanings of the negative infinitive constructions themselves, as well as in relation to genre, register, and discourse situations. The paper presents the comparative frequency of using each form in discourse and discusses the pragmatic implications and applications of the observed distribution.

As per COCA database, the frequency of to not VERB constructions is significantly higher in the spoken register (13.33 per million) as opposed to academic texts (4.10 per million). The newspaper genre usage is second high (7.02 per million), while the frequency in fiction and magazine genres is similar to academic texts (4.24 and 4.45 per million respectively).

Chronologically, the frequency of to not VERB shows a steady growth each decade since

1990, rising from 4.56 per million quotations in 1990-1994 to 9.59 per million in 2010-2015.

The paper explores the relationship between the form and communicative function, as well as the role of the additional lexical and non-lexical devices in generating the negative meaning. Theoretically and methodologically, this analysis relies on the concept of markedness and non-random distribution, treating split negative infinitive as a marked item in discourse. Since the two negative infinitive constructions co-exist, the speakers are presented with a subconscious choice, which is contextually motivated.

The non-synonymy assumption claims that different signals/forms indicate different meanings/messages, however subtle such a difference may be, and that discourse situations and communicative goals serve as a motivating factor.

(1) KING: Bob Grant, what can you tell us? GRANT: Well, I would just caution folks not to lose confidence. (CNN, 1997)

(2) After 9/11, U.S. President Bush asked Americans to carry on with their lives, to not lose confidence, and to continue spending. (Murray, 2013)

The study takes a closer look at specific contexts and meanings of the negative infinitive considering the criteria that go beyond the formal linguistic aspects of language, such as the communicative situation and interplay between the interlocutors. As follows from the corpus-based analysis, one of the constructions is significantly less frequent overall, but has shown dominance in certain types of discourse and increase in popularity over the last decade.

References

Fauconnier, G., & Turner, M. (2003). The way we think: Conceptual blending and the mind's hidden complexities. Basic Books.

Givón, T. (2005). Context as other minds: The pragmatics of sociality, cognition and communication. Amsterdam/Philadelphia: John Benjamins.

Napoli, E. (2006). Negation. Grazer Philosophische Studien 72(1), 233-252

Tobin, Y. (1990). Semiotics and linguistics. London: Longman.

Translation patterns of recurrent n-grams in English-to-Polish translation: a corpus-informed study

Łukasz Grabowski[a]

[a]*Department of Translation Studies and Lexicography, University of Opole, Poland;*

***Keywords**: recurrent n-grams, parallel corpus, English-to-Polish translation, textual fit, discoursal functions*

Abstract

In this descriptive study, we will use both parallel and comparable reference corpora in the English-Polish language pair to explore, first, how translators have dealt with recurrent multi-word items performing specific discoursal functions; second, whether the observed tendencies overlap with the ones found in native texts originally written in contemporary Polish; and third, our aim is to explore to what extent the discourse functions realised by n-grams under scrutiny are "preserved" in translation. By extending an earlier preliminary study (Grabowski 2018), we will focus on a set of pre-selected recurrent multi-word items (n-grams) that can be divided, given their discoursal functions, into two groups, namely those signaling stance-taking on the one hand, and those functioning as textual, discourse-structuring devices on the other. In other words, we aim to explore whether and to what extent English recurrent n-grams with specific discoursal functions reflect the same level of formulaicity and regularity in Polish translations and native Polish texts.

More precisely, using selected recurrent n-grams performing discourse-organizing functions and/or expressing stance and found in various text types and genres collected in Paralela corpus (Pęzik 2016), a parallel English-Polish and Polish-English translational corpus, I will explore the use – largely in terms of frequency distributions – of the Polish equivalents of selected English multi-word items and – using the National Corpus of Polish (NKJP) - check whether the Polish equivalents, which may include single and multi-word units, are typical of contemporary Polish as found in native texts. The latter aspect is essential since – as argued by Kranich (2016:10), apart from cultural differences and interference from the source language "a tendency to „say what seems normal or safe‟ should be also kept in mind as a potential explanation for differing behaviour of translated texts compared to the source and target language originals". The unit of analysis are recurrent n-grams performing discourse-organizing functions and expressing stance. The linguistic items were randomly selected from among the ones identified by Granger (2014) and Grabar and Lefer (2015) as English lexical bundles in the European parliament debates. Then, eight recurrent n-grams were randomly selected, four expressing attitudinal stances and four expressing discourse-organizing functions with the caveat that they occur at least 20 times in the EPP subcorpus (i.e. on average 1 per 1,000 translation segments) and at least 80 times in the entire Paralela corpus. This is to ensure that the recurrent items provide sufficient data for both quantitative and qualitative analyses. Some of the preliminary results (Grabowski 2018) revealed a high number of Polish equivalents, both single- and multi-word units, expressing attitudinal and epistemic stance; it was also reported that sometimes entirely different sentence fragments in the original and in the translation conveyed attitudinal and epistemic stance; finally, the results obtained from NKJP corpus revealed a number of Polish equivalents which are very rare or do not occur in native (i.e. non-translated) Polish texts. In the course of the study a number methodological problems were also addressed, as they are inevitable when dealing with recurrent phraseologies in typologically different languages (such as English and Polish) or extracting recurrent phraseologies from specific parallel and comparable (reference) corpora.

The motivation to pursue a study like this one is manifold. One can find a number of studies of various types of recurrent n-grams conducted from a cross-linguistic perspective, e.g. comparisons of native texts originally produced in two languages (e.g. Cortes 2008; Forchini & Murphy 2008; Granger 2014) or comparisons of native versus translational texts in the same language (e.g. Baker 2004; Lee 2013; Ebeling & Ebeling 2018). However, such functional items have been rarely explored from the perspective

of translation, i.e. from the viewpoint of source texts and target texts in two languages. Notable exceptions are the studies conducted by Berūkštienė (2016) and Noreika and Seskauskiene (2017), who looked at English-to-Lithuanian translation patterns yet focusing more on structural than functional aspects. Also, Grabar and Lefer (2015) also argue that terminological databases used by translators rarely, if ever, include recurrent multi-word sequences that express writer‟s stance or structure texts. This observation provided another motivation to pursue this research. Hence, I believe that it is worthwhile exploring in a more comprehensive manner whether recurrent multi-word sequences described by corpus linguists (e.g. n-grams, lexical bundles, phrase frames) performing discourse-organizing functions and expressing stance - and their textual equivalents in a target language - should be included into translation memories or terminological glossaries and whether they could help enhance the textual fit (Chesterman 2004) of translated texts.

References

Berūkštienė, D. (2017). "A corpus-driven analysis of structural types of lexical bundles in court judgments in English and their translation into Lithuanian". *Kalbotyra*, 70: 7-31.

Chesterman, A. (2004). Hypothesis about translation universals. In: G. Hansen, K. Malmkjaer & D. Gile (Eds), *Claims, Changes and Challenges in Translation* Studies. Amsterdam: John Benjamins, 1–13.

Ebeling, Jarle & Signe O. Ebeling (2018). "Comparing n-gram-based functional categories in original versus translated texts". *Corpora* 13:3, 347-370.

Forchini, P. & Murphy, A. (2008). N-grams in comparable specialized corpora. Perspectives on phraseology, translation and pedagogy. *International Journal of Corpus Linguistics*, 13(3), 351–367.

Grabowski, Ł. (2018). "Stance bundles in English-to-Polish translation: a corpus-informed study". *Russian Journal of Linguistics*, 22 (2), 404-422

Granger, S. (2014). "A lexical bundle approach to comparing languages. Stems in English and French". *Languages in Contrast*, 14 (1): 58-72.

Grabar, N. & Lefer, M-A. (2015). "Building a lexical bundle resource for CAT and MT". Presentation delivered at Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT2015) of EUROPHRAS 2015. 29 Jun-1 Jul 2015, Malaga, Spain.

Kranich, S. (2016). *Contrastive Pragmatics and Translation: Evaluation, epistemic modality and communicative styles in English and German*. Amsterdam: John Benjamins.

Lee, C. (2013). "Using lexical bundle analysis as discovery tool for corpus-based translation research". *Perspectives* 21 (3), 378–395.

Noreika, M. & Seskauskiene, I. (2017). "EU Regulations: Tendencies in Translating Lexical Bundles from English into Lithuanian". *Vertimo Studijos*, 10, 156-174 [DOI: 10.15388/VertStud.2017.10.11302]

Oksefjell Ebeling, S. & Ebeling, J. (2017). "A Cross- Linguistic Comparison of recurrent word combinations in a comparable corpus of English and Norwegian

Fiction". In: M. Janebova, E. Lapshinova-Koltunski & M. Martinkova (Eds), *Contrasting English and Other Language through Corpora*. Newcastle upon Tyne: Cambridge Scholars Publishing, 2-31.

Pęzik, P. (2016). "Exploring phraseological equivalence with Paralela". In: E. Gruszczyńska & A. Leńko-Szymańska (Eds), *Polish-Language Parallel Corpora*. Warszawa: Instytut Lingwistyki Stosowanej UW, 67-81.

Acknowledgements

This research has been funded by the Polish National Agency for Academic Exchange (NAWA) under the agreement no: PPN/BEK/2018/1/00081/

Parallel Corpus Examples for Language Learning Applications

Johannes Graën[ab]

[a]*Institutionen för svenska språket, Göteborgs universitet, Sweden;* [b]*Departament de Traducció i Ciències del Llenguatge, Universitat Pompeu Fabra, Spain*

***Keywords**: parallel corpora, word alignment, language learning, data-driven learning*

Abstract

The use of corpora has proven beneficial to language learners. Cobb and Boulton (2015) performed a meta-analysis of corpus use studies and concluded that "corpus-based learning is more efficient than traditional treatments" (p. 491). This can be seen as a direct response to (Meunier 2011), who holds the "lack of empirical studies exploring the actual impact of corpus methods on the learning outcomes" (p. 463) responsible for the little use that corpora saw (and probably still see) in language learning, especially in classroom contexts. Cassany (2016) investigates how technology in language learning is received by college students and concludes that they prefer popular interfaces that are easy to use, which are mostly automatic translators, electronic dictionaries and spell checkers.

Pilán, Volodina and Borin (2016) address the question, which properties make a corpus example adequate for language learners to present them "pedagogically more relevant results". Their notion of good corpus examples follows the idea of 'Good Dictionary Examples' by (Kilgarriff et at, 2008) to illustrate the contextual use of lexical units. Good corpus sample sentences for learners are context independent and match the respective learner's proficiency in terms of vocabulary and syntactical complexity.

Parallel corpora have seen many uses besides serving as training material for machine translation algorithms. We see a great potential in using parallel sentences as scaffolding for language learning applications. Similar to good monolingual examples, parallel corpus examples (i.e. pairs of sentences in two different languages) should be context-independent and match the learner's proficiency. Additionally, we want the sentences to be good translations of each other, so that learners can acquire language skills by exploring authentic language examples on their own, following the concept of data-driven learning (see Boulton 2017).

We have previously worked on a large parallel corpus with several layers of annotation

and alignment (Graën 2018). Word alignment, in particular, has revealed many more use cases than we had considered in the first place. In this work, we make use of word alignment to estimate the complexity of translation. We expect closer translations to be more helpful for learners. Structurally divergent translations of a particular part (e.g. a single-word verb that is translated with a relative clause) may, however, also be of interest in particular learning situations.

We combine this measure with measures of syntactical complexity based on syntactic relations, distributions of part of speech and information on CEFR levels of the respective content words. With this combination, we classify all parallel sentences for a selection of language pairs in our corpus. The resulting lists consist of the respective sentence pair, values for the aforementioned measures, an overall rating on the CEFR level scale and suggestions for which lexical units we consider the respective example adequate and helpful. In a future step, we plan to evaluate our corpus examples in a crowsourcing task with language learners and teachers.

References

Boulton, A. (2017). Data-Driven Learning and Language Pedagogy. *Language, Education and Technology*, 1–12.

Cassany, D. (2016). Recursos lingüísticos en línea: Contextos, prácticas y retos. *Revista Signos*, *49*, 7–29.

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 478–497). Cambridge University Press.

Graën, J. (2018). *Exploiting Alignment in Multiparallel Corpora for Applications in Linguistics and Language Learning* (PhD Thesis). University of Zurich.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlỳ, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In J. D. Elisenda Bernal (Ed.), *Proceedings of the 13th EURALEX International Congress*.

Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Meunier, F. (2011). Corpus linguistics and second/foreign language learning: exploring multiple paths. *Revista Brasileira de Linguí stica Aplicada*, *11*(2), 459–477.

Pilán, I., Volodina, E., & Borin, L. (2016). Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *Traitement Automatique Des Langues*, *57*(3), 67–91.

CORDIA-VEN-s. XX: un subcorpus dialectal para el estudio sociolingüístico del español hablado en Venezuela

Krístel Guirado

Universidad Central de Venezuela; Universidad de Zaragoza

***Palabras clave**: Reingeniería de Corpus, habla urbana en Venezuela, microdiacronía, corpus sociolingüístico, Lingüística de Corpus.*

Abstract

La definición de *corpus* propuesta por Parodi (2008) es una de las más útiles en el ámbito de las investigaciones que tienen como propósito el desarrollo, diversificación e integración de muestras de habla, ya que advierte sobre la pertinencia que tienen los metadatos en el tratamiento futuro del corpus: "debe cumplir con aportar detalles relevantes acerca de su recolección y procedencia" (106). Según Guirado (2015), la *Reingeniería de Corpus* incluye el proceso y las tareas de reconfiguración de materiales de habla (orales y escritos), recopilados y estructurados en diversas bases y cuerpos de datos cuando su productividad se torna vulnerable, para su reutilización en diversos ámbitos. De acuerdo con el criterio de representatividad, los corpus pueden ser objeto de diversas tipologías. Atkins, Clear y Ostler (1992) diferencian entre *corpus* y *subcorpus*. Para Sinclair (1996) y Torruella y Llisterri (1999), un subcorpus es cualquier porción seleccionada de un corpus mayor. No obstante, para Pearson (1998), los corpus especializados no constituyen un subcorpus porque la selección no tiene todas las propiedades del corpus mayor, razón por la cual prefiere identificar con el término *corpus de propósito especial* (*especial purpose corpus*) toda selección de materiales que responde a un objetivo específico de investigación, el cual motiva y le da sentido a su creación. El objetivo de la presente investigación fue crear un subcorpus para el estudio dialectal del habla venezolana en una *microdiacronía* del siglo XX (Pons Bordería 2014), a partir de la reingeniería de tres corpus orales de esta variedad, recogidos y transcritos durante la década 1986-1996: *Corpus sociolingüístico de Maracaibo 1986* (cf. Chela-Flores y Gelman 1988), *Corpus sociolingüístico de Caracas 1987* (cf. Bentivoglio y Sedano 1993) y el *Corpus sociolingüístico de Mérida* 1995 (Cf. Domínguez y Mora 1998). Esta idea tiene su antecedente en el proyecto *Gramática del Español Hablado en Venezuela*, GREHV (cf. Bentivoglio 1996). La metodología para el rediseño incluyó: i. describir la estructuración de los corpus originales; ii. determinar los aspectos quebrantados, los sólidos y los puntos de intervención de cada arquitectura; iii. evaluar los metadatos de las entrevistas para una posible recodificación; y, iv. crear el nuevo diseño y estimar su representatividad. La estratificación social de los corpus originales facilitó la creación de una nueva arquitectura para los materiales. Como resultado se obtuvo el *Corpus dialectal para el estudio sociolingüístico del español hablado en Venezuela a finales del siglo XX* (CORDIA-VEN-s. XX), el cual consta de 144 hablantes distribuidos equitativamente por grupo etario (1: 14-29 años; 2: 30-45 años; 3: 46-60; 4: 61 años o más), nivel socioeconómico (alto, medio, bajo), sexo y ciudad (Maracaibo, Caracas, Mérida). Su representatividad garantiza la proyección de algunas tendencias de uso y, en algunos casos, el análisis exhaustivo de fenómenos del lenguaje asociados a factores socio-culturales específicos, ya que su proporción permite hacer generalizaciones e incluso acercamientos desde algunos modelos estadísticos inferenciales. Esto confirma que un corpus sociolingüístico debidamente estratificado, transcrito y digitalizado se puede instrumentalizar en otra base de datos con fines y enfoques similares. Se concluye que la *Reingeniería de Corpus* constituye un mecanismo eficaz de la Lingüística de Corpus para la diversificación y reincorporación

de los materiales en nuevos contextos interpretativos para la comunidad lingüística.

Referencias

Atkins, S., J. Clear y Ostler, N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing*, 7(1), 1-16.

Bentivoglio, P. y Sedano, M. (1993). Investigación sociolingüística: sus métodos aplicados a una experiencia venezolana. *Boletín de Lingüística*, 8, 3-35.

Bentivoglio, P. (1996). La gramática del español hablado en Venezuela. *Lengua y habla*, 1(1), 21-39.

Chela-Flores, B. y Gelman, J. (1988). *El habla de Maracaibo: Materiales para su estudio*, Material mimeografiado, Maracaibo: Universidad del Zulia.

Domínguez, C. L. y Mora, E. (1998). *El habla de Mérida*, Mérida: Universidad de Los Andes.

Guirado, K. (2015). *Reingeniería de Corpus* en Venezuela: una propuesta metodológica para diversificar el análisis de los corpus del español hablado en Caracas. *Letras*, 92, 55-87. (http://revistas.upel.edu.ve/index.php/letras/article/view/5962/3239).

Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito. *Revista de lingüística teórica y aplicada*, 46(1), 93-119.

Pearson, J. (1998). *Terms in Context. Studies in Corpus Linguistics* 1.

Amsterdam/Philadelphia: John Benjamins.

Pons Bordería, S. (2014). El siglo XX como diacronía: intuición y comprobación en el caso de *o sea*. *RILCE*, 30(3), 985-1016.

Sinclair, J. (1996). Preliminary recommendations on Corpus Typology. EAG-TCWG-CTYP/P", en *EAGLES.*

(http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html)

Torruella, J. y Llisterri, J. (1999). Diseño de corpus textuales y orales. En J. M. Blecua,

G. Clavería, C. Sánchez y J. Torruella (eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos*, 45-77. Barcelona: Milenio. (http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf)

Perspectiva diacrónica contrastiva del Paisaje Lingüístico de inmigración en el barrio de Russafa a partir de un corpus fotográfico

Ana Gómez-Pavón Durán

University of Valencia

Palabras clave: *Paisaje Lingüístico, PL, Paisaje Lingüístico de inmigración, estudio diacrónico contrastivo, fuentes digitales, corpus fotográfico, Paisaje Lingüístico de Russafa, diacronía en Paisaje Lingüístico*

Abstract

Este estudio es parte de un trabajo de fin de máster actualmente en desarrollo y presenta un análisis diacrónico del Paisaje Lingüístico (PL) de inmigración en el barrio de Russafa de la ciudad de Valencia que se ha llevado a cabo por medio de un estudio diacrónico contrastivo a partir de fuentes digitales recogidas en un corpus fotográfico. El tema de este trabajo surge de la necesidad de conocer el impacto del proceso de gentrificación del barrio en el uso de las lenguas de inmigración (principalmente chino y árabe) y del castellano desde el año 2008 hasta la actualidad. Su objetivo es contrastar las diferentes relaciones que se establecen entre las lenguas que son objeto de este estudio, así como analizar la evolución de dichas relaciones en los últimos diez años y determinar en qué medida estos datos revelan las actitudes lingüísticas de la población del barrio de Russafa en el pasado y en el presente. Con este análisis se pretende también demostrar el reflejo de la decreciente presencia de las lenguas de inmigración en el PL de Russafa. Para elaborar el marco teórico de este estudio ha sido necesario obtener datos acerca de la población inmigrante de Russafa de la Oficina de Estadística del Ayuntamiento de Valencia (2016) y de los estudios de Torres (2006) y de Romero y Lara (2015) de este barrio. Algunos estudios previos sobre PL como los de Backhaus (2007) y Pons (2012), junto con el que hace Ma (2017) también de una zona de la ciudad de Valencia, entre otros, han servido de modelo para determinar qué aspectos son relevantes en un estudio de PL o en qué medida algunos factores como las traducciones de los signos o la prominencia visual de ciertas lenguas son un reflejo de las actitudes lingüísticas de la sociedad estudiada. Asimismo, se ha tomado la metodología de estudio y la clasificación de signos que se lleva a cabo en algunos trabajos, como el de Comajoan (2013) o Piotr (2016), entre otros, como modelo para este trabajo. Para el estudio de signos lingüísticos ha sido necesario recurrir a fuentes digitales y para ello se ha elaborado un corpus integrado por fotografías de rótulos y signos privados de algunos negocios de este barrio y por fotografías obtenidas del programa Google Street View para los signos del año 2008. Las fotografías contienen muestras de lenguas sobre las que se ha realizado el análisis de este trabajo. Los resultados de este estudio reflejan no solo el descenso en la presencia de las lenguas de inmigración en el PL de Russafa, sino también la creciente importancia que han ido adquiriendo en los últimos diez años el español y algunas lenguas de prestigio como el inglés. Este fenómeno se puede apreciar en el análisis de algunos aspectos como la combinación de lenguas, las traducciones y la prominencia visual y la preferencia de código en los signos estudiados; ya que es posible observar cómo los signos de las fotografías del año 2008, en los que la presencia de las lenguas de inmigración era mayor, pasan a ser signos en los que el español es la lengua predominante en el año 2018.

Referencias

AYUNTAMIENTO DE VALENCIA (2016): «Población de nacionalidad extranjera en la ciudad de Valencia: Estudio del Barrio de Russafa», Oficina de Estadística Ayuntamiento de Valencia.

BACKHAUS, Peter (2006): «Signs of Multilingualism in Tokyo: A Diachronic Look at the Linguistic Landscape», *International Journal of the Sociology of Language*, vol. 2005, n.º 175/176, pp. 103-121.

COMAJOAN, Llorenç (2013): «El paisaje lingüístico en Cataluña: caracterización y percepciones del paisaje visual y auditivo en una avenida comercial de Barcelona», *Revista internacional de lingüística Iberoamericana,* vol. 11, n.º 1 (21), pp. 63-88.

DE LUNA VILLALÓN, María Eugenia (2013): «El uso del español en el paisaje lingüístico de una pequeña ciudad Canadiense», *Estudios de lingüística aplicada*, vol. 31, n.º 58.

MA, Yuying (2017): «El paisaje lingüístico chino-español de la ciudad de Valencia: una aproximación a su estudio». *Lengua y migración/Language and Migration*, vol. 9, n.º 1, pp. 63-84.

MUÑOZ CARROBLES, Diego (2013): *Lenguas y culturas en contacto en contexto urbano: el caso de la comunidad rumana de Madrid*, Madrid, Universidad Complutense.

PIOTR ROSA, Krzysztof (2016): «La influencia del turismo en el paisaje lingüístico de la ciudad de Valparaíso: un estudio empírico», Radboud Universiteit.

PONS RODRÍGUEZ, Lola (2012): *El paisaje lingüístico de Sevilla. Lenguas y variedades en el escenario urbano hispalense*, Sevilla, Diputación de Sevilla.

TORRES PÉREZ, Francisco (2006): «Las dinámicas de la convivencia en un barrio multicultural. El caso de Russafa (Valencia)», *Papeles del CEIC, International Journal on Collective Identity Research*, n.º 1.

A computer-assisted approach to identifying terrorism frames in Chinese media

Jingxuan Guo

Linguistics and Literary Studies (LIST), Vrije Universiteit Brussel, Belgium

***Keywords**: corpus linguistics, media representation, framing theory, cluster analysis, terrorism*

Abstract

Media frames refer to discursive constructions of news narrative, which can be seen as structures of meaning, involving central ideas and organizing principles. As such, media frames involve elements of 'selection' and 'salience' (Entman,1993), which may affect how audiences perceive reality. Thus, identifying frames used by media is key to studying the nature and possible societal impact of media discourse.

However, previous studies on deriving media frames have two shortcomings. First, the majority of the studies adopt qualitative methods by manually coding small samples, rather than adopting a quantitative approach based on large corpora. Moreover, quantitative studies still rely on manual coding to a greater extent

(Mathes&Kohring,2008). Second, quantitative framing research is oftentimes limited to an identification of general themes. Our study, in a final step, will move beyond this and look deeper into the argumentative structure of the (issue-specific) frames.

To fill those gaps, in the first step, we will attempt to use a quantitative approach, namely cluster analysis, to explore the extent to which we can automatically detect and identify frames based on a relatively large data set. More specifically, using complete news articles as the research object, we will cluster articles based on k-means algorithm (Burscher et al., 2016). Next, we will first adopt a deductive approach to list a priori defined news frames used by previous studies, and then classify them into two categories (De Vreese, et al,2001):

1) Generic frames are suitable to all kinds of specific news events with different topics, times, and cultural contexts, mainly including conflict, human interest, attribution of responsibility, and economic consequences frames (Semetko&Valkeburg,2000). Generic frames can also be classified into episodic and thematic frames (Iyengar,1991).

2) Issue-specific frames are tied to a particular topic, such as war on terror frames. Finally, we will further analyse and elaborate the findings of the cluster analysis by a combination of deductive and inductive approaches.

In this study, we used the topic of terrorism as an example to measure whether our method is feasible. We extracted 6518 texts which include the keywords "恐怖主义[terrorism]", "恐怖 事件 [terrorist attack(s)]", or " 恐怖分子 [terrorists]" from the Dynamic Circulating Corpus (DCC) for 2015. DCC is a corpus collected by the Institute of Language Monitoring and Social Computation at Beijing Language and Culture University and includes text data of the 18 most popular Chinese newspapers from 2002 till the present. As we took the complete news texts as our object, after preprocessing (including segmentation), we created a document-term matrix based on the weight of token. Then, k-means clustering was used to cluster those texts into 8 clusters using the 'elbow method'. Next, those clusters can be summarized as 5 issue-specific frames, namely narrative of attacks, description of agent, international relations and war, economic issues, and security, and then assigned them into 4 generic frames, namely conflict, morality, economic consequences, and human-interest frames. Additionally, it is worth mentioning that those cluster can also fall into thematic and episodic frames: the former one refers to putting terrorism into a wider context, such as politics, war; the later one focuses on reporting on specific terrorist attacks.

References

Burscher, Bjorn & Vliegenthart, Rens & De Vreese, CH, 2016, Frames beyond words: applying cluster and sentiment analysis to news coverage of the nuclear power issue, Social Science Computer Review, 34:530-545.

De Vreese, CH., Peter, Jochen, Semetko, HA, 2001, Framing politics at the launch of the Euro: a cross-national comparative study of frames in the news, Political communication, 18:107-122.

Entman, RM, 1993, Framing: towards clarification of a fractured paradigm, Journal of Communication 43(3):51-58.

Iyengar, Shanto, 1991, Is anyone responsible?: how television frames political issues, University of Chicago Press: Chicago and London.

Matthes, Jorg & Kohring, Matthias, 2008, The content analysis of media frames:

toward improving reliability and validity, Journal of communication, 58:258-279.

Semetko, HA & Valkenburg, PM, 2000, Framing European politics: a content analysis of press and television news, International communication association,93-109.

Un protocolo para la transcripción y revisión de un corpus oral de conversaciones coloquiales: el Corpus Val.Es.Co. 2.0

Alejandra Hernández Ruiz

Filología Española, Universitat de València, España

***Keywords**: corpus Val.Es.Co., transcripción, etiquetado, conversación coloquial, protocolo, sistematización*

Abstract

En el seno del Grupo Val.Es.Co. (Universitat de València) y, concretamente, en el marco del Proyecto UDEMADIS (FF120216-77841-P), se está llevando a cabo un proceso de revisión y ampliación del corpus que dicho grupo ha compilado desde su fundación y que está recogido en formato papel (Briz y Grupo Val.Es.Co., 2002) y en formato digital (Cabedo y Pons, 2013).

La recopilación de un corpus oral, que recoja los intereses de todos los miembros de un grupo de investigación, requiere de una toma de decisiones metodológicas y de la aplicación de estas a la (re)construcción del mismo. Así, nos encontramos ante la necesidad de establecer un protocolo para, por una parte, revisar las conversaciones ya incluidas en el Corpus 2.0 (Cabedo y Pons, 2013) y, por otra, sistematizar las que se vayan incluyendo en este. Para ello, ELAN se erige como el programa principal de trabajo del grupo en la actualidad.

Por tanto, los objetivos de esta comunicación son (1) reflejar los pasos seguidos para filtrar y revisar las transcripciones ya recogidas, a las que se incorporan las nuevas convenciones del grupo, (2) presentar el proceso desde que se graban hasta que se suben a la página web las nuevas interacciones coloquiales, (3) dar cuenta de las peculiaridades

–permisos de grabación, formatos de audio, conversión de signos a etiquetas, alineación del texto con el audio– y programas informáticos –ELAN, Audacity, Exmaralda o Textpad, entre otros– requeridos para los dos puntos anteriores, y (4) mostrar los documentos elaborados para el seguimiento de esta fase unificadora –fichas técnicas, hojas de verificación y tablas de datos sociolingüísticos–.

Para ello, en primer lugar, se ha obtenido una visión global de la situación que presentan los corpus tanto escritos como orales (Torruella y Llisterri, 1999; Briz, 2005; Briz, 2012; Briz y Albelda, 2009). De las referencias extraídas, se ha acudido a algunos de ellos para tomar ideas sobre metodología y disposiciones prácticas. Es el caso del

corpusESLORA (Recalde y Vázquez, 2009; Vázquez, 2014), con quienes se comparten algunos objetivos en cuanto al diseño y protocolo de transcripción, o del corpus AMERESCO (Briz, 2016), del que se toman las etiquetas incorporadas en esta nueva etapa.

Como resultado de este procedimiento, contamos ahora con una plantilla de ELAN en la que incorporar las transcripciones de las que disponemos en Word, así como en la que realizar las siguientes, con el fin de unificar todas bajo un mismo formato. Se ha redactado también un documento de protocolización, en el que se detallan y formalizan todos los pasos necesarios y al que podrán acudir futuros investigadores cuando deban enfrentarse al filtrado. En tercer lugar, se ha adoptado un sistema de etiquetas que complementa el sistema de signos del Grupo Val.Es.Co., con el fin de poder informatizar y facilitar las búsquedas sobre el corpus web. Por último, el proceso incluye la tokenizacion, todavía

en fase experimental, de cada una de las transcripciones, en línea con la digitalización mencionada anteriormente.

References

Briz, Antonio y grupo Val.Es.Co. (2002). *Corpus de español coloquial*. Madrid, Arco/Libros.

Briz, Antonio (2005). "Los corpus de español hablado. Presentación", Oralia. Análisis del discurso Oral. Almería, Vol.8, Madrid, Arco/Libros, 40884

Briz, Antonio (2012). "Los Déficits de los corpus orales del español (y de algunos análisis)", Cum corde el in nova grammatica. Estudios ofrecidos a Guillermo Rojo, Servizo de Publicacións da Universidade de Santiago de Compostela, 115- 137

Briz, Antonio (2016). "El proyecto AMERESCO. La idea de un corpus de conversaciones coloquiales del español de América", en Antonio Bañón y otros (eds.), Oralidad y análisis del discurso. Homenaje a Luis Cortés Rodríguez, Almería, EDUAL, pp. 81-103

Briz, Antonio y Albelda, Marta (2009). Estado actual de los corpus de lengua española hablada y escrita: I+D". En El español en el mundo. Anuario del Instituto Cervantes 2009. Consultado en http://cvc.cervantes.es/lengua/anuario/anuario_09/briz_albeida/p01.htm

Cabedo, Adrián y Pons, Salvador (eds.) (2013). Corpus Val.Es.Co 2.0. Valencia: Val.Es.Co. (Valencia, Español Coloquial), Departamento de Filología Española, Universidad de Valencia. Consultado en http://www.valesco.es

Recalde, Montserrat & Vázquez, María Victoria (2009). "Problemas metodológicos en la formación de corpus orales". En P. Cantos Gómez y A. Sánchez Pérez (eds.), A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus. Murcia: AELINCO, 37-49. Consultado en

<http://www.um.es/lacell/aelinco/contenido/titulos.html>

Torruella, Joan y Llisterri, Joaquim (1999). "Diseño de corpus textuales y orales", en

Blecua, JM., Clavería, G., Sánchez, C. y Torruella, J. (eds.) Filología e informática. Nuevas tecnologías en los estudios filológicos. Barcelona: Seminario de Filología e Informática, Departamento de Filología Española, Unviersidad Autónoma de Barcelona – Editorial Milenio. Pp. 45-77. Consultado en https://gramatica.usc.es/~gamallo/aulas/lingcomputacional/biblio/LinguisticaDe Corpus.pdf

Vázquez, María Vázquez (2014). "ESLORA: Diseño, codificación y explotación de un corpus oral de español de Galicia". II Workshop de Procesamiento Automatizado de Texto y Corpus (WOPATEC-2014). Pontificia Universidad Católica de Valparaíso, Viña del Mar.

How do mental models affect the way we evaluate things? A cross-cultural study of Peninsular-Spanish and British-English opinion articles

Francisco Miguel Ivorra Pérez

English and German Department, University of Valencia, Spain

***Keywords**: mental models; evaluative language; Peninsular Spanish; British English; opinion articles*

Abstract

Cross-cultural studies on the use of evaluative language in English and Spanish on different genres abound in the literature (Marín-Arrese & Núñez-Perucha, 2006; Alba-Juez & Martínez-Cano, 2011; Carretero & Taboada, 2014). However, this paper aims at analysing how 'mental models' (van Dijk, 2006; 2008) impact on the evaluative dimension in opinion articles from Peninsular-Spanish and British-English digital newspapers. More particularly, we draw on the framework of Appraisal Theory (Martin & White, 2005; White, 2004) and, more specifically, on the resources of 'engagement'. Following Martin and White (2005: 94), these include "all those locutions which provide the means for the authorial voice to position itself with respect to, and hence to 'engage' with, the other voices and alternative positions construed as being in play in the current communicative context". As such, the objective of the present study is twofold: (a) to establish comparisons and similarities of the patterning of engagement resources in opinion articles from both Peninsular-Spanish and British-English digital newspapers and (b) to determine the types of mental models that could affect the use of particular engagement resources in the corpora analysed.

The corpora of this research consist of 80 opinion articles drawn from the digital Peninsular-Spanish newspapers El País and El Mundo as well as from their digital British-English counterparts The Guardian and The Telegraph (20 articles were selected from each newspaper). All these belong to the year 2017 and mainly cover political and social news. In addition, we take into account the socio-political orientation of the medium of publication. Due to the different length in the total number of words and the extension of the opinion article in both corpora, the raw frequencies

found in all the categories and subcategories analysed in each subcorpus have been divided by the total amount of words included in that subcorpus and then multiplied by 1,000 words, which is considered a conventionalised way to standardize small corpora of a different size. A quantitative analysis through the Sketch Engine Software (2003) has been conducted and we have mainly focused on the word lists and frequencies analysis. An observational analysis has also been required to explore more in depth the function of some engagement resources, particularly the co-text in which these appear.

The findings reveal that the engagement resources commentarists make use of in opinion columns tend to diverge reflecting mental models, which are either sensitive to the socio-political orientation or the culture of the medium of publication. For instance, it has been found out that the more centre-left political newspapers The Guardian and El País include a higher frequency of dialogic expansive resources than their centre-right counterparts The Telegraph and El Mundo. Some similarities between the four subcorpus have also been found, which seem to be constrained by the particular conventions of this professional genre.

References

Alba, L., & Martínez, E. (2011). The evaluative function as part of the hidden pragmatic meaning of certain expressions in English and Spanish. *Rocky Mountain Review*, 65(2), 188-207.

Carretero, M., & Taboada, M. (2014). Appraisal in English and Spanish consumer reviews of books and movies: a contrastive study of the expression of graduation within the scope of attitude. In G. Thompson & L. Alba-Juez (Eds.), *Evaluation in Context* (pp. 221-236). Amsterdam: John Benjamins.

Marín, J., & Núñez, B. (2006). Evaluation and engagement in journalistic commentary & news reportage. *Revista Alicantina de Estudios Ingleses*, 19: 225-248.

Martin, J., & White, P. (2005). *The language of evaluation*. Palgrave: MacMillan.

van Dijk, T. (2006). Discourse, context and cognition. *Discourse Studies*, 8(1), 159-177. van Dijk, T. (2008). *Discourse & context. A sociocognitive approach*. Cambridge: Cambridge University Press.

White, P. (2004). Subjectivity, evaluation & point of view in media discourse. In C. Coffin, A. Hewings, & K. O'Halloran (Eds.), *Applying English Grammar: Corpus & Functional Approaches* (pp. 229-246). Arnold: London.

"The Queen Backs Brexit!" Discursive Construction of News Values in British Quality Press Coverage of Brexit Referendum

Arash Javadinejad

Department of English and German Studies, Universitat de València, Spain

***Keywords**: Brexit, News Values, Broadsheets, Discourse analysis, Corpus linguistics, CADS, DNVA*

Abstract

The Brexit referendum of 2016 was the pinnacle of the tumultuous and always problematic relation between the UK and the EU (cf. Simms, 2016; Clarke et al., 2017). The media coverage of this referendum was specifically controversial due to a number of sensitive and important discussions it brought up in British and European politics (Moore & Ramsay, 2017) and consequently, the coverage of this referendum has started to be studied using discourse analysis approaches. The existing studies on Brexit covered some specific topics, from the language of Brexit per se (Buckledee, 2018), to immigration (Cap, 2017; Share, 2018), and Racism (Virdee & McGeever, 2018), but clearly more research needs to be done in this regard. More interestingly, numerous studies have been carried out on Brexit coverage in different countries and regions such as Dekavalla (2018) in Scottish media, Roberts (2018) in Wales, Baker (2018) in Northern Ireland, Alameda Hernández (2018) in Gibraltar, Bijsmans et al. (2017) in France, Germany and the Netherlands, Martin and Binet (2018) in France, Müller (2018) in quality newspapers in Austria, Germany, and Switzerland, Fonn (2018) in Norway, León-Solís et al. (2018) in Spain, Simões-Ferreira (2018) in Portugal, Cere (2018) in Italy, Katsambekis and Souvlis (2018) in Greece, Way (2018) in Turkey, Samuel-Azran (2018) in Israel , Waddell (2018) in Canada, and Bassil-Morozow (2018) in Russia. Despite all these studies, a corpus- based discursive analysis of major British quality press (broadsheets) is still lacking.

The aim of this study is to contribute to the existing literature on discursive studies of Brexit referendum by analyzing how news values are discursively constructed in the British quality press. The study tries to answer two main questions: 1. What were the main topics covered by British broadsheets before the referendum? And 2. How was newsworthiness of these topics discursively constructed in their coverage along different ideological lines and political affiliations?

To do so, a corpus of more than 8 million words from the coverage of four British quality press (The Guardian, The Independent, The Times, The Daily Telegraph) on referendum (from February 22 to June 2016) was compiled using LexisNexis. The data were analyzed using Bednarek and Caple's (2017) recently developed framework of *Discursive News Values Analysis (DNVA)*. DNVA is a model of corpus assisted discourse analysis inspired by Baker et al.'s (2008) model of *Corpus-assisted Critical Discourse Analysis* (Bednarek and Caple, 2014, p.7).

Following this model, the analytical procedure includes using frequency analysis to extract main coverage topics in each paper, performing collocation analysis for each discourse topic, coding statically significant collocations according to the framework's categories of news values and linguistics resources of discursively constructing the newsworthiness of the topics and finally comparing the relative frequency of major news values in each newspaper. The results would reveal how newsworthiness is constructed in the press coverage of the Brexit referendum in the quality press in general, and more specifically, how this construction is different along ideological and political lines followed distinctively by each newspaper.

References

Alameda Hernández Á. (2018) The 2016 EU Referendum in Gibraltar: Opinion Articles in Gibraltarian News. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2

http://trobes.uv.es/record=e1000337*val

Baker S. (2018) Whither the 'Hand of History'?: Northern Ireland Newspaper Coverage of the 2016 EU Referendum Campaign. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978- 3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & Society, 19(3), 273-306.

Bassil-Morozow H. (2018) Russian Media and the EU Referendum 2016. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

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.

Bednarek, M., & Caple, H. (2017). The discourse of news values: How news organizations create newsworthiness. New York: Oxford University Press. Retrieved from https://ebookcentral.proquest.com/lib/univalencia/detail.action?docID=4792791 http://trobes.uv.es/record=e1000135*val

Bijsmans, P., Galpin, C., & Leruth, B. (2017). 'Brexit'in transnational perspective: an analysis of newspapers in France, Germany and the Netherlands. Comparative European Politics, 1-18.

Buckledee, S. (2018). The language of brexit: How britain talked its way out of the european union. Bloomsbury Publishing.

Cap, P. (2017). Immigration and Anti-imigration Discourses: The Early Rhetoric of Brexit. In The Language of Fear (pp. 67-79). Palgrave Macmillan, London. Cere R. (2018) 'Little England Beats Great Britain': Italian Media Coverage of the EU Referendum 2016. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Clarke, H. D., Goodwin, M. J., & Whiteley, P. (2017). Brexit: Why britain voted to leave the european union. Cambridge, United Kingdom; New York, NY: Cambridge University Press.

Dekavalla M. (2018) The EU Referendum 2016 on Scottish Television. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave

Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Fonn B.K. (2018) It's the Economy, Stupid: Coverage of the British EU Referendum in Norway. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Katsambekis G., Souvlis G. (2018) EU Referendum 2016 in the Greek Press. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

León-Solís F., Castelló E., O'Donnell H. (2018) Spanish Media and the EU Referendum 2016: The Assault on an Enlightened Project. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978- 3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Martin T., Binet L. (2018) Left Versus Right, or Mainstream Versus Margins? Divisions in French Media and Reactions to the 'Brexit' Vote. In: Ridge-Newman A., León- Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Moore, M., & Ramsay, G. (2017). UK media coverage of the 2016 EU Referendum campaign. King's College London.

Müller K.P. (2018) The 2016 EU Referendum Stories in Austrian, German, and Swiss Media: Catastrophes, Characterizations, Challenges. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Roberts S.G. (2018) 'A Pit We Have Dug Ourselves': The EU Referendum and the Welsh Democratic Deficit. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Samuel-Azran T., Galily Y. (2018) Israeli Media and the EU Referendum 2016: Political or Economic Story?. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Share O. (2018) Mobilizing Migration: Analysing the Role of the 'Migrant' in the British Press During the EU Referendum 2016 Debate. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from

http://dx.doi.org/10.1007/978- 3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Simms, B. (2016). Britain's Europe: A Thousand Years of Conflict and Cooperation. Penguin UK.

Simões-Ferreira I. (2018) Discursive Dimensions of the EU Referendum 2016 Press Coverage in Portugal. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Virdee, S., & McGeever, B. (2018). Racism, crisis, brexit. Ethnic and Racial Studies, 41(10), 1802-1819.

Waddell C. (2018) Whose News? How the Canadian Media Covered Britain's EU Referendum. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Way L.C.S. (2018) Turkish Newspapers: How They Use 'Brexit' for Domestic Political Gain. In: Ridge-Newman A., León-Solís F., O'Donnell H. (eds) Reporting the Road to Brexit. Palgrave Macmillan, Cham. Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-73682-2 http://trobes.uv.es/record=e1000337*val

Making corpus-based searching accessible for non-expert users: the case of Hansard

Lesley Jeffries[a]; Hugo Sanjurjo-González[a]

[a]*Department of English, Linguistics and History, University of Huddersfield, United Kingdom*

Keywords*: corpus linguistics; data visualisation; parliamentary debates; Hansard*

Abstract

The SAMUELS project (grant reference AH/L010062/1) released a corpus composed by UK House of Commons and Lords debates from 1803 to 2005 with a grammatical and semantic annotation (www.hansard-corpus.org) based on the Historical Thesaurus Semantic Tagger (Piao, Dallachy, Baron, Rayson & Alexander, 2014; Alexander, Dallachy, Piao, Baron & Rayson, 2015). The corpus and interface are designed for a specialist audience related to the corpus linguistics field. This means that a level of expertise is required to carry out complex queries that include semantic annotation. Parliamentary debates are also available using the official UK Parliament site (https://hansard.parliament.uk) which allows users to search specific debates and members of parliament but does not allow for searches based on linguistic parameters

that may discover patterns of debate to interest some users who lack the requisite expertise to search in the SAMUELS corpus.

One of the main goals of the Hansard at Huddersfield project is to bring corpus linguistic methods to the general public in an easy and straighforward manner. So, we are trying to overcome the weaknesses (for the general end-user) of the previously mentioned websites: required expertise in the Hansard Corpus and the lack of corpus linguistics methods on the official Hansard website. To achieve this, first we carried out a series of consultations with potential users such as politicians, journalists or historians, in order to establish the optimum set of searches and functionalities that they demand as an alternative to the current Hansard website. To make corpus linguistic methods easy to use and understand we make use of visualisations that allow us to explain results of common corpus linguistics functions such as keywords, list of frequencies or collocations in a visually appealing way.

From a technical perspective, we have developed a new accessible and web-based front end to enable the general public to access and make use of the corpus, including its semantic annotation. We also transformed the Hansard corpus into a relational data model to facilitate future developments, compatibility with other front ends and linking new data with external databases. For visualisations, we employ D3.js library, a JavaScript library for manipulating documents using common Web technologies.

References

Alexander, M., Dallachy, F., Piao, S., Baron, A., & Rayson, P. (2015). Metaphor, Popular Science and Semantic Tagging: Distant Reading with the Historical Thesaurus of English. *Digital Scholarship in the Humanities*, 30(1), 16-27. doi: /10.1093/llc/fqv045

Piao, S., Dallachy, F., Baron, A., Rayson, P., & Alexander, M. (2014). Developing the Historical Thesaurus Semantic Tagger. In *Digital Humanities Congress 2014* (pp 4-6). Sheffield, United Kingdom: Digital Humanities Institute - Sheffield. Retrieved from https://www.dhi.ac.uk/dhc/2014/paper/31

In which clause do subordinate conjunctions prosodically belong?

Zuzana Komrsková; Petra Poukarová

Institute of the Czech National Corpus, Czechia.

Keywords: *conjunction, subordination, stress group, spoken corpus*

Abstract

This paper focuses on the position of three most frequent Czech subordinating conjunctions *že* (’that’), *když* (‘when’), and *až* (‘when’) within the stress group. It compares two types of data, namely spontaneous speech and read texts. Both data contain segmentation into stress groups (phonetic words) besides traditional segmental transcription. The spontaneous speech, collected within the ORTOFON corpus (Kopřivová et al., 2017), was recorded in a private environment with fluctuating sound quality. On the other hand, the read speech was recorded in a professional studio for

detailed phonetic research.

The position of subordinating conjunctions is traditionally (e.g. Hrbáček, 1967) described as initial within the subordinate clause. According to the orthographical rule, it is necessary to write a comma before subordinating conjunctions in Czech. The rhythmical structuring of read texts does not need to be dependent on particular words rather on punctuation (by extension, the intonation patterns are more closely related to syntax, cf. Wichmann, 2000: 123). It means that the subordinating conjunctions could stay at the beginning of the stress group more often and can, but does not have to be stressed. On the contrary, Wichmann (2000: 21, 39) claims that the prosodic boundary occurs after rather than before a conjunction more often within spontaneous, unprepared speech. The online syntax theory (Auer, 2000) offers pragmatic reasons for that. The prosodic boundary occurring after a conjunction enables the speaker to signal her/his intent to continue the speech with some kind of new or known but rephrased information within the following subordinate clause. At the same time, the speaker indicates that s/he needs more time in order to think her/his next statement through. Therefore, the other speakers do not start to talk and wait for another opportunity for turn-taking.

However, Wichmann's claim also invites some doubts about the speaker's ability to predict what s/he will say. Does the meaning of the conjunction correspond to the information within main and subordinate clauses, or does it rather freely connect the clauses as a kind of discourse marker? Concerning the issue of discourse markers, the position within the stress group could be taken as a cue to determine the semantic bleaching of conjunctions (cf. the subordinate conjunction *protože*, 'because', in Čermáková et al., 2017).

References

Auer, P. (2009): On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences 31*, pp. 1–13.

Čermáková, A. et al. (2017): Between Syntax and Pragmatics: The Causal Conjunction *Protože* in Spoken and Written Czech. *Corpus Pragmatics 1(4)*, pp. 393–414.

Hrbáček, J. (1967): K poměru mezi spojovacími prostředky členskými a větnými (Podřadicí spojky v jednoduché větě) / The relation between the means connecting phrases and sentences (Subordinating conjunctions in a clause). *Naše řeč 50(3)*, pp. 138–144.

Kopřivová, M. el al. (2017): *ORTOFON: korpus neformální mluvené češtiny s víceúrovňovým přepisem, verze 1 z 2. 6. 2017*. Ústav Českého národního korpusu FF UK, Praha.

Wichmann, A. (2000): *Intonation in Text and Discourse. Beginnings, middles and ends*. Pearson Education Limited.

El tratamiento de mexicanismos en las lexicografías española e hispanoamericana

Sabrina Lafuente Gimenez

Foreign Languages Department, Universidad Federal de Sergipe, Brazil

Keywords: *mexicanismos; léxico; diccionarios*

Abstract

El objetivo de esta propuesta es difundir la investigación que se está llevando a cabo en la Universidad Federal de Sergipe -Brasil, como parte del Programa de Iniciación Científica de este centro, en el grado de Letras -español. Nuestro corpus de estudio es oral y sincrónico (Cruz Piñol, 2015), constituido por 50 unidades léxicas recolectadas *in situ* en México, en la localidad de Puebla, en el año 2016. Nos proponemos a averiguar en qué medida estas unidades de significado se encuentran plasmadas en las obras lexicográficas de referencia del español, y si sus definiciones corresponden a aquellas observadas en su contexto de uso. Para ello, nos apoyaremos en las teorías de los campos de la variación linguística, de la lingüística de corpus y de la lexicografía (Labov, 2008; Lara, 1990; Máynez, 2010; Payrató y Alturo, 2002; Torruella y Llisterri, 1999; Welker, 2014, entre otros). La metodología se hará de forma analíticacomparativa entre los lemas encontrados en las obras: Diccionario de la Lengua Española (en línea); Diccionario de Hispanoamericanismos (RICHARD, 2006); Diccionario de mexicanismos de la Academia Mexicana de la Lengua (2010) y el Diccionario del Español de México (en línea). En una segunda etapa, se realizará la búsqueda de las unidades también en el Corpus del Español del Siglo XXI - CORPES (en línea). Si bien aún nos encontramos en la fase de análisis de los datos, con este proyecto se espera: (i) que los dicentes tengan contacto con la variante mexicana, poco explorada en las clases de español como lengua extranjera; (ii) ofrecer formación científica a los discentes tanto sobre la lexicografía como sobre la utilización de corpus para este fin; (iii) contribuir con los estudios sobre las variedades del español actual.

References

Academia Mexicana de la Lengua. Diccionario de mexicanismos. México: Academia Mexicana y Siglo XXI, 2010.

Cruz Piñol, M. Lingüística de corpus y enseñanza del español como 2-L. Madrid: Arco/Libros, 2012.

el Colegio de México. Diccionario del español de México. Disponible en: http://dem.colmex.mx, Acceso en: 20/02/2017.

Labov, W. Padrões sociolinguísticos. São Paulo: Traducción de Bagno, M; Scherre, M. P.; Cardoso, C. R. Parábola, 2008.

Lara, L. F. Dimensiones de la lexicografía. A propósito del Diccionario del español de México. México: El Colegio de México, 1990.

Máynez, P. En torno al concepto y uso de" mexicanismos". Estudios de cultura náhuatl, v. 41, p. 217-230, 2010.

Medina Gerra, A. M. Lexicografía española. Barcelona: Ariel, 2003.

Payrató, L.; Alturo, N. Corpus oral de conversa col·loquial: materials de treball. Barcelona: Publicacions Universitat de Barcelona, 2002.

Real Academia Española: Corpus del Español del Siglo XXI (CORPES). Disponible en:

http://web.frl.es/CORPES/org/publico/pages/consulta/entradaCompleja.view. Acceso en: 20/02/2017.

Real Academia Española: Diccionario de la Lengua Española. Disponible en: http://dle.rae.es/?w=diccionario. Acceso en: 20/02/2017.

Richard, R. (org). Diccionario de hispanoamericanismos no recogidos por la Real Academia Española. Madrid: Cátedra, 2006.

Sardinha, T. B. Linguística de corpus. Barueri: Manole, 2004.

Torruella, J.; LListerri, J. Diseño de corpus textuales y orales. In: BLECUA, J.

M.; Clavería, G.; Sánchez, C.; Torruella, J. (Eds.). Filología e Informática. Nuevas Tecnologías en los Estudios Filológicos. Barcelona: Editorial Milenio, 1999. p. 45–77.

Welker, H. A. Dicionários: uma pequena introdução à lexicografia. Brasília: Thesaurus, 2004.

Supporting ERPP: pedagogical uses of a corpus-based lexicographic tool (*SciE-Lex*) to assist biomedical writing

Natalia Judith Laso

Department of Modern Languages and Literatures and English Studies, University of Barcelona, Spain

***Keywords**: corpus-based lexicographic tool; SciE-Lex; pedagogical applications; English for Research Publication Purposes (ERPP); biomedical writing*

Abstract

Due to the widespread dominance of English on academic writing for publication (Lillis & Curry 2010; Pérez-Llantada 2016), there is an increasing need for tools to assist English scholarly writing. This is particularly relevant for NNES writers, who face the further challenge of gaining access to and acceptance within a discourse community (Etherington 2008; Matarese 2013). Within this context, electronic dictionaries and lexical databases are worth noting lexical resources. Most of them focus on either general English or technical and scientific language. Despite being very useful lexical resources, they often lack contextual information on the lexicogrammatical patternings and usually associated with general terms frequently used in specialized registers.

English for Specific Purposes (ESP) writing research has shown that non-native English speakers (NNES) specialized discourse communities already have a good command of specialized terminology, typical of their field of research, but the use of general vocabulary tends to pose more difficulties, especially with regard to lexicogrammatical and combinatorial patterns.

With the aim of filling this lexicographical gap and helping Spanish researchers write phraseologically-competent research publications in English, the GRELIC Research Group has developed SciE-Lex (http://www.ub.edu/grelic /eng/scielex2/scielex.html), a lexical database that includes morphological, semantic, syntactic, collocation and

phraseological information and is aimed to assist the production of biomedical texts in English.

This paper serves the purpose of checking the pedagogical benefits of SciE-Lex by means of a qualitative survey distributed among a group of university researchers and professors who participated in a "Writing for Publication Workshop" and used the abovementioned lexicographic tool in the process of writing a research paper in English. The results of the survey are expected to inform the revision process of SciE-Lex as well as contribute to verifying its pedagogical benefits for the English for Research Publication Purposes (ERPP) biomedical community.

References

Etherington, S. (2008). Academic writing and the disciplines in Friedrich, P. (Ed.) Teaching Academic Writing. London: Continuum, pp. 26-58.

Lillis, Theresa & Curry, M. L. (2010). Academic writing in a global context: The politics and practices of publishing in English. London and New York:

Routledge.

Matarese, V. (Ed). (2013). Supporting Research Writing. Roles and challenges in multilingual settings. Oxford: Chandos Publishing.

Pérez-Llantada, C. (2016). Globalization and Applied Linguistics. In A. Linn (Ed.), Investigating English in Europe. Contexts and agendas (pp. 117-124). Berlín: Mouton de Gruyter. Available from: https://www.researchgate.net/publication/307632959_Globalization_and_the_c ont ribution_of_Applied_Linguistics [accessed Dec 26 2017].

Interpersonal conditionals in spoken English: Expressing stance and engagement through conditional clauses

Cristina Lastres-López

University of Santiago de Compostela, Spain

***Keywords**: conditional; interpersonal; stance; engagement; conversation*

Abstract

Conditionals are "an area of language use where the interaction of form, meaning, and context is exceptionally complex and fascinating" (Dancygier, 1998: 2). Delving into this complexity, previous studies have shown that conditional constructions can express many different pragmatic functions in discourse (Ford & Thompson, 1986; Ford, 1997; Warchal, 2010; Mato-Míguez, 2016; Puente-Castelo, 2017; Author, 2019; to name but a few), in addition to their prototypical use to indicate a cause-consequence relation

between the protasis (or subordinate clause) and the apodosis (or matrix clause), as illustrated in example (1) below.

(1) If he answers me, I'll drop you a note <ICE-GB: S1A-078 #176: 3: A>

This paper sets out to explore the uses and functions of conditional clauses in a subcorpus of conversations in English, with the hypothesis that these constructions are pragmatically richer in colloquial face-to-face conversation than in other types of spoken interaction, such as courtroom or parliamentary discourse (Author, 2019), for example. The framework proposed in this paper combines the approaches postulated in cognitivelyoriented studies on conditionality (Sweetser, 1990; Dancygier, 1998; Dancygier & Sweetser, 2000, 2005) with the three metafunctions distinguished by Halliday and Matthiessen (2014) in Systemic Functional Linguistics (SFL): (i) ideational, (ii) interpersonal, and (iii) textual. Special attention will receive interpersonal conditionals, which will be analysed in relation to the notions of stance and engagement (Hyland, 2005; cf. Evans, Bergqvist & San Roque, 2018). Interpersonal conditional constructions do not conform to the prototypical pattern of the protasis indicating the cause and the apodosis the consequence, but rather they are used to interact with the addressee and to express speakers' opinions and judgements or to exert an influence on others, for instance. Examples illustrating stance and engagement in conditional clauses are presented in (2) and (3), respectively.

(2) And also I'd be very interested in travelling and working abroad if possible <ICEGB: S1A-035 #006: 1: B>

(3) I'm just going to top up my tea again if you don't mind <ICE-GB: S1A-067 #169: 1: B>

The methodology employed in this study is corpus-based. The data are extracted from the face-to-face conversations subcorpus of the British component of the *International Corpus of English* (Nelson, Wallis & Aarts, 2002), which contains 90 conversations totalling approximately 180,000 words. The results from the sample of conversations analysed show that conditionals expressing interpersonal meaning outnumber *if*constructions which only express cause-consequence patterns in discourse. Corpus findings also reveal that, within the interpersonal category, conditionals may express the stance of the speaker or try to engage with the addressee, displaying a wide variety of discourse-pragmatic functions which range from metalinguistic comments to expressions of politeness, for instance.

References

Dancygier, B. (1998). Conditionals and prediction: Time, knowledge and causation in conditional constructions. Cambridge: Cambridge University Press.

Dancygier, B., & Sweetser, E. (2000). Constructions with 'if', 'since', and 'because': Causality, epistemic stance, and clause order. In E. Couper- Kuhlen & B. Kortmann (Eds.), Cause, condition, concession, contrast: Cognitive and discourse perspectives (pp. 111- 142). Berlin: De Gruyter Mouton.

Dancygier, B., & Sweetser, E. (2005). Mental spaces in grammar: Conditional constructions. Cambridge: Cambridge University Press.

Evans, N., Bergqvist, H., & San Roque, L. (2018). The grammar of engagement I: framework and initial exemplification. Language and Cognition, 10, 110-140.

Ford, C. (1997). "Speaking conditionally: Some contexts for if-clauses in conversation". In A. Athanasiadou, & R. Dirven (Eds.), On conditionals again (pp. 387-413). Amsterdam: John Benjamins.

Ford, C. & Thompson, S. A. (1986). Conditions in discourse: A text-based study from English. In E. C. Traugott, A. Meulen, J. S. Reilly, & C. A. Ferguson (Eds.), On conditionals (pp. 353-372). Cambridge: Cambridge University Press.

Halliday, M. A. K. & Matthiessen, C. (2014). Halliday's Introduction to Functional Grammar. London: Routledge.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. Discourse Studies, 7(2), 173-192.

Lastres-López, C. (2019). Conditionals in spoken courtroom and parliamentary discourse in English, French and Spanish: A contrastive analysis. In T. Fanego, & P. Rodríguez-Puente (Eds.), Corpus-based research on variation in English legal discourse (pp. 51-78). Amsterdam: John Benjamins.

Mato-Míguez, B. (2016). The expression of directive meaning: A corpus-based study on the variation between insubordinated if-clauses, imperatives and conditionals in Contemporary Spoken British and American English. Unpublished PhD dissertation, University of Santiago de Compostela.

Nelson, G., Wallis, S., & Aarts, B. (2002). Exploring Natural Language: The British Component of the International Corpus of English. Amsterdam: John Benjamins.

Puente-Castelo, L. (2017). On conditionality: A corpus-based study of conditional structures in Late Modern English scientific texts. Unpublished PhD dissertation, University of A Coruña.

Sweetser, E. (1990). From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure. Cambridge: Cambridge University Press.

Warchal, K. (2010). Moulding interpersonal relations through conditional clauses: Consensus- building strategies in written academic discourse. Journal of English for Academic Purposes, 9, 140- 150.

A Corpus-based Study on Polimorphemic Proper Names in Hungarian

Noémi Ligeti-Nagy

MTA-PPKE Hungarian Language Technology Research Group, Hungary

***Keywords**: Hungarian; proper names; rule-based algorithm*

Abstract

The main focus of the research presented here is on the polymorphemic proper names in Hungarian. We present a corpus-based study which is mainly motivated by the needs of a parser for Hungarian (AnaGramma (Prószéky et al. 2016)). As it processes the texts from left to right, we need to provide algorithms resolving small linguistic tasks during the parsing process that work, as well, from left to right.

The most important background of the research presented here is an algorithm called *nom-or-not* (Ligeti-Nagy et al. 2019), which is a segment of the parser and is responsible for the disambiguation of caseless nominals. The main problem still unsolved here is the recognition of structures like (1) representing a polymorphemic proper name:

(1)

Angela#N.Nom	Merkel#N.Nom	német#Adj.Nom	kancellár#N.Nom
Angela	Merkel	German	chancellor

'Angela Merkel German chancellor'

All four nominals in this example are caseless (marked by a *Nom* in their morphological annotation). This, however, does not ensure them being in nominative case which marks the subject of the sentence. To make a rule-based algorithm be able to handle cases like (1), these proper names require further inquiry.

Although we refer to the subject of this study as proper names, this simplification needs an explanation. Several studies are concerned with named entities and named entity recognition. It is stated, for example in Simon (2017), that two types of polymorphemic proper names may be distinguished: the one where the head of the structure is a common noun modified by a proper name (*Roosevelt pub*), and the one where the proper name consists of two or more proper nouns (*Angela Merkel*). However, the algorithm mentioned above needs to handle another type of polymorphemic names, which, in contrast to the two types described before, do not contain any proper name – neither as a head, nor as a modifier: *elnök elvtárs* ('comrade president') or *tanár bácsi* ('Mr. teacher'). These are titles, inscriptions, consisting of two or more common nouns.

The corpus research needed to answer the following questions: 1) what kind of common nouns can be the head of a polymorphemic proper name, 2) which type of proper names can be modifiers of these heads, 3) are there any rules how the proper names and the common nouns are allowed to connect, 4) which common nouns can combine into a title or inscription consisting of more than one common noun. We used the Szeged Treebank 2.0 (Csendes et al. 2005) as a syntactically annotated corpus to collect the possible heads of polymorphemic proper names. We extended our query on the Hungarian Gigaword Corpus (Oravecz et al. 2014). Finally, we acquired some knowledge from the word embedding models presented in Siklósi-Novák (2016).

The results indicate that there is a clear, well-definable structure in the polymorphemic proper names in Hungarian. However, the rules allowing proper names and common nouns joining into one proper name are not syntactical or morphological; they can be derived solely from world knowledge and semantic information. The head of the polymorphemic common names is always an element of a relatively small set of common nouns. This latter phenomenon may be solved by the algorithm mentioned earlier by some well-written rules. Polymorphemicproper names, on the other hand, need to be identified at a higher level of the parsing process (or by using world knowledge information).

References

Csendes, D., Csirik, J., Gyimóthy, T., and Kocsor, A. (2005). The Szeged Treebank. In: Matoušek, V., Mautner, P. and Pavelka, T., editors, Text, Speech and Dialogue:

8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings, Berlin, Heidelberg. Springer Berlin Heidelberg, 123–131.

Ligeti-Nagy, N., Dömötör, A., Vadász, N. (2019). What does the Nom say? An algorithm for case disambiguation in Hungarian. In Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages. Accepted, in press.

Oravecz, Cs., Váradi, T. and Sass, B. (2014). The Hungarian Gigaword Corpus. In: Calzolari, N. et al., editors, Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland. ELRA.

Prószéky, G., Indig, B., Vadász, N. (2016). Performanciaalapú elemző magyar szövegek számítógépes megértéséhez [A Performance-based Parser to the Comprehensive Understanding of Hungarian Texts]. In Kas, B., editor, „Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére, MTA Nyelvtudományi Intézet, Budapst, 223-232.

Simon, E. (2017). The definition of named entities. In Gyuris, B., Mády, K., and Recski, G., editors, K + K = 120. Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays. Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest.

Siklósi, B., Novák, A. (2016). Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra [Embedding models for lexical categorization tasks]. In Tanács, A. Varga, V., Vincze, V., editors, XII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: JatePress. 3-14.

A quantitative and qualitative analysis of errors occurring in the codeswitched utterances of two bilingual siblings

Cathy Lonngren-Sampaio

University of Hertfordshire, England

***Keywords**: code-switching; errors; CLAN*

Abstract

This paper reports on some of the results of a wider investigation of code-switching in a computerised corpus of child bilingual language, the LOBILL Corpus, which consists of twenty-five hours of recordings of naturalistic interactions between two bilingual Brazilian/English siblings (JAM, 3;6 and MEG, 5:10) and their family members). Collected over three years, the data was transcribed and coded using the CHAT (Codes for the Human Analysis of Transcripts) transcription system developed by MacWhinney and colleagues (MacWhinney, 2000). Errors (see Gagarina, 2013, Gillam et al, 2013 and James, 2013 for a discussion on the use of this term) were identified and coded, and information regarding the target form and possible origin of the error were also included in the transcripts.

Focusing on the production of the bilingual siblings, the CLAN (Computerized Language Analysis) software (MacWhinney, 2018) was used to perform several analyses. First, frequency analyses were performed on the siblings' output in order to determine the number of times errors occurred (i) overall and (ii) only in code-switched utterances. The

overall frequency results showed more error codes for JAM (956) than for his older sister MEG (371). However, for both children, proportionately more error codes were seen to occur in their codeswitched utterances than in their monolingual utterances. An examination of the tokens coded as errors revealed a marked difference in the type of errors each child produced while in bilingual mode: whereas the older sibling's errors were mostly lexically related, JAM's errors were mostly grammatical in nature. In addition, it was noted that while MEG's errors consisted of mostly singular occurrences (i.e. she rarely repeated the same error twice), the top two most frequent tokens in JAM's frequency list, 'which' and 'to' occurred 19 times and 15 times respectively, accounting for over 17% of the errors he produced when code-switching.

A second more qualitative analysis involved using CLAN to retrieve all the codeswitched utterances containing errors, thereby enabling their examination in a wider linguistic context. Due to the longitudinal nature of the corpus, it was also possible to track JAM's reoccurring errors over the three years and investigate the data from a developmental perspective. In many cases it was necessary to examine the siblings' monolingual output in order to determine the extent to which an error could be attributed to the act of code-switching itself (i.e. caused by the surface juxtaposition of two languages) or the result of the manifestation of underlying cross-linguistic influence (see Jarvis, 2009 and Treffers-Daller, 2009). By also considering developmental aspects and the influence of contextual factors, such as the language of the environment (which changed over the course of the data collection), this paper provides a holistic examination of two bilingual siblings' errors, the results of which offer a unique insight into the possible outcomes of bilingual language use.

References

Gagarina, N.V. (2013). Acquisition and loss of L1 in a Russian-German bilingual child: A case study. Путь в язык. Одноязычие и двуязычие, 137.

Gillam, R. B., Peña, E. D., Bedore, L. M., Bohman, T. M., & Mendez-Perez, A. (2013). Identification of Specific Language Impairment in Bilingual Children: I. Assessment in English. Journal of Speech, Language, and Hearing Research, 56(6), 1813-1823.

James, C. (2013). Errors in language learning and use: Exploring error analysis. Routledge.

Jarvis, S. (2009). Lexical Transfer. In A. Pavlenko (Ed.), The Bilingual Mental Lexicon: interdisciplinary Approaches (pp.99-124). Bristol: Multilingual Matters.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2018). The CHILDES Project, Tools for Analyzing Talk – Electronic Edition. Part 2: The CLAN Programs. Carnegie Mellon University. Available online: https://talkbank.org/manuals/CLAN.pdf

Treffers-Daller, J. (2009). Code-switching and transfer: an exploration of similarities and differences. In B. E. Bullock & J. Toribio (Eds.), The Cambridge Handbook of Linguistic Code-switching (pp.58-74). Cambridge: Cambridge University Press.

Diseño de un corpus de estímulos para el estudio psicolingüístico de la ambigüedad léxica

Natalia López Cortés

Departamento de Lingüística General e Hispánica, Universidad de Zaragoza, España

Keywords: *homonimia; polisemia; monosemia; ambigüedad; sincronía; psicolingüística*

Abstract

La ambigüedad léxica, producida cuando un único lexema lleva aparejado más de un significado, es un fenómeno que ha sido ampliamente estudiado. La bibliografía establece la existencia de dos tipos de ambigüedad léxica: la polisemia y la homonimia. Desde un punto de vista experimental, cobran especial importancia los estudios, llevados a cabo en los últimos años, que encuentran un procesamiento diferencial para homonimia y polisemia (Rodd et al., 2002; Klepousniotou & Baum, 2007, entre otros). Lo interesante de este dato es que apunta a un almacenamiento igualmente diferencial de estas unidades en el lexicón mental.

Normalmente, la homonimia y la polisemia se definen desde la diacronía, es decir, teniendo en cuenta la evolución histórica de las unidades léxicas (Ulmmann, 1986). Sin embargo, diversos autores afirman que los hablantes no son sensibles a esta distinción histórica (Gutiérrez Ordoñez, 1989). Esta investigación parte de la hipótesis de que lo verdaderamente relevante a la hora de estudiar la ambigüedad léxica es la interpretación subjetiva de las unidades léxicas que hacen los hablantes. Así, para poder llevar a cabo un estudio sobre el procesamiento y el almacenamiento de las unidades homónimas y polisémicas, se han de redefinir los conceptos de homonimia y polisemia teniendo en cuenta cómo los hablantes interpretan los diferentes significados.

Para ello, en los últimos años se ha diseñado un corpus de palabras ambiguas, que pretende servir de base para nuestra propia investigación psicolingüística sobre polisemia y homonimia, así como material para cualquier investigador interesado en el tema.El punto de partida de este corpus es, como se ha mencionado antes, la interpretación subjetiva de los hablantes. Es un corpus basado, por tanto, en la sincronía. Para obtener estos datos sincrónicos se ha seguido una metodología basada en cuestionarios, que nos permiten recabar la información necesaria y conocer qué piensan los hablantes sobre palabras con más de un significado en español.

Además, puesto que este corpus pretende ser una base de datos de estímulos, se recogen también otros valores de interés para el diseño de experimentos y pruebas de laboratorio. Resulta de especial interés el proceso de selección de palabras monosémicas adecuadas, balanceadas en significado y frecuencia con las unidades ambiguas y que sirvan como estímulos y controles en el diseño de tareas de decisión léxica.

El objetivo de esta comunicación es doble. Por un lado, se presentará el trabajo de diseño del corpus, explicando en qué consisten los cuestionarios, cómo se interpretan los resultados y cómo se clasifican los estímulos en monosémicos, homónimos y polisémicos. Por otro lado, de manera descriptiva, se ofrecerán los datos actuales del corpus que, aun siendo provisionales, sirven para ilustrar el trabajo de diseño que se está

llevando a cabo. Además, se compartirán resultados de investigaciones sobre el procesamiento de la homonimia y la polisemia, realizadas con los materiales del corpus.

Los resultados significativos de estos estudios preliminares prueban que el trabajo de diseño de corpus de estímulos es un paso preliminar necesario para cualquier investigación experimental.

References

Gutiérrez Ordoñez, S. (1989). Introducción a la semántica funcional. Madrid: Síntesis.

Klepousniotou, E., & Baum, S.R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. Journal of Neurolinguistics, 20, 1-24.

Rodd, J., Gaskell, M. G. & Marslen-Wilson, W.D. (2002). Making sense of semantic ambiguity: semantic competition in lexical access. Journal of Memory and Language, 46, 245-266.

Ullmann, S. (1986). Semántica: introducción a la ciencia del significado. Madrid: Aguilar.

El papel de la lingüística en la detección automática de errores en el ámbito biosanitario: hacia una propuesta de tipología de errores

Jésica López-Hernández[a]; Ángela Almela[b]

[a]*Departamento de Informática y Sistemas, Universidad de Murcia, España;*

[b]*Departamento de Filología Inglesa, Universidad de Murcia, España*

Keywords*: tipología de errores; detección automática de errores; corrección ortográfica automática; lenguaje biomédico*

Abstract

El objetivo de este trabajo es realizar una tipificación de errores en español para el ámbito biosanitario. Se han llevado a cabo estudios sobre identificación y clasificación de errores en otros idiomas (Kukich, 1992; Paggio, 2000; Gimenes, Roman y Carvalho, 2015; Siklósi, Novák, y Prószéky, 2016, entre otros), así como la creación de tipologías de errores en español (Díaz, 2005; Ramírez y López, 2006), pero no contamos con datos cuantitativos para el lenguaje de especialidad médico en español. Además, no es posible hablar de una tipología universal de errores, sino que los tipos de error identificados pueden variar dependiendo del ámbito y contexto de uso (Díaz, 2005), de ahí la necesidad de elaborar una tipología propia para el lenguaje médico, pues en estos textos el uso de términos técnicos específicos del dominio es constante.

El análisis de patrones de error está ligado al diseño de sistemas de detección y corrección automática; asimismo, también es útil para reconocimiento óptico de caracteres, extracción de información y otras tecnologías del lenguaje natural (Jurafsky y Martin, 2014). Concretamente, mediante la detección y la corrección ortográfica automática

puede agilizarse una parte del proceso de normalización de documentos, lo que posibilita el acceso, recuperación e intercambio de información.

El corpus objeto de análisis está compuesto por miles de informes clínicos de diferentes especialidades, compilado con ayuda de profesionales de cada área y que ha sido utilizado para el desarrollo de un sistema de reconocimiento de voz en español para entornos médicos. La tasa de acierto de estos sistemas de reconocimiento de voz está estrechamente relacionada con la calidad de los modelos lingüísticos que se emplean para su funcionamiento; estos, a su vez, dependen del grado de corrección y normalización de los corpus utilizados para construirlos, por ello es de gran importancia detectar y corregir errores como mecanismo para incrementar la tasa de acierto.

Con el fin de contribuir a la automatización del proceso de corrección, en este proyecto se ha llevado a cabo un análisis estadístico de errores en el corpus objeto de estudio. Para la delimitación y clasificación de errores se han tomado como punto de partida las propuestas de análisis de Díaz (2005) y Ramírez y López (2006). Específicamente, se han tenido en cuenta parámetros como frecuencia de aparición y tipo de error (omisión, sustitución, adición o trasposición), posición del error en la palabra, longitud de la palabra o contexto en el que se produce. Los resultados preliminares son prometedores, por lo que con esta clasificación de errores específica del ámbito biosanitario se pretende contribuir a la descripción y organización de errores y a la mayor cobertura de casos para detección y corrección ortográfica automática.

References

Díaz Villa, A. (2005). Tipología de errores gramaticales para un corrector automático, Procesamiento del Lenguaje Natural, 35, 409-416.

Gimenes, P. A., Roman, N. T., y Carvalho, A. M. (2015). Spelling Error Patterns in Brazilian Portuguese. Computational Linguistics, 41(1), 175–183. DOI:10.1162/coli_a_00216.

Jurafsky, D. y Martin, J. (2014). Speech and Language Processing, London: Pearson Education.

Kukich, K. (1992). Technique for automatically correcting words in text. ACM Computing Surveys, 24(4), 377-439. DOI:10.1145/146370.146380.

Paggio, P. (2000). Spelling and grammar correction for Danish in SCARRIE. Proceedings of the Sixth Conference on Applied Natural Language Processing. DOI:10.3115/974147.974182.

Ramírez, F. y López, E. (2006). Spelling Error Patterns in Spanish for Word Processing Applications, Proceedings of Fifth international conference on Language Resources and Evaluation, LREC, 93-98.

Siklósi, B., Novák, A., y Prószéky, G. (2016). Context-aware correction of spelling errors in Hungarian medical documents, Computer Speech & Language, 35, 219-233.

Using corpus tools to unveil ideological dynamics in the anti-vax campaign rhetoric.

Stefania M. Maci

Foreign Languages and Literatures Department, University of Bergamo, Italy

Keywords: *Twitter; #antivax; ideology; Corpus tools; fake news*

Abstract

Antivaccine controversial debates have been occurring for almost a century. As the debate has moved onto social media, the issue has been amplified by the interplay between the visual and the verbal components: for anti-vaccine campaigners the use of Twitter means giving them voice and massively amplifying their message. Yet, precisely because social media advertisements and news work on the basis of an algorithm that brings people to see similar news to those they have read before, anti-vax campaigners tend to read and to always believe in the same type of news, be it fake (Fernández-Luque and Bau 2018; Tandoc et al., 2017) or real. In other words, they cannot discern real and fake news (Balmas, 2014), as they do not realize that scientific fake news is the result of a decontextualization of the medical sources.

This paper, by drawing on on sociosemiotics (Kress, 2010) and CDA (Fairclough, 1989, 2014a, 2014b) aims at analyzing fake news ideological dynamics and discourse strategies related to the anti-vax campaing to unveil cognitive, social and institutional constructs of misinformation and to create a possible linguistic frameworks for fake news detection across scientific genres and cultures.

9705 tweets (roughly 40,000 running words) will be analized on a corpus-based approach (Tognini-Bonelli, 2001). The quantitative analysis will be carried out with Atlas.TI, WordSmith Tools (Scott, 2012) and WMatrix (Rayson, 2006) to detect the main aspects related to socio-semiotic, lexico-pragma and semantic features. Qualitative analysis will follow to reveal the ideological framework behind the #anti-vax narrative and its rhetoric.

Data suggest that the #antivax narrative is negatively-oriented and negatively-constructed both through the use of images and the way in which discourse is formulated. As to images, the use of Atlas.TI to images has revealed that are employed in the following cases:

- To create emotive response
- To enumerate disease, conditions, negative side effects
- To emphasize the mandatory aspect of vaccination as a negative issue
- To stress on the necessity of getting informed
- To underline statistic data
- To Focus on the economic profits for the pharmaceutical industry (and related labs/scientific teams, medics included)

WordSmith Tools has revealed that Tweets are characterized there is a great frequency of Noun Verb and Adjectives pivoting around the discourse of vaccination and their negative effects or adverse reactions on children, the necessity of being *informed* parents and the need to use social media to stop Big Pharma and eventually sue these corporation and, if this is the case, also the Government.

In addition, the semantic profile elaborated through WMatrix evidences the existence of a bias within the #anti-vax narrative: objectivity in science is not taken into consideration

as it is implied the presence of conspiracies and financial interests which lead to chaos and disinformation.

To conclude, the discourse of #antivax discourse on Twitter, constructed in an accessible way, seems to be founded on moral and scientific grounds, supported by visual metaphors amplifying the verbal component of the tweet.

References

Balmas, Meital (2014). When Fake News Becomes Real. Communication Research 41/3, 430-454

Fairclough, Norman (1989). Language and Power. London: Longman.

Fairclough, Norman (2014a). Language and Power (3rd edition). London: Longman.

Fairclough, Norman (2014b). Critical Language Awareness. London: Routledge

Fernández-Luque, Luis / Bau, Teresa (2015). Health and Social Media: Perfect Storm of Information. Healthcare Informatics Research 21/2:67-73

Kress, Gunther (2010). Multimodality. A Social Semiotic Approach to Contemporary Communication. London: Routledge

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

Scott, Mike (2012), WordSmith Tools version 6, Stroud: Lexical Analysis Software.

Tandoc, Edson C. Jr./ Lim, Zheng Wei/ Ling, Richard (2017). Defining "Fake News". Digital Journalism

Tognini-Bonelli, Elena. (2001). Corpus Linguistics at Work. Amsterdam: John Benjamins.

Syntactic complexity in argumentative essays of advanced learners and native speakers of German

Christos Makrodimitris

Institute for Psycholinguistics and Didactics of German, Goethe University Frankfurt, Germany

***Keywords**: syntactic complexity; German as a second language; proficiency level; argumentative writing*

Abstract

This study examines syntactic complexity in argumentative essays of adult second language (L2) learners of German (*n*=45, evenly distributed among levels B2, C1, and C2 of the Common European Framework) as well as of native (L1) German speakers (*n*=15). One writing sample from each L1 and L2 author is included in the analysis. All texts were produced under identical conditions and are part of the Falko-essay-corpus (Reznicek et al., 2012). The main hypothesis of the study is that syntactic complexity is an aspect of

L2 performance that concretizes L2 proficiency and therefore more proficient learners are expected to produce more complex syntactic structures than learners of lower proficiency level (cf. Bulté & Housen, 2012; Michel, 2017). The objective of the study is to examine at which level L2 learners can produce equally complex texts as native speakers. In order to account for different dimensions of syntactic complexity, the following measures were applied to each individual text (cf. Norris & Ortega, 2009): i) Mean Length of T-unit (MLU) for overall complexity, ii) Mean Length of Clause (MLC) for complexity via phrasal elaboration, and iii) Subordination Index (SI), defined as the mean number of subordinate clauses per T-unit, for complexity via subordination. The latter measure was further differentiated in first subordination grade (SubI), i.e., the mean number of subordinate clauses depending directly on a main clause per T-unit, and second subordination grade (SubII), i.e., the mean number of subordinate clauses depending on a further subordinate clause per T-unit. Clauses were defined as grammatical units containing either a finite or an infinitive predicate, whereas T-units were operationalized, following Hunt (1965: 20), as text segments consisting of one main clause and all the subordinate clauses attached to it. Group means (μ) in the different measures were compared using the independent samples *t*-test for normally distributed data and the Wilxocon rank-sum test for not normally distributed data. In respect of the MLU, the results show that learners at B2 level (μ_{B2}=11.89) differ significantly from both the more advanced learners (μ_{C1}=14.41, μ_{C2}=14.28, p<0.05 in both cases) and the native speakers (μ_{L1}=15.09, p<0.001). As regards the MLC, the B2 group achieves lower scores (μ_{B2}=6.83) than the more advanced L2 groups (μ_{C1}=7.54, μ_{C2}=7.42) and the native speakers (μ_{L1}=7.76), but it differs significantly only from the latter (p<0.05). On the other hand, the MLU and MLC scores of the C1 and C2 level learners do not differ from one another nor from those of the L1 writers. As far as the three subordination measures are concerned, differences are not significant across the board, although the B2 learner group again achieves the lowest scores overall (e.g. for SI: μ_{B2}=0.71, μ_{C1}=0.86, μ_{C2}=0.88, μ_{L1}=0.86). In sum, B2 learners differ from native speakers with regard to both MLU and MLC and from more advanced learners only with regard to MLU. At the same time, C1 and C2 learners exhibit homogenous and native-like performance. These findings suggest that the transition from B2 to C1 level is in terms of syntactic complexity an important juncture in instructed L2 acquisition of German.

References

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA (pp. 21-46). Amsterdam, Netherlands: John Benjamins.

Hunt, K. W. (1965). Grammatical structures written at three grade levels (NCTE Research Report No. 3). Champaign, IL: National Council of Teachers of English.

Michel, M. (2017). Complexity, accuracy, and fluency in L2 production. In S. Loewen & M. Sato (Eds.), The Routledge handbook of instructed second language acquisition (pp. 50-68). New York, NY: Routledge.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. Applied Linguistics, 30(4), 555-578. doi: 10.1093/applin/amp044

Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., Hirschmann, H., & Andreas, T. (2012). Das Falko-Handbuch: Korpusaufbau und Annotationen (Version 2.01). Retrieved from https://www.linguistik.huhttps://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falkoberlin.de/de/institut/professuren/korpuslinguistik/forschung/falko

Objectiveness and detachment in *legalese*: is it justified to take them for granted? A corpus-driven profiling of the vocabulary of *appraisal* in legal English.

María José Marín Pérez [a]; Camino Rea Rizzo[b]

[a]*Departamento de Inglés, Universidad de Murcia, Spain;* [b]*Departamento de Métodos Cuantitativos, Ciencias Jurídicas y Lenguas Modernas, Universidad Politécnica de Cartagena, Spain*

Keywords: *corpus linguistics; discourse studies; legal language; appraisal theory*

Abstract

The objective and impersonal nature of legal language has traditionally been emphasised by scholars (Alcaraz, 1994; Bathia, 1982; Cao, 2007; Mellinkoff, 1963; Tiersma, 1999), who consider that this English for Specific Purposes (ESP) variety displays authorial detachment on the lexical, syntactic and even discursive levels. Such detachment is often achieved at the expense of clarity and effectiveness as regards the communicative function of legal text, whose major aim, that of resulting precise and unambiguous, is often blurred by the convoluted and obscure character of its language.

The expression of dissatisfaction on the part of laymen and even specialists in this respect can be traced back to the 16th century, when Edward VI, king of England, wished that "the superfluous and tedious statutes were (…) made more plain and short, to the intent that men might better understand them" (as cited in Bathia, 1982: 4). Several initiatives have been carried out to date to make legal language less obscure and more accessible such as the Plain English Campaign[1], which aims at making public documents more accessible to the general public.

Impersonality has thus been envisaged as a common feature to all legal genres which manifests itself through different linguistic mechanisms. One of the most widespread options to accomplish this function is the use of agentless passive forms, participle clauses, or cleft sentences, amongst other (Orts, 2018). A feasible consequence of constructing legal discourse around the axes of impersonality and detachment, which necessarily increases the difficulty for legal texts to be deciphered by the laymen, is the tendency for the "ruling class (the law-makers) (…) to maintain its advantageous social and cultural hegemony upon its audience (the law-takers)" (as cited in Orts, 2018: 808).

Nevertheless, specifically within the field of legal argumentation in judicial decisions, subjectivity also plays a relevant role in building such arguments. Judges' discourse must

[1] http://www.plainenglish.co.uk/

not only appear fair and objective to all the parties but also be justified on the basis of evaluation, that is, the strengths and weaknesses of the arguments in a judicial decision necessary "express their stances and align themselves with other institutional interactants" (Goźdź-Roszkowski, 2017: 93).

Therefore, the present research introduces a corpus-driven analysis of a 3.7 million-word collection of 600 judicial decisions issued by British courts which aims at demonstrating that the presence of evaluative vocabulary in judicial decisions on immigration is more than incidental. Using the *Appraisal Theory* framework (White, 1999; Martin 2003; Eggins and Slade, 1997; Rothery & Stenglin, 2000; and Kaltenbacher, 2006) as reference for the analysis, it was found that, out of the most frequent 2,500 types extracted from the corpus, 10% of them could be deemed evaluative.

In order to reduce the subjectivity levels implied in identifying this type of vocabulary, an inter-rater reliability test was designed for two specialists to decide whether the items on the list could be regarded as evaluative. Then, the items were classified following the taxonomy suggested by the above-cited authors and it was found that the system *appreciation* contained the largest amount of these items, 46%, followed by *judgment*, 26%, *affect*, 19%, and *amplification*, 8%.

References

Alcaraz Varó, E. (1994). El inglés jurídico: textos y documentos. Madrid: Ariel Derecho.

Bhatia, V.K. (1982). An Investigation into Formal and Functional Characteristics of Qualifications in Legislative Writing and its Application to English for Academic Legal Purposes. PhD thesis. University of Aston in Birmingham.

Cao, Deborah. 2007. Translating Law. Clevedon/Buffalo/Toronto: Multilingual Matters.

Goźdź-Roszkowski, S. (2017). 'Signalling sites of contention in judicial discourse. An exploratory corpus-based analysis of selected stance nouns in us supreme court opinions and Poland's constitutional tribunal judgments'. Comparative Legilinguistics, 32: 91-115.

Kaltenbacher, M. (2006). 'Culture-Related Linguistic Differences in Tourist Websites: the Emotive and the Factual. A Corpus Analysis within the Framework of Appraisal'. In G.Thompson and S. Hunston (eds.). System and Corpus. Exploring Connections. London: Equinox.

Martin, J.R. (2003). 'Beyond Exchange: APPRAISAL Systems in English'. In Hunston, S. and Thompson, G. (eds.). Evaluation in text. Authorial Stance and the Construction of Discourse. Oxford: Oxford University Press.

Mellinkoff, D. (1963). The Language of the Law. Boston: Little, Brown & Co.

Orts Llopis, M.A. (2018). 'A bilingual, bicultural approach to detachment and appraisal in the law: tracing impersonality and interaction in English and Spanish legal opeds'. International Journal for the Semiotics of Law, 31 (4): 805-828.

Rothery, J. and Stenglin, M. (2000). 'Interpreting Literature: the role of APPRAISAL'. In Unsworth, L. (ed.). Researching Language in Schools and Communities: Functional Linguistics Perspectives. London: Casell.

Tiersma, P. (1999). Legal Language. Chicago: The University of Chicago Press.

White, P.R.R. (1999). The Language of Attitude, Arguability and Interpersonal Positioning. The appraisal website: http://www.grammatics.com/appraisal/

The lemmatisation of a parallel corpus of Old English. Automatic searches and knowledge bases

Javier Martín Arista

Departamento de Filologías Modernas, Universidad de La Rioja

Keywords: *parallel corpus; Old English; lemmatisation; knowledge-base*

Abstract

The most widely used corpora of Old English do not provide parallel texts with word-for-word glosses, neither are they annotated by inflectional form and lemma. The YorkToronto-Helsinki Parsed Corpus of Old English Prose and *The York-Helsinki Parsed Corpus of Old English Poetry*, nevertheless, offer morphological tagging (without lemmatisation) and are parsed syntactically. Therefore, the state of play in Old English corpus design and compilation can be summarised by pointing out that more annotation is required, which, in turn, crucially depends on lemmatisation. The need is even more pressing considering that the written records present a remarkable degree of spelling variation which can be attributed to dialectal and diachronic variation and, above all, to the lack of a fixed spelling. Against this background, the aim of this paper is to present the two procedures that are being implemented to lemmatise a parallel corpus of Old English and to assess the results obtained so far. Inflectional forms in the selected texts are assigned to the lemmas listed in the lexical database of Old English Nerthus by means of automatic searches launched on the lemmatiser Norna and by retrieving information from lexicographical and textual sources. Norna is one of the building blocks of a relational database implemented in Filemaker software, which comprises the lemmatiser as well as two knowledge bases (*Nerthus* and *Freya*), a concordance and an index to *The Dictionary of Old English Corpus* and The Whirl, a layout that compares the results of the analysis of primary sources with the information retrieved from secondary sources (fig. 1). Beginning with automatic searches, query strings are defined on the basis of the target category and aimed to the derivational affix, the inflectional affix, and/or the stem. Turning to the retrieval of information, two types of sources are relevant for the research: lexicological and lexicographical sources (these include grammars, glossaries and the available dictionaries, Bosworth-Toller, Sweet, HallMerritt, and *The Dictionary of Old English*), and textual sources (these are corpora that provide annotation at some level, which is tantamount to saying the York corpora). The results obtained so far with these procedures are assessed as to quantity and accuracy. Two conclusions are drawn. Firstly, it is necessary to refine the lemmatisation procedure in order to maximise the automatisation of the annotation of the parallel corpus. This can be partly achieved with the feedback from automatic searches. Secondly, in order to enhance the retrieval of information from the knowledge bases, it is necessary to align the headword spellings of *Nerthus* and *Freya*.

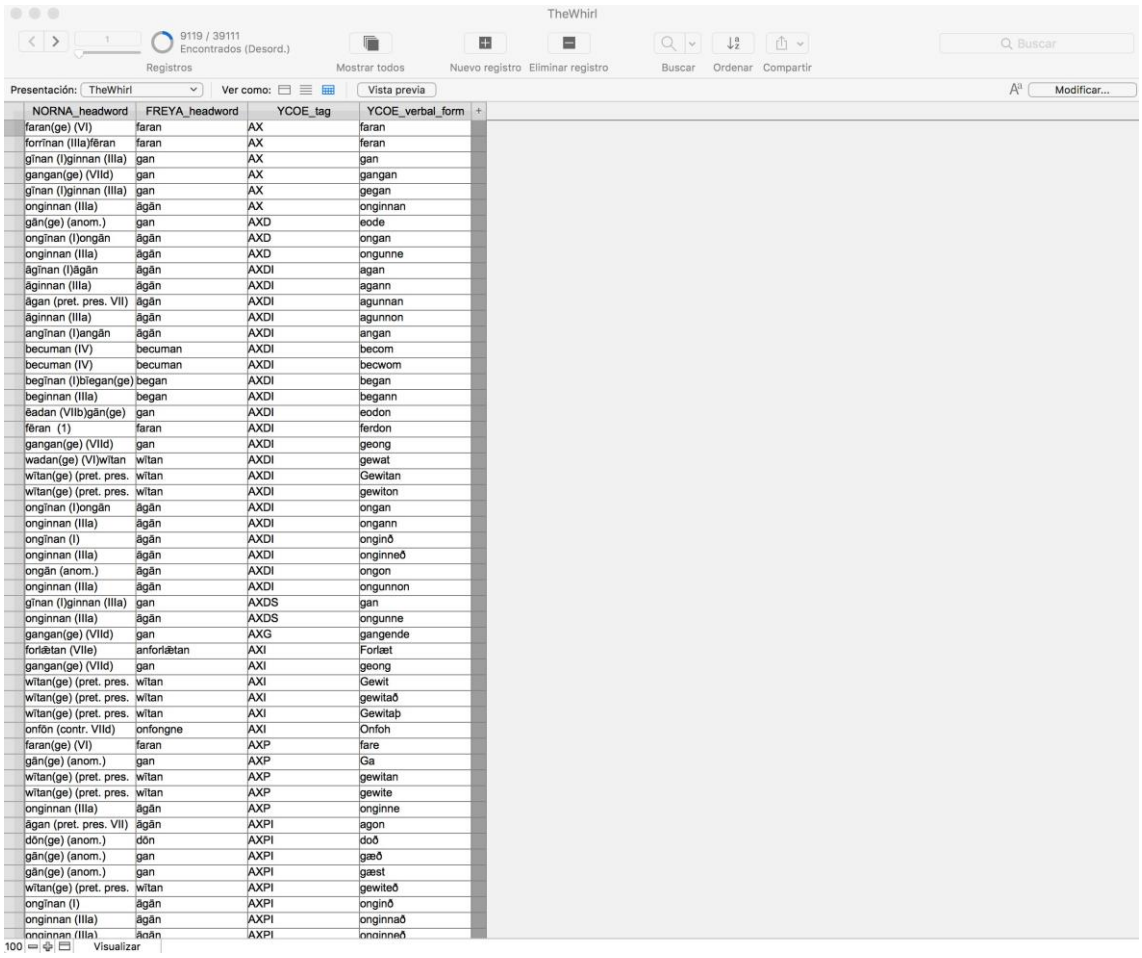


Fig. 1. The Whirl.

References

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

Clark Hall, J. R. (1996). A Concise Anglo-Saxon Dictionary. Toronto: University of Toronto Press.

Healey, A. diPaolo (Ed.) with J. Price Wilkin & X. Xiang (2004). The Dictionary of Old English Web Corpus. Toronto: Dictionary of Old English Project.

Healey, A. diPaolo (Ed.) (2016). The Dictionary of Old English in Electronic Form A-H. Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

Pintzuk, S. & L. Plug (ed.) 2001. The York-Helsinki Parsed Corpus of Old English Poetry. Department of Language and Linguistic Science, University of York.

Sweet, H. 1976 (1896). The student's Dictionary of Anglo-Saxon. Cambridge: Cambridge University Press.

Taylor, A., A. Warner, S. Pintzuk & F. Beths (2003). The York-Toronto-Helsinki Parsed Corpus of Old English Prose. Department of Language and Linguistic Science, University of York.

A corpus-based approximation to non-financial corporate discourse in English

José Manuel Martínez Martínez [a]; Lianet Sepúlveda Torres [a]; Donato Calace [b]

[a]*Datamaran, Valencia, Spain;* [b]*Datamaran, London, UK*

Keywords: *corporate discourse; corporate reports; financial reports; non-financial reports; English, terminology; greenwashing*

Abstract

Non-financial reporting (NFR) is today a global and established phenomenon on its way towards institutionalization through dedicated legislation. NFR consists in disclosing information covering at least the following topics: (a) environmental matters, (b) the company's employees, (c) social matters, (d) respect for human rights, and (e) anti-corruption and anti-bribery matters. The latest KPMG Survey of Corporate Responsibility Reporting (KPMG, 2017) indicates that 93% of the 250 largest companies by revenue based on the Fortune Global 500 2016 index adopted NFR.

Policies, regulations, stock exchanges, and investors are increasingly requiring disclosure of non-financial information in annual reports or stand-alone documents. Although, there is a body of legislation regulating disclosure of non-financial information like the Sarbanes-Oxley Act in the US, or the Directive 2014/95/EU in the EU, the policy making process is still unfolding and different actors are calling policy-makers and regulators for further action (Williams & Fisch, 2018).

There is a wide and shared consensus that NFR offers a contribution towards the enactment of sustainability development principles at business level. Indeed, compiling a report through collecting a firm's ESG performance should raise the organizational awareness on such matters, ensuring that they are taken into account in decision-making processes. Nevertheless, factual evidence refutes the NFR-sustainability performance equation, suggesting that there is more than meets the eye. Most of the leading companies in NFR practices are indeed organizations operating in industries with major impacts on social and environmental systems, such as oil & gas, metals & mining, chemicals or are often involved in major scandals.

This raises the question of how reliable the information is disclosed through NFR and to what extent NFR offers a fair account of a company's material non-financial risks and opportunities, and their centrality in the business strategy. In other terms, understanding when NFR is "greenwashing" corporate image. A crucial knowledge gap exists in identifying how and when companies greenwash. We propose to address this gap by looking at the discursive patterns showed by corporations in their disclosures concerning non-financial issues along time, across report types, both internally—the consistency of the discourse of a single organization—and externally—the similarity of the discourse of a single organization compared with the common practices of its peers.

Previous corpus-based studies on corporate discourse have mainly focused on financial reports like Wang, Li and Cao (2012), the work done in the frame of the CFIE Project (see Pimenta Alves, El-Haj, Rayson, Walker & Young, 2016; Vasiliki, El-Haj, Rayson, Walker, & Young, 2014; El-Haj, Rayson, Young, Walker, Moore, Athanasakou, &

Schleicher, 2016), and the Jena Organization Corpus (Händschke, Buechel, Goldenstein, Poschmann, Duan, Walgenbach, & Hahn, 2018) which also includes social responsibility reports. We go beyond previous proposals by annotating non-financial disclosures on a corpus of corporate reports in English containing financial and non-financial reports.

In this paper, we will describe the corpus, its annotation, and we will discuss the variation of corporate non-financial discourse using this annotation (a) across different types of reports, (b) along time, (c) within industries. We conclude with a reflection on the meaning of the patterns identified and those features that could be potential indicators of "discursive greenwashing" practices.

Our research contribution is multidisciplinary: (1) to the non-financial accounting field proposing a methodology to identify disclosure patterns which can lay the foundations to characterize "discursive greenwashing" practices and showing how NLP techniques can be applied to the analysis of corporate disclosures, and (2) to the corpus linguistics and discourse analysis field by introducing the first study that compares and contrasts corporate discourse around NFR topics in financial versus non-financial reports.

References

El-Haj, M., Rayson, P. E., Young, S. E., Walker, M., Moore, A., Athanasakou, V., & Schleicher, T. (2016). Learning tone and attribution for financial text mining. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation. (pp. 1820-1825). European Language Resources Association (ELRA). (PDF; annotated corpus and guidelines)

Händschke, S. G. M., Buechel, S., Goldenstein, J., Poschmann, P., Duan, T., Walgenbach, P., & Hahn, U. (2018). A Corpus of Corporate Annual and Social Responsibility Reports: 280 Million Tokens of Balanced Organizational Writing. In Proceedings of the First Workshop on Economics and Natural Language Processing, pages 20–31 Melbourne, Australia, July 20, 2018 Retrieved from http://aclweb.org/anthology/W18-310 3

KPMG. (2017). The KPMG Survey of Corporate Responsibility Reporting 2017. Retrieved from www.kpmg.com/crreporting

Pimenta Alves, P. A., El-Haj, M., Rayson, P., Walker, M., & Young, S. (2016, July 1). Heterogeneous Narrative Content in Annual Reports Published as PDF Files: Extraction, Classification and Incremental Predictive Ability. Retrieved from https://ssrn.com/abstract=2803275

Vasiliki A., El-Haj, M., Rayson, P., Walker, M., & Young, S. (2014) Computer-based Analysis of the Strategic Content of UK Annual Report Narratives. American Accounting Association Annual Meeting, August 2-6, 2014, Atlanta, USA.

Wang, H., Li, L., & Cao, J. (2012). Lexical features in corporate annual reports: a corpus-based study. EUROPEAN JOURNAL OF BUSINESS AND SOCIAL SCIENCES, 1(9), 55–71. Retrieved from http://www.ejbss.com/Data/Sites/1/decemberissue2012/ejbss-12-1183-lexicalfeaturesi ncorporateannualreports.pdf

Williams, C. A., & Fisch, J. E. (2018). Petition to SEC for Rulemaking on

Environmental, Social, and Governance (ESG) Disclosure. Harvard Law School Forum on Corporate Governance and Financial Regulation. Retrieved from https://corpgov.law.harvard.edu/2018/10/09/petition-to-sec-for-rulemaking-on-enviro nmental-social-and-governance-esg-disclosure/

Análisis de la complejidad estilística de las opiniones de los Abogados Generales de las Cortes de Justicia de la Unión Europea: una demostración de la funcionalidad de la lingüística de corpus para el alcance de resultados cualitativos.

Virginia Mattioli

Departamento de Traducción e Interpretación, Pontificia Universidad Católica de Valparaíso, Chile

***Keywords**: estudios cualitativos basados en corpus; metodología de corpus; complejidad estilística; Cortes de Justicia de la Unión Europea; Universales de Traducción*

Abstract

Partiendo de postulados teóricos sobre los rasgos estilísticos de los textos, se diseñó una metodología basada en corpus que permitiese medir la complejidad estilística de un conjunto de textos. A continuación, se comprobó su funcionalidad mediante su aplicación práctica. En concreto, se realizó un estudio comparativo entre dos corpus que representan las opiniones de los Abogados Generales de la Cortes de Justicias de la Unión Europea (CJEU). Específicamente, se compararon las opiniones escritas antes de 2004 con aquellas producidas después de esa fecha. Efectivamente, en 2004 se introdujo una reforma en el sistema lingüístico de las CJEU que, de acuerdo con la hipótesis de la investigación, conllevaría la simplificación estilística de las opiniones de los Abogados Generales y la consiguiente disminución de su elocuencia y función persuasiva. La reforma consistió en reducir el número de los idiomas de redacción de las opiniones obligando a los Abogados Generales a escribir los textos en lenguas diferentes de sus idiomas nativos. La metodología propuesta para el análisis se compone de 3 fases. En primer lugar, se estudió la literatura disponible para determinar los rasgos lingüísticos responsables de la complejidad estilística y la fluencia de un texto. Frente a la carencia, en el marco de los estudios lingüísticos, de aportaciones concretas que permitieran identificar rasgos unívocos relacionados con el fenómeno en examen, se reparó en los Estudios de Traducción. En efecto, una de las hipótesis de los Universales de Traducción (Baker, 1993) es la mayor simplicidad de los textos traducidos frente a los originales que, por lo tanto, presentarían una mayor complejidad estilística. Sustentándose en estas aportaciones, se determinaron cuatro rasgos específicos que, a continuación, se habrían buscado en los corpus en examen: densidad y variedad léxica (Laviosa, 1998), tamaño de las oraciones (Tai, 2015) y presencia de estructuras hipotácticas.

En la segunda fase, se identificaron en los textos en análisis los rasgos determinados empleando diferentes herramientas proporcionadas por los buscadores de concordancias *AntConc* (Anthony, 2014) y *WordSmith* (Scott, 2016): listados de palabras, para conocer el número de *types* y *tokens* necesarios para el cálculo de la densidad y la variedad léxica;

resultados estadísticos, para conocer la longitud media oracional; listados de concordancia, para la identificación de las estructuras subordinadas.

Tras aplicar los exámenes propuestos a cada uno de los dos corpus analizados, se compararon los resultados. Los datos corroboran la hipótesis principal, demostrando que las opiniones escritas en lengua nativa antes del 2004 presentan una mayor complejidad estilística y una mayor fluencia que aquellas producidas después del 2004 en una segunda lengua. Estos resultados, además de despertar interés, desde una perspectiva conceptual, por el impacto que tuvo la introducción de la reforma lingüística de 2004 en la función de las opiniones de los Abogados Generales y, por consiguiente, en la jurisprudencia europea, permitieron comprobar la funcionalidad de la metodología adoptada. El alcance del objetivo propuesto en cuanto a la complejidad estilística y la fluencia de los textos, comprueba la funcionalidad de la aplicación de la lingüística de corpus para el estudio de fenómenos cualitativos. Considerado el estado aún muy debatido de la cuestión, esta investigación pretende defender esta posición y sugerir ulteriores trabajos desde la misma perspectiva para ampliar las posibilidades de aplicación de la lingüística de corpus como metodología funcional para el alcance de objetivos cualitativos.

References

Anthony, L. (2014). AntConc (Version 3.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/.

Baker, M. (1993) Corpus Linguistics and Translation Studies – Implications and Applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.) (1993) Text and Technology: In Honour of John Sinclair (pp. 233-250). Amsterdam and Philadelphia: John Benjamins.

Laviosa, S. (1998). Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. META 43(4), 557-570.

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

Tai, H. Y. (2015). Writing development in syntactic complexity, accuracy and fluency in a content and language integrated learning class. International Journal of Language and Linguistics, 2(3), 149-156.

El corpus como herramienta para la comparación de los aspectos lingüísticos entre las traducciones y el texto original de *Viaje al Oeste*

Tian Mi

Nankai University, China

Keywords: *Viaje al Oeste; traducción; corpus; español-chino*

Abstract

Viaje al Oeste es una de las obras literarias más clásicas y conocidas de China. Trata del viaje de un grupo de peregrinos formado por un monje y sus tres discípulos, un mono poderoso, un cerdo y un espíritu que van hacia al Oeste para recibir los sutras budistas,

en el proceso del cual tienen que vencer 81 problemas. La historia se basó en un hecho histórico real de la dinastía Tang (618-907 d.C.) y se transmitía boca a boca, incorporándose la imaginación creadora del pueblo chino, hasta que finalmente se recopiló y se elaboró en forma narrativa por el literato Wu Cheng'en en la dinastía Ming (1368-1644).

Esta novela china contiene una variedad de culturesmas propios y formas narrativas únicas, lo que se constituye en diversas dificultades en la transmisión intercultural. A pesar de eso, hoy en día, tenemos dos traducciones completas accesibles en los mercados chino y español, las cuales nos sirven como buenos materiales para realizar una investigación comparativa. Además, debido a la gran cantidad de datos a los que se tenga que enfrentar, la construcción de un corpus nos ha presentado como una metodología válida para sacar las características de las traducciones.

Este estudio se basa en una comparación, tanto entre las dos versiones existentes de traducción al español de la dicha obra, como entre cada una de estas dos y la novela original en chino, mediante el análisis sobre el corpus paralelo y bilingüe español-chino construido por nuestro equipo de investigación. Deseamos, en primer lugar, compartir los problemas y dificultades que nos hemos hallado en el proceso de la construcción del mismo corpus, así como en la digitalización de los textos escritos, la depuración de los textos digitalizados, la alineación de los textos paralelos y las anotaciones necesarias para el fin de investigación. En segundo lugar, estamos interesados en sacar los datos estadísticos desde el corpus para comparar los aspectos lingüísticos del texto origen y los textos meta y, en tercer lugar, con base en esta comparación, analizar las diferencias estilísticas de los distintos traductores y centrarnos en los rasgos de simplificación y explicitación, así como otros enfoques traductológicos.

References

Almela, M., From words to lexical units. A corpus-driven account of collocation and idiomatic patterning in English and English-Spanish. Frankfurt a. M.: Peter Lang, 2006.

Anónimo, Viaje al Oeste: las aventuras del Rey Mono. Trad. Enrique P. Gatón e Imelda Huang-Wang. Madrid: Ediciones Siruela, S. A., 2006.

Baker, M., "Towards a Methodology for Investigating the Style of a Literary Translator", Target, 12:2, 2000, pp. 241-266.

-------, "The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators", International Journal of Corpus Linguistics, 4:2, 1999, pp. 281-298.
-------, "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research", Target, 7:2, 1993, pp. 223-243.

-------, (1993). "Corpus Linguistics and Translation Studies—Implications and Applications." In Baker et al., eds., pp. 233-250.

-------, (ed.), The Routledge Encyclopedia of Translation Studies, Londres, Routledge, 1998.

Corpas Pastor, Gloria, Investigar con corpus en traducción: los retos de un nuevo paradigma. Fráncfort: Peter Lang, 2008.

Halliday, M. A. K., An Introduction to Funtional Grammar. London: Edward Arnold, 1976.

Laviosa, S., Corpus-Based Translation Studies: Theory, Findings, Applications. Amsterdam: Rodopi, 2002.

Munday, J., Introducing Translation Studies: Theories and Applications. London and New York: Routledge, 2001.

Newmark, P, A Textbook of Translation. Shanghai: Shanghai Foreign Language Education Press, 2002.

Nida, E. A. and Taber, C. R., The Theory and Practice of Translation. Leiden: E. J.

Brill: 164-165, 1969.

Cheng'en Wu, Peregrinación al Oeste. Trad. María Lacea y Carlos Trigoso Sánchez. Beijing: Ediciones en Lenguas Extranjeras, 2010.

Latest developments in Western Balkan discourses on Europe and how to study them?

Piotr Mirocha

Institute of Slavonic Philology, Jagiellonian University, Krakow, Poland

***Keywords**: Serbia; Croatia; European discourse; political discourse; collocation analysis; LDA topic modelling; exploratory factor analysis*

Abstract

While corpora are widely employed to study syntax or lexicology, their use in analysis of political discourses is rarer. Here one of the most important tasks is to identify topoi occurring in an examined set of texts. In case of this presentation the question posed was what topoi are used to talk about Europe (Croat. *Europa*, Serb. *Evropa*) in the most recent Croatian and Serbian press discourses of various ideological profiles. With this goal in mind, diverse corpus linguistic tools were applied to the collection of texts: collocation analysis (Cf. e.g. Baker et al. 2008), LDA topic modelling (Cf. e.g. Törnberg / Törnberg 2016) and exploratory factor analysis (EFA) (Cf. e.g. Ajšić 2016). The obtained results were interpreted and evaluated in order to find out to what extent it is possible on the basis of the above-mentioned methodological procedures to reconstruct what topoi are employed in the Croatian and Serbian press of differing ideological profiles to write about Europe.

The analysed corpus consisted of articles scrapped from 2 online issues of Croatian newspapers – *Večernji list,* presenting an articulately conservative ideological profile, and *Novi list*, Rijeka-based left-liberal, pro-European newspaper. Two online issues of Serbian newspapers were also taken into account – *Politika*, of a centre-right orientation, and *Danas,* Vojvodina-centred liberal, pro-European newspaper. This gave a randomly downsampled total of ca. 40,000 articles published from 2012 to 2017 included in a corpus and lemmatised.

Then corpus linguistic tools were employed in order to reconstruct the 'European' topoi appearing in the text collection. Collocation analysis focused on the discursive prosody of a lexeme *Europa/Evropa.* LDA topic model was trained on articles containing at least 2 occurrences of a character string *-europ/evrop-*. In EFA the most common collocations

of the lexeme *Europa/Evropa* were a starting point to generate factors and find observations (texts) most typical to them. Each of the methods was applied in a temporalised manner, i.a. the whole corpus was subdivided into yearly subcorpora, which should enable to observe temporal change in Croatian and Serbian European discourses of conservative and liberal ideological profiles. The goal of the presentation is to identify advantages and disadvantages of the discussed approaches.

Last but not least, it was also one of the research objectives to find out how change in topoi use is determined by external, non-linguistic factors. In the analysed time-period, it seems that 2015, marked with an acute refugee crisis, was a turning point, when many previously less represented topoi (e.g. xenophobic or strongly antiEuropean) started to emerge in the mainstream press. The conclusion will show which of the described methods depicts this finding best and, in result, how to improve methodological instruments for corpus-driven study of political discourses.

References

Ajšić, A. (2016). English, 'polyglot' politicians and polyglot businessmen: Language ideologies in contemporary Bosnian press. In L. Buckingham (Ed.), The position of English in Bosnia-Herzegovina (pp. 159-202). Bristol: Multilingual Matters.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & Society, 19(3), 273-306.

Törnberg, A. & Törnberg, P. (2016). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum, Discourse & Society, 27(4), 401-422.

Compiling a corpus of financial reports in Spanish

Antonio Moreno-Sandoval[a]; Ana Gisbert;[b] Helena Montoro Zamorano[a]

[a]*Computational Linguistics Lab, Universidad Autónoma Madrid; Accounting Department, Universidad Autónoma Madrid*

***Keywords**: financial texts, communicative narrative, Spanish*

Abstract

The process of automatically analyzing textual content in Computational Linguistics is called Text Analytics. This process can be widely applied to different fields: from analyses of social network messages to information extraction from legal, medical or financial texts. Text Analytics' main challenge is to understand the content of the linguistic utterances and to show relevant information. In order to achieve these goals, different techniques are used, including statistics (data mining) or rule-based procedures. Our approach is based on the Computational Linguistics traditional method: we annotate the relevant information appearing in non-structured texts through domain-specific rules

and lexicons. Then, we analyze such information in terms of quantity and quality by using corpus linguistics tools.

The first step then is to collect a corpus of financial reports. This paper describes the process we have followed to compile raw material in pdf from the websites, to export into txt format, to clean the output and create a textual dataset with metadata (Company name, year, sector). The standard financial report in Spanish does not follow fixed style guidelines. Out of 350 downloaded reports, they all share at least several sections: Chairman's Statement, Corporate Government Report and Financial Review. However, the last two sections differ greatly in size and format between reports. In addition, they contain a lot of numerical data. Our goal is to study financial narrative. For this reason, we decided to start with letters from the Chairman. They belong to an informative and communicative register, in contrast to the other two much more technical sections.

For the selection of the reports, the 125 companies listed on the Spanish stock exchange were chosen. 32 companies did not show the annual report on their website. To process the pdf documents, we used the CFIE-FRSE tool (El-Haj et al. 2014, El-Haj et al. 2016) developed by UCREL for English. We adapted it to Spanish section names. In general, acceptable results were achieved, but all texts had to be processed manually. Over a period of four months, 10 linguists, including 7 students, cleaned the texts and copied parts that were not processed by the tool.

The final corpus contains all the Chairman's letters from 82 companies in the annual reports from 2014 to 2017. The final corpus comprises 318 texts (includes some texts of companies that do not have a complete set of 4 years) and over 600000 words. Each text has one sentence per line, which will facilitate its processing by a PoS tagger (Grampal) and a semantic tagger (USAS, http://ucrel.lancs.ac.uk/usas/). In addition to the morphosyntactic and semantic annotation, an online search web is planned for next year.

This new corpus will be the basis for terminological studies and analyses of financial discourse in Spanish.

References:

El-Haj, Mahmoud, Rayson, Paul, Young, Steven, and Walker, Martin (2014). "Detecting Document Structure in a Very Large Corpus of UK Financial Reports". In The 9th edition of the Language Resources and Evaluation Conference, 26-31 May 2014, Reykjavik, Iceland.

El-Haj, Mahmoud and Rayson, Paul and Walker, Martin and Young, Steven and Alves, Paulo (2016). "Heterogeneous Narrative Content in Annual Reports Published as PDF Files: Extraction, Classification and Incremental Predictive Ability". Available at SSRN: https://ssrn.com/abstract=2803275 or http://dx.doi.org/10.2139/ssrn.2803275

Possibility and necessity in financial narrative: a study of modal adverbs in Spanish

Antonio Moreno Sandoval

Department of Linguistics, Universidad Autónoma Madrid, Spain

***Keywords**: modality, financial texts, communicative narrative, Spanish*

Abstract

The study of modality in terms of possibility and necessity has a long tradition in Linguistics and Logic. Recently, it has become a topic of great interest in Computational Linguistics for its applications to textual analysis (*Text Analytics*), both in the analysis of messages in social networks (*Opinion Mining and Sentiment Analysis*) as well as in the information extraction of relevant content. Detecting modality markers that express the possibility or need for a given event helps greatly in understanding the goals and intentions of the text's author.

Of the four modality markers (auxiliary verbs, verbal mood, adjective and adverbs) we will focus on the latter. From a computational point of view, adverbs are simple to recognize automatically because they are unambiguous, especially those ending in *-mente*. Adverbs can express both necessity and possibility, but only epistemic modality, that is, defining the certainty of the event. On the other hand, deontic modality is carried out in the form of a command, prohibition or suggestion, something that is typically expressed with the verb in imperative or by means of verbal periphrasis.

In this article we will study the manner adverbs in a corpus of financial reports. To do this, we will use two tools. First, OJIME (Herrero 2017) is an online modality tagger that annotates and classifies those auxiliaries, adverbs, adjectives and verb mood that encode information about possibility and necessity. OJIME also compares the normalized frequency (*Fn*) of the text against the Fn in a spoken corpus of Spanish. On the other hand, we use LYNEAL, a concordancer with statistical measures, to analyze the cases and quantify the differences with respect to the Fn of those words in the CREA.

We will use two types of financial texts:

1. The "Chairman's letters" subcorpus: taken from the first sections of companies' annual reports. These are the informative-style texts written by the presidents of the enterprises in order to communicate to their shareholders the annual results of the company. It consists of about 300 texts and half a million words.
2. The financial analysis subcorpus: taken from the rest of the annual reports and written in a more technical register. It contains over 20 million words.

Both corpora are part of the FinT-esp project (Financial Texts in Spanish), which aims to develop, on the one hand, tools for terminology annotation and semantic tagging and, on the other, the analysis of financial discourse.

On the results, as expected, the frequency of the modal markers in general (as measured by OJIME) is much lower in financial texts than in spoken texts. We have found 806 different types of adverbs in *-mente* in the large corpus. From them, those expressing abrupt changes in behavior are barely used: "desgraciadamente" (unfortunately), "súbitamente" and "repentinamente" (suddenly), "forzosamente" (forcibly). On the other side, adverbs indicating regular temporality and safety are overused: "adicionalmente" (additionally), "periódicamente" (periodically), "anualmente" (annually), "razonablemente" (reasonably) [2] (see attached table and figure).

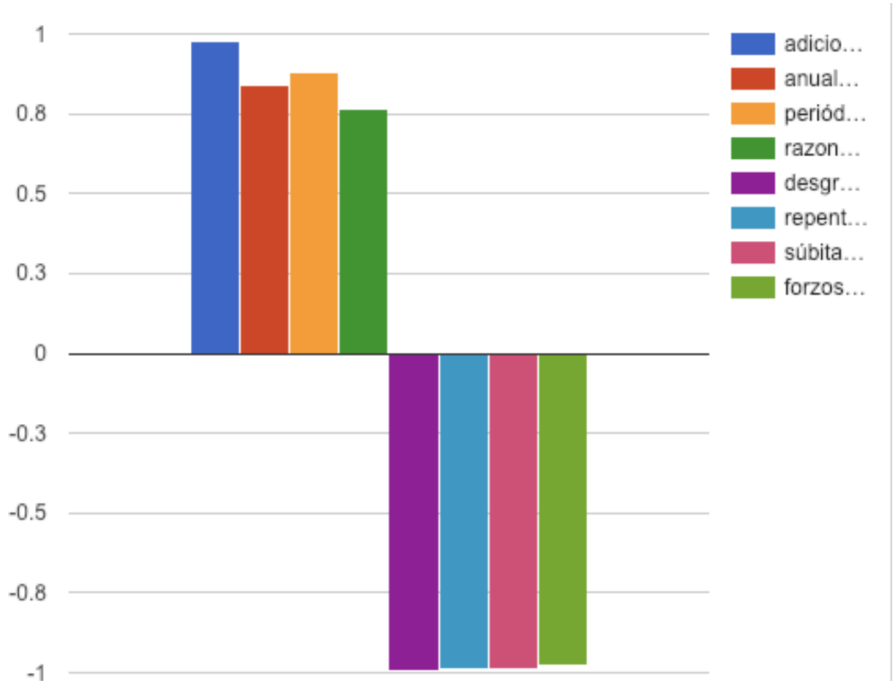
[2] The comparison is made as a function of the normalized frequency per million words of CREA, using the formula (N – C) / (N + C), where N is the normalized frequency of the word *i* in the financial corpus and C is the normalized frequency of the same word *i* in CREA. The range goes from 1 to -1, being 1 the most

■ Números: 20. (N-C) / (N+C) - 1. Puntuación

Número de filas: 8 / Número de columnas: 1

Ventana / Copiar con formato / Copiar sin formato

→	(N-C)/(N+C)	Todos	*Reiniciar*
1	adicionalmente	0.979	adicionalmente
2	anualmente	0.842	anualmente
3	periódicamente	0.879	periódicamente
4	razonablemente	0.766	razonablemente
5	desgraciadamente	-0.992	desgraciadamente
6	repentinamente	-0.986	repentinamente
7	súbitamente	-0.989	súbitamente
8	forzosamente	-0.980	forzosamente



References

Herrero Zorita, Carlos (2017): *Contrastive Study of Modality and Negation in Spoken Spanish and Japanese.* PhD thesis, Universidad Autónoma de Madrid.

Gendered voices in EFL course books: A multimodal CADS analysis

extreme value for a word very used in the financial corpus and -1 the most extreme value for a word much more used in the CREA corpus. The results were automatically generated by LYNEAL.

Laura Mercé Moreno Serrano

Department of English and German, University of Valencia, Spain

***Keywords**: Feminist CDA; EFL course books; Compulsory Secondary Education; multimodality; Appraisal Theory; interpersonal communication*

Abstract

As a result of a fast-paced globalised western world, all means of communication have experienced an increase of visual semiotic resources over the past years (Kress, 2010; Kress & Van Leeuwen, 2006). Correspondingly, scholarly research has also paid special attention to multimodality in pedagogical materials, as it is the case of EFL course books (Yassine, 2014). Against the background of free-stereotyped, non-sexist education fostered by Spanish and Valencian legislation (Spanish Official Gazette nº7899, 2006; nº12886, 2013; nº 37, 2015; Valencian Department of Education, 2015), the present study aims to analyse the actual representation of gender equality in course materials, as promoted in the aforementioned legal documents. Thus, I analyse multimodal resources as interpersonal meaning-making actors of gender stereotypes in three EFL course books used in some Compulsory Secondary Education (CSE) schools in the region of Valencia. Following CADS, I combine Corpus Linguistic tools with Systemic-Functional Multimodal Discourse Analysis (Kress & van Leeuwen 2006; O'Halloran & Lim Fei, 2014), and Appraisal Theory (Martin & White, 2005) in order to examine the gendered (linguistic and visual) discourses of EFL course books as modal affordances (Benzemer & Kress, 2008), whose analysis may result in a powerful asset for a "holistic feminist critique of the discursive construction of gender" (Lazar, 2005).

After reviewing the course books of 86 educational centres in the region of Valencia, I selected the three most used EFL books in the 4th year of CSE as a synchronic corpus. I

then focus on the semiotic complex of ten reading sections from each course book, leading to a total amount of thirty modal affordances. Methodologically, I adopt a mixed-methods approach to unveil traces of (gendered) discourse prosodies in these thirty texts and gain insights into the way linguistic elements lead to a dialogic engagement with the addressee. Qualitatively, Kress and Van Leeuwen's multimodal framework (1996) is used to uncover how gendered subjects are represented in these visuals and how they interact with the textual resource (Norris, 2004). Results point towards the existence of polarised discourses in the representation of men and women in terms of agency, social roles and personality.

References

Bezemer, J. & Kress, G. (2018). Writing in Multimodal Texts A Social Semiotic Account of Designs for Learning. Written Communication, 25(2). Doi: 10.1177/0741088307313177.

Chisholm, L. (2018). Representations of Class, Race, and Gender in Textbooks. In E. Fuchs & A. Bock, The Palgrave Handbook of Textbook Studies (pp. 225-237). New York: Palgrave Macmillan. Retrieved from https://link.springer.com/book/10.1057%2F978-1-137-53142-1.

Generalitat Valenciana (2015). DOGV 87/2015, Currículum General de l'Educació Secundària Obligat ria i Batxillerat a la Comunitat Valenciana [Ebook]. Retrieved from http://www.gva.es/downloads/publicados/2015_5410.pdf.

Jefatura del Estado (2006). BOE-A-2006-7899 [Ebook]. Retrieved from https://www.boe.es/buscar/pdf/2006/BOE-A-2006-7899-consolidado.pdf

Jefatura del Estado (2013). BOE-A-2013-12886 [Ebook]. Retrieved from https://www.boe.es/buscar/pdf/2013/BOE-A-2013-12886-consolidado.pdf

Klerides, E. (2010). Imagining the Textbook: Textbooks as Discourse and Genre. Journal of Educational Media, Memory, and Society, 2(1). Doi: 10.3167/jemms.2010.020103.

Kress, G., & van Leeuwen, T. (2006). Reading Images - The Grammar of Visual Design (2nd ed.). New York: Rouledge.

Kress, G. (2010). Multimodality, A social semiotic approach to contemporary communication. New York: Routledge.

Lazar, M. (2005). Feminist critical discourse analysis: gender, power, and ideology in discourse. New York: Palgrave Macmillan.

Lim Fei, V., & O'Halloran, K. (2014). Systemic functional multimodal discourse analysis. In U. Jessner & C. Kramsch, Interactions, Images and Text, A Reader in Multimodality. Germany: De Gruyter. Retrieved from https://www.researchgate.net/publication/306003898_Systemic_Functional_Multimodal_Discourse_Analysis.

Martin, J., & White, P. (2005). The Language of Evaluation: Appraisal in English. New York: Palgrave Macmillan.

Ministerio de Educación, Cultura y Deporte. (2015). BOE-A-2015-37 [Ebook]. Retrieved from https://www.boe.es/buscar/pdf/2015/BOE-A-2015-37-consolidado.pdf

Norris, S. (2004). Analyzing multimodal interaction: a methodological framework [Ebook]. New York: Routledge. Retrieved from http://ebookcentral.proquest.com/lib/univalencia/detail.action?docID=182714.

Segunda Lengua Extranjera ESO y Bachillerato (2015) (1st ed.). Retrieved from http://www.ceice.gva.es/documents/162640733/162655315/Segunda+Len gua+Extr.+%28PDF%29/7dd177b9-2455-44c9-b859-b54900de3313.

Yassine, S. (2014). Multimodal Design of EFL Textbooks: A Social Semiotic Multimodal Approach. Anglisticum Journal, 3. doi: 10.0001/(aj).v3i12.1523.g2031.

Addressing the audience: second person pronouns in CHET and CECheT

Isabel Moskowich[a]; Begoña Crespo[a]

[a]*Department of Letters, University of A Coruña, Spain*

Keywords: *Late Modern English; scientific discourse; readership; personal pronouns*

Abstract

One of the many linguistic features that reveal the presence of the audience in writing (Biber, 1988; Hyland, 1996; Atkinson, 1999) is the use of pronouns. Second person pronouns are particularly important as they, as central pronouns (Quirk, 1985; Chamonikolasová, 1991) are especially revealing of the author`s intention to get the reader involved in the message conveyed.

Stemming from the idea that scientific English is detached and object-centred (Atkinson, 1999) our aim is twofold: on the one hand, we want here to ascertain the inclusive character of the pronoun *you* and whether this characteristic is maintained across time in general terms. For the purpose of this piece of research, we have not considered dialectal forms of the pronouns such as *thou* (Wales, 2003) as this variant is not expected to be found in standardised texts as scientific ones are supposed to be. As a matter of fact, authors such as Gramley & Patzold (1992) have to admit that *thou* has not completely disappeared but is clearly 'marginalised' from 'common core' usage. On the other hand, being more specific, we will compare the results obtained in both the History and Chemistry texts compiled for the Coruña Corpus of English Scientific Writing. The reason for this is that we are conscious that they represent different writing traditions, the soft and the hard sciences in Hyland's words (1996), and want to show the differences. As the tendency to depersonalise scientific discourse seems to grow as time goes by, we will compare the evolution of the use of the second person pronoun in the eighteenth and nineteenth centuries. The results might be indicative of the transition from a "personalised" and audience-inclusive discourse to a more detached writing style. We will also consider other variables that may exert some kind of influence on the use of pronouns such as the sex of the author, the sort of discipline in which the forms under study occur. In fact, we expect to find a decreasing frequency of occurrence of *you* forms in the nineteenth century. Likewise, our initial hypothesis is that the *Corpus of History English Texts* (CHET) will contain a higher proportion of these forms than the *Corpus of English Chemistry Texts* (CECheT). This hypothesis is based on the idea that the disciplines in the Humanities are generally more dialogic that those in the Hard Sciences.

The material we will be using in this study forms part, as already mentioned, of the socalled *Coruña Corpus of English Scientific Writing*. Two subcorpora have been selected on this occasion, namely, The Corpus of History English Texts (CHET) and the Corpus of English Chemistry Texts (CECheT). They contain text extracts belonging to works published in the eighteenth and nineteenth centuries. Both of them are still beta versions at the moment of writing this abstract. Some of the characteristics of these subcorpora such as structure, organisation and markup will be also described in detail (Crespo and Moskowich, 2010).

References

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

Bozic, L. D. (2016). Gender differences in the personal pronouns usage on the corpus of congressional speeches. Journal of Research Design and Statistics in Linguistics and Communication Science, 3(2), 161-188. doi: 10.1558/jrds.30111.

Chamonikolasová, J. (1991). Pronouns, functional sentence perspective and

imntonation. Brno Studies in English, 19, 55-64.

Crespo, B. & Moskowich, I. (2010). CETA in the Context of the Coruña Corpus. Literary and Linguistic Computing, 25(2), 153-164.

Gramley, S. & Patzold, K.M. (1992). A Survey of Modern English. London: Routledge.

Hyland, K. (1996). Talking to the Academy: Forms of Hedging in Science Research Articles. Written Communication, 13(2), 251–281.

Quirk, R. et al. (1985). A Comprehensive English Grammar. London: Longman.

Wales, K. (2003). Second Person Pronouns in Contemporary English: The End of a Story or Just the Beginning? Franco-Bristish Studies, 33-34, 172-185.

A corpus based contrastive analysis: the use of boosters in engineering and medicine academic articles in English

Rut Muñiz Calderón[a]; Chiara Tasso[a]

[a]*Departamento de Idiomas, Universidad Católica de Valencia San Vicente Mártir, Spain*

Keywords*: academic writing; boosters; metadiscourse; native speakers*

Abstract

This study presents a contrastive analysis of the use of verbs and adverbs acting as boosters in engineering and medicine research papers in English and has been carried out within the frame of the *IAMET project where a large contrastive corpus of Spanish and English journal articles from different disciplines has been gathered. A great number of authors such as Mauranen (1993), Hyland and Tse (2004), Dahl (2004), Hyland (2005), Mur Dueñas (2011) and Carrió (2015) have analyzed the importance of the metadiscourse devices in the written academic discourse.

In this paper we focus on boosters which are metadiscourse interactional devices used to express the authors' certainty and conviction about a proposition and which may be the result of certain pragmatic conventions in academic writing.

The reason behind this study is the lack of research in contrasting the use of boosters in these specific disciplines. Taking this into consideration the hypothesis of this contrastive analysis is to verify that the use of boosters varies according to the specific field of study.

With this in mind, the main objective of this paper is to identify and compare the use and distribution of the above mentioned metadiscourse interactional features using Hyland's (2005) and Mur-Dueñas (2011) taxonomy as a study model in order to be able to analyse and differentiate the use of boosters in engineering and medicine research articles. The specific objectives are the following: to tag all boosters used in both corpora, to detect the most frequently used boosters and detect similarities and differences in both corpora and to demonstrate Metool tagging tool efficiency.

In order to meet these objectives, a corpus of one hundred and thirty-two scientific papers was compiled. The corpus was composed by sixty-five full research papers from engineering and sixty seven from medicine, all of them written by English native authors

with a total of 992,943 words. The research articles were selected from JCR academic journals published in the last two years in the fields of medicine and engineering.

The methodology used to achieve the stated objectives was a combination of quantitative and qualitative analysis of the corpus. Thus, the occurrences were extracted from the corpora and contrasted to show synchronic variation in academic discourse using the tagging tool Metool. Within the analysis of boosters, a distinction between two typologies was made: the analysis of adverbs and that of verbs taking Hyland's (2005) and Mur-Dueñas (2011) classifications as a reference. Both corpora were used to analyze variation and identify the specific features of the use of both adverbs and verbs acting as boosters in medicine and engineering. The overall results obtained from the analysis of full research papers from engineering and medicine disciplines verified our initial hypothesis: both quantitative and qualitative variation exists in the use of boosters adverbs and verbs depending on the scientific domain and this may be the result of conceptual and organizational differences among disciplines. Finally, conclusions were drawn.

*IAMET Project is a competitive project granted by the Spanish Ministry of Economy and Competitiveness (Proyecto: FFI2016-77941-P) entitled: Identification and analysis of rhetoric elements in Spanish and in English: study of metadiscursive strategies.

El contexto interactivo concreto de la actividad atenuadora en el corpus AMERESCO: acerca de los segmentos desencadenantes, atenuantes y atenuados

Danny Fernando Murillo Lanza

Hispanic Studies Department, Universitat de València, Spain

Keywords: *Atenuación; Corpus Ameresco; Contexto Interactivo Concreto; Desencadenantes; Atenuantes; Atenuados; Español de España e Hispanoamérica; Conversación coloquial; Análisis del Discurso*

Abstract

Uno de los problemas más importantes que ha supuesto el estudio de la atenuación lingüística tiene que ver con su reconocimiento. Dado que se necesita un marco interactivocomunicativo real, como cualquier fenómeno pragmático, la atenuación se reconoce en el contexto (Briz & Albelda, 2013). Sin embargo todavía se observa cierta debilidad para reconocer dicho fenómeno (Briz & Albelda, 2013).

Ante tal panorama, autores como Albelda Marco (2010), Briz & Albelda (2013) y, recientemente, Albelda & otros (2014) han establecido una serie de criterios y directrices que pueden seguirse para identificar con mucha mayor certeza aquellos casos que realmente manifiestan un valor atenuante. Al respecto, estos autores señalan que para reconocer la atenuación es necesario, entre otras cuestiones, describir el contexto interactivo concreto en el que se desarrolla la actividad atenuadora. Se entiende por contexto interactivo concreto a la combinación de los siguientes segmentos: el miembro del discurso atenuado, el miembro del discurso desencadenante, la expresión atenuante concreta y, en su caso, el efecto que produce (Briz & Albelda, 2013).

De ahí que, esta comunicación tiene como objetivo responder las siguientes preguntas: (1) qué patrones interactivos, más o menos regulares, se forman entre los elementos que forman el marco interactivo concreto de la actividad atenuadora en la conversación coloquial de diversas zonas de España e Hispanoamérica; y, (2) qué elementos desempeñan con mayor frecuencia dicho marco interactivo.

En concreto, el análisis de este estudio parte del corpus Ameresco (Albelda & Estellés, 2017). Puntualmente, se ha elaborado un subcorpus de aproximadamente 10 conversaciones coloquiales de las siguientes zonas: Tucumán, Santa Cruz, Iquique, Medellín, La Habana, Valencia, Ciudad de México, Ciudad de Panamá. Seguidamente, se ha analizado el corpus prestando especial atención a los elementos que componen el contexto interactivo concreto de la actividad atenuadora. Finalmente, para obtener los resultados se ha efectuado un análisis cualitativo de la atenuación y se han aplicado algunas pruebas estadísticas como el Análisis Múltiple de Correspondencias, entre otras.

Nuestros resultados, quizá, pueden ser importantes para establecer una tipología y una sistematización inicial de los patrones interactivos y elementos más frecuentes que aparecen en los contextos interactivos concretos de la actividad atenuadora en la conversación coloquial de algunas zonas de España e Hispanoamérica. Por otro lado, permitirían observar algunos puntos de convergencia y divergencia entre las variedades lingüísticas estudiadas.

References

Albelda Marco, M. (2010). ¿Cómo se reconoce la atenuación? Una aproximación metodológica. En F. Orletti, & L. Mariottini, (Des)cortesía en español. Espacios teóricos y metodológicos para su estudio. (págs. 47-70). Roma-Estocolmo: Università degli Studi Roma Tre-EDICE.

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

Albelda Marco, M., & Estellés Arguedas, M. (2017). Corpus Ameresco. Recuperado de: www.esvaratenuacion.es

Briz, A., & Albelda, M. (2013). Una propuesta teórica y metodológica para el análisis de la atenuación lingüística en español y portugués. La base de un proyecto común (ES.POR.ATENUACIÓN). Onomazéin, 288-319.

Briz Gómez, A. (2002). Corpus de conversaciones coloquiales. Arco Libros. Recuperado de: https://dialnet.unirioja.es/servlet/extlib?codigo=89115

Cabedo, A., & Pons, S. (2013). Corpus Val.Es.Co 2.0. Recuperado de www.valesco.es

Gómez, J. R. (2001). El español hablado de Valencia. Materiales para su estudio. I. Nivel sociocultural alto. Valencia: Universitat de València.

Gómez, J. R. (2005). El español hablado de Valencia. Materiales para su estudio. II. Nivel sociocultural medio. Valencia: Universitat de València.

Mini-Corpus del español oral: un recurso para el estudio de la articulación de la información

Carlota Nicolás Martínez

Dipartimento di Formazione, Lingue, Intercultura, Letterature e Psicologia, Università degli Studi di Firenze, Italia

***Keywords**: estructura de la información; lingüística de corpus; español oral; segmentación; análisis prosódico; pragmática*

Abstract

El aspecto prosódico para el estudio del habla es esencial, si bien es uno de los aspectos menos estudiados. Poniendo esencialmente en relación prosodia y acto de habla se ha desarrollado la teoría *Language into Act Theory L-AcT* (Cresti 2000; Cresti 2018) que no solo se basa en reflexiones teóricas, sino que ha evolucionado mediante la comprobación práctica en miles de enunciados en varias lenguas.

El objetivo de esta ponencia es mostrar el contenido y el uso del Mini-corpus del español oral que en octubre del 2018 se incluyó en la base de datos DB-IPIC junto a los Minicorpus del italiano (Panunzi & Gregori 2012) y del portugués brasileño (Panunzi & Mittmann, 2014; Raso & Mello 2010, 2012; Mittmann & Raso 2012). Estos tres minicorpus comparables están etiquetado gracias al trabajo del grupo de investigación de LABLITA siguiendo la teoría *Language into Act Theory L-AcT*. En este marco teórico el Enunciado corresponde a un Acto de Habla; para la teoría *Language into Act Theory* (Cresti 2000; Cresti & Moneglia 2010) un Enunciado es la unidad de referencia para el estudio del habla espontánea; el Enunciado tiene propiedades prosódicas y, por tanto, "The criterion for the segmentation of the speech flow into utterances is prosodic. Each prosodic break qualified as terminal defines the utterances limits in the speech flow" (Moneglia & Cresti 2006: 92) (Autor 2018).

En el proceso de elaboración del Minicorpus primero se ha escuchado la grabación para hacer la transcripción ortográfica y delimitar y etiquetar las unidades prosódicas; se ha observado que existen secuencias prosódicas terminadas, que pueden estar compuestas por secuencias prosódicas no terminales y no independientes. Sucesivamente se ha decidido qué unidad informativa, según sus características, correspondía a cada una de estas unidades prosódicas. Para el etiquetado de las unidades informativas, ha sido imprescindible la observación en la pantalla de la curva melódica de cada unidad prosódica, ya que las transcripciones están alineadas con la voz con el programa WinPitch.

Por último, se han analizado todos los textos con PoS antes de cargarlos en la Base de datos de IPIC de manera que el Mini-corpus del español sea interrogable.

Como se ha dicho a cada una de estas unidades prosódicas, le corresponde una unidad de la articulación de la información. Las secuencias terminadas pueden ser Enunciados y Estancias, estas son las dos únicas unidades articulatorias referenciales del habla según la *Language into Act Theory L-AcT*.

Las unidades informativas o de la articulación de la información pueden ser de tres tipos:
- las unidades informativas textuales nucleares cuyo papel es realizar el contenido semántico y proposicional del Enunciado o de la Estancia;

- las unidades informativas textuales no nucleares, que forman parte de la composición del contexto semántico del enunciado;

- las unidades informativas dialógicas, cuyo papel está relacionado con el desarrollo de la interacción, entendida como la realización del intercambio comunicativo.

Hay otras unidades que marcan partes del discurso, pero no tienen ningún valor informativo.

Se verá como a partir de este análisis de la estructura de la información de la lengua oral se pueden hacer estudios muy interesantes.

References

Cresti, E. (2000). Corpus di italiano parlato, 2 voll., CD-ROM. Firenze, Italia: Accademia della Crusca.

Cresti, E., & Moneglia, M. (2010). Informational patterning theory and the corpus-based description of spoken language. The compositionality issue in the topic-comment pattern. En M. Moneglia, & A. Panunzi (Eds.). Bootstrapping Information from Corpora in a Cross-Linguistic Perspective. Firenze, Italia: Firenze University Press.

Cresti, E., & Moneglia, M. (2018). The definition of the Topic within Language into Act Theory and its identification in spontaneous speech corpora. En M. Borreguero Zuloaga, V. Atayan, & S. Grosse (Eds.), Revue Romane, 53(1), 30-62.

Moneglia, M., & Cresti, E. (2006). C-ORAL-ROM. Prosodic Boundaries for Spontaneous Speech Analysis. En Kawaguchi, Y., S. Zaima, & T. Takagaki (Eds.), Spoken Language Corpus and Linguistics Informatics (pp. 89-112). Amsterdam, Holanda: Benjamins.

Nicolás Martínez, C., & Lombán Somacarrera, M. (2018). MINI-CORPUS del español para IPIC. CHIMERA. Romance Corporta and Linguistic Studies 5 (2) 197-215. http://doi.org/10.15366/chimera2018.5.1.002

Panunzi, A., & Gregori, L. (2012). DB-IPIC: An XML database for the representation of information structure in spoken language. En Mello H. R., A. Panunzi, & T. Raso (Eds.), Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation (pp.133-150). Firenze, Italia: Firenze University Press.

Panunzi, A., & Mittmann, M. (2014). The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. En T. Raso, & H. Mello (Eds.), Spoken corpora and linguistic studies (pp. 189-227). Amsterdam, Holland: Benjamins.

Nueva versión de C-Or-DiAL (Corpus Oral Didáctico Anotado Lingüísticamente). Alineado y ampliado

Carlota Nicolás Martínez[a]; Marina Lombán Somacarrera[b]

[a]*Dipartimento di Formazione, Lingue, Intercultura, Letterature e Psicologia, Università degli Studi di Firenze, Italia;* [b]*Universidad Complutense de Madrid, Spain*

Keywords: *Corpus orales; Didáctica del español; Corpus alineados; Análisis del discurso*

Abstract

C-Or-DiAL (Corpus Oral Didáctico Anotado Lingüísticamente) es un corpus de habla de español espontáneo para ser utilizado en el aula de ELE. La primera versión es de 2012 http://lablita.it/app/Chttp://lablita.it/app/cordial/Or-DiAL/. Ese mismo año se publicó un libro de análisis detallado del corpus que también contenía reflexiones y propuestas sobre su uso, ofreciendo para ello las transcripciones y los audios (Autor 2012a). También se han publicado dos manuales de ELE usando como material los audios y transcripciones de C-Or-DiAL ((Autor 2012b) y (Autor *et al.* 2016)).

En los últimos años los corpus han empezado a ser considerados de enorme utilidad para la enseñanza del español, tanto es así que en ocasiones se olvidan los requisitos mínimos que un corpus lingüístico tiene que cumplir (Cresti & Panunzi 2016), ya que debe tener un soporte informático que facilite el acceso a los textos que deben estar adecuadamente etiquetados.

Muchos de los artículos o libros sobre la didáctica relacionada con los corpus que han sido publicados en los últimos años (Bailina 2014; Barros García *et al.* 2015; Barros 2012; Piñol 2012 y 2017; Yanira 2013; Tolchinsky 2014); algunos son los precursores anteriores a cuando se publicó C-Or-DiAL (Quiroz Herrera Pitkowski 2009; Campillos Llanos *et al.* 2007; Véliz Campos 2008; Barros 2011; Autor 2004, 2005, 2008, 2011, 2012).

A pesar del reconocimiento actual que se otorga a los corpus, no han llegado a ser material de uso frecuente en las aulas. Esta es una de las razones por las que se ha creído necesario crear una nueva versión de C-Or-DiAL, con las siguientes características respecto a la primera versión:

- Contiene aprox. 250.000 formas de lengua oral espontánea, lo que significa que ha aumentado 100.000 formas respecto a la primera versión.

- Cada uno de los enunciados transcritos se han alineado con el audio con ayuda de Winpitch. Para hacer este trabajo de etiquetado se siguieron las teorías desarrolladas por Cresti y Moneglia (Cresti

2000; Moneglia & Cresti 2006; Cresti & Moneglia 2010) y su grupo de investigación en el laboratorio LABLITA.

- Todas las palabras tienen su etiqueta de PoS, para este etiquetado se ha utilizando FreeLing: http://nlp.lsi.upc.edu/freeling/node/1.

- Todo el corpus es interrogable por forma o lema, y las concordancias ofrecen el acceso directo al audio. Para esto se ha adaptado la versión libre de SketchEngine. Con este analizador se ofrecen otras funciones como por ejemplo la consulta de la lista de palabras del corpus, la ampliación del contexto de las concordancias o la consulta de los

metadatos. Las búsquedas se pueden hacer con todas las posibilidades ofrecidas por este analizador.

Por lo tanto, el nuevo C-Or-DiAL mejora su accesibilidad, tanto a las palabras como al audio, para facilitar el uso en el aula y para poder crear material didáctico. La innovación principal es la extracción de concordancias con enunciados transcritos que contengan una determinada palabra o categoría gramatical pudiendo escuchar inmediatamente el audio. Además, se puede ver ese enunciado en un contexto más amplio, incluso ver en otra ventana la conversación entera en el que está la concordancia y consultar los metadatos de la conversación.

La otra razón de peso por la que se ha hecho el enorme esfuerzo de alinear y ampliar los textos de COr-DiAL es el deseo de ofrecer a quien esté interesado en el análisis de la lengua oral un recurso que facilite la investigación.

References

Bailini, S. 2014. Los corpus como recursos didácticos para la enseñanza de las variedades diatópicas del español. En FIAPE. V Congreso internacional: ¿Qué español enseñar y cómo? Variedades del español y su enseñanza: Cuenca 25-28/06-2014, 1-17. w.mecd.gob.es/dam/jcr:9879b92d9541-4b3f-b2bd-67a58a86f26e/3--los-corpus-como-recursos-didacticos-para-la-ensenanza-delas-variedades-diatopicas-del-espanol--bailinisonia-pdf.pdf.

Barros García, M. J., J. Morales Cabezas & P. Barros García 2015. Enseñanza de español real: corpus y guía didáctica, Cuadernos CANELA 26: 93-106. ISSN 1344-9109 http://cuadernoscanela.org/index.php/cuadernos/article/view/30/28.

Barros, Mª. J. 2011. La enseñanza del papel y rasgos de la cortesía valorizadora a partir del análisis de corpus. En J. de Santiago Guervós, H. Bongaerts, J.J. Sánchez Iglesias & M. Seseña Gómez (eds.), Del texto a la Lengua: la aplicación de los textos a la enseñanza/aprendizaje del español L2/LE. Actas del XXI Congreso Internacional de la Asociación para la Enseñanza del Español como Lengua Extranjera (ASELE), 213-224.

Barros, P. (ed.) 2012. COGILA. Español oral conversacional: corpus y guía didáctica. Granada: Editorial Universidad de Granada.

Campillos Llanos, L., P. Gozalo Gómez & A. Moreno Sandoval 2007. Las destrezas orales en la enseñanza del español L2-LE. XVII Congreso Internacional de la Asociación del Español como lengua extranjera (ASELE). Logroño 27-30 de septiembre de 2006, 1115-1128. https://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/17/17_1115.pdf. (acceso 19 de junio de 2018).

Cresti, E. & A. Panuzi 2013. Introduzione ai corpora dell'italiano. Il Mulino: Bologna.

Cresti, E. & M. Moneglia 2010. Informational patterning theory and the corpus-based description of spoken language. The compositionality issue in the topic-comment pattern. En M. Moneglia & A. Panunzi (eds) Bootstrapping Information from Corpora in a Cross-Linguistic Perspective. Firenze: FUP.

Cresti, E. 2000. Corpus di italiano parlato, 2 vol., CD-ROM. Firenze: Accademia della Crusca.

Curto García-Nieto N., A. González Ledesma, C. Nicolás Martínez, A. Valverde Mateos 2008. Diseño de una unidad didáctica con corpus del español oral. ¿Y si tú fueras el profesor?. En Nicolás Martínez, M. C. (ed.), 2008: 243-252.

Moneglia, M. & Cresti, E. 2006. C-ORAL-ROM. Prosodic Boundaries for Spontaneous Speech Analysis. En Kawaguchi, Y., S. Zaima & T. Takagaki (eds) 2006, 89-112.

Nicolás Martínez, C. 2016 Corte prosódico al principio del enunciado: la salida nula. En Sainz González, M. & Maria E. Sainz Gonzalez, I. Solís García, F. del Barrio de la Rosa, I. Arroyo Hernández (eds) Geométrica explosión Estudios de lengua y literatura en homenaje a René Lenarduzzi, Venezia: Biblioteca di Rassegna Iberistica 1: 1-22.

Nicolás Martínez, C. 2015 Análisis de las concordancias con etiqueta de reinicio en C-Or-DiAL: contextos dialógicos, LEA XVIII/2: 173-197.

Nicolás Martínez, C. 2015 Las funciones comunicativas en la lengua oral: validez y adecuación al alumnado. En Morimoto Y. & M.V. Pavón Lucero, R. Santamaría Martínez (eds) 2015, 1197-1207.

Nicolás Martínez, C. 2011. C-Or-DiAL corpus oral didáctico anotado lingüísticamente. Un corpus para la didáctica y la investigación. En A. M. Cestero, I. Molina, F. Paredes (eds) 2011. La lengua, lugar de encuentro: Actas del XVI Congreso Internacional de la ALFAL (Alcalá de Henares, 6-9 de junio de 2011). Universidad de Alcalá, Servicio de Publicaciones: Alcalá de Henares, 3217-3226.

Nicolás Martínez, C. 2014 Propuestas prácticas para el uso en el aula de C-Or-DiAL (Corpus Oral Didáctico Anotado Lingüísticamente). En Contreras Izquierdo N. M. 2014. La enseñanza del español como LE/L2 en el siglo XXI: XXIV Congreso Internacional ASELE. 903-913.

Nicolás Martínez, C. 2012. C-Or-DiAL (Corpus Oral Didáctico Anotado Lingüísticamente) y la enseñanza del español. En H. Mello, P. Massimo & T. Raso (eds) 2012. Proceedings of the VIIth GSCP International Conference. Speech and Corpora, Firenze University Press, Firenze. 93-97.

Nicolás Martínez, C. 2008. Creating an oral corpus for teaching purposes - with students of Spanish as authors. En 8th Teaching and Language Corpora Conference, Associação de Estudos e de Investigação Científica do ISLA, Lisboa, 2008, 247-250.

Nicolás Martínez, M. C. (ed.) 2008. Ricerche sul Corpus del parlato romanzo C-ORAL-ROM. Studi linguistici e applicazioni didattiche per l'insegnamento di L2, Firenze University Press, Firenze. http://epress.unifi.it/nicolas.htm

Nicolás Martínez, C. 2005. Una propuesta de utilización de corpus orales en el enseñanza de segundas lenguas. Lamusa digital n. 3 Congreso CLiP, Computers, Literature and Philology. http://www.uclm.es/lamusa/lista_articulos–mono.asp

Nicolás Martínez, C. 2004. Come affrontare lo studio di testi

provenienti da corpora. En M. C. Nicolás Martínez & S. Staton (eds) 2004, 243-252.

Nicolás Martínez M. C. & S. Staton (eds) 2004. Studi per l'insegnamento delle lingue Firenze University Press, Firenze, 243-252. http://epress.unifi.it/nicolas.htm.

Piñol M. C. 2017. Lingüística de corpus y enseñanza del español como 2/L. Arco libros: Madrid.

Pitkowski, E. F. & J. Vásquez 2009. El uso de los corpus lingüísticos como herramienta pedagógica para la enseñanza y aprendizaje de ELE. TINKUY 11: 31-51.

Quiroz Herrera, G. Á. 2007. Preparación y procesamiento de un corpus para la creación de materiales en clase de español para propósitos específicos. Actas del X Congreso Brasileño de profesores de español, Natal, 16 a 20 de setembro de 2003. Brasil: Embajada de España en Brasil - Consejería de educación, 131-151.

Tolchinsky, L. 2014. El uso de corpus lingüísticos como herramienta pedagógica. Textos de Didáctica de la Lengua y de la Literatura 65: 9-17.

Véliz Campos, M. 2008. La lingüística de Corpus y la Enseñanza del Inglés (como lengua extranjera): ¿Un matrimonio forzado?. Literatura y Lingüística 19: 251-263.

Yanira, P. 2013. Aplicación pedagógica de la lingüística de corpus a una clase de gramática avanzada de español como segunda lengua. marcoELE REVISTA DE DIDÁCTICA ELE, 16: 1-17.

Análisis de palabras clave en *La Fontana de Oro*, de Benito Pérez Galdós

Guadalupe Nieto Caballero

Departamento de Filología Hispánica y Lingüística General, Universidad de Extremadura, Spain

***Keywords**: Pérez Galdós; La Fontana de Oro; palabras clave; verbos de comunicación*

Abstract

En esta comunicación se presenta un análisis de corpus sobre los hábitos estilísticos que caracterizan *La Fontana de Oro*, primera novela de Benito Pérez Galdós. Aunque este tipo de enfoques no resultan desconocidos en los estudios literarios en general, la implementación de metodologías de corpus sí supone una novedad en el análisis de textos literarios en lengua española. En concreto, analizamos las palabras clave en la ópera prima de Galdós en comparación con el resto de novelas de su producción, con el fin de detectar hábitos del Galdós más joven. Como se podrá comprobar, uno de los elementos que más descuella en su primera novela con respecto al resto de su producción son los verbos de comunicación (*decir*, *contestar*, etc.). El uso que Galdós hace de estos verbos en su primera novela es fundamentalmente como parte de la estrategia de representación de discurso de estilo directo, o lo que es lo mismo, para dar voz directamente a los

personajes que pueblan su universo ficticio, como se puede advertir en el siguiente ejemplo:

—Señora —dijo Clara, poniéndose de rodillas junto al lecho y estrechándole las manos a la devota—. Señora, usted me defenderá; usted que es tan buena, que es una santa; usted que ya me defendió otra vez. (Pérez Galdós 1906 [1870], 263)

Como varios estudios han demostrado, el uso que Galdós hace de esta estrategia de representación del discurso está muy extendido en la primera mitad de su producción, para luego atenuarse paulatinamente a favor de otras estrategias por las que goza de gran reconocimiento desde un punto de vista literario, como el estilo indirecto libre o el monólogo interior (Gullón 1960; Andreu 1989). En el caso de su primera obra, sin embargo, el estilo directo comporta un hábito estilístico especialmente destacado con respecto al resto de su producción. Además de calibrar de manera precisa la presencia de estilo directo en *La Fontana de Oro* con respecto al resto de su producción narrativa, en esta comunicación analizaremos el uso que Galdós hace de esta estrategia. En lugar de centrarnos, como ha solido hacer la crítica especializada, en el habla de los personajes —algunos de los aspectos más celebrados del uso del estilo directo en Galdós se encuentran en el parlamento mismo de sus personajes, como las muletillas que sirven para caracterizarlos ("en toda la extensión de la palabra" de Doña Lupe en *Fortunata y Jacinta*, por ejemplo)—, el análisis girará en torno a cómo Galdós utiliza a la figura del narrador para introducir las palabras de los personajes a través de las proposiciones proyectoras. Como veremos, además de para darles voz, Galdós también se apoya en el narrador para crear efectos literarios concretos, como la definición del plano psicológico de los personajes, el efecto de simultaneidad entre el habla y el lenguaje gestual de estos o la sensación de tensión con pausas narradas retrospectivamente.

Además de analizar el estilo del Galdós más joven, esta comunicación pretende ilustrar el potencial de los enfoques de corpus en el análisis literario de obras en lengua española, y abrir así nuevas vías de análisis que hasta ahora no han sido exploradas en la exégesis de autores en este idioma.

References

Andreu, A. G. (1987). Modelos dialógicos en la narrativa de Benito Pérez Galdós.

Amsterdam: John Benjamins Publishing

Gullón, R. (1960). Galdós. Novelista moderno. Madrid: Taurus.

Pérez Galdós, B. (1906 [1870]). La Fontana de Oro. Madrid: Perlado, Páez y Compañía, Consultada a partir de la edición digitalizada por la Biblioteca Virtual Miguel de Cervantes en 2001. 15 de diciembre de 2018. <http://www.cervantesvirtual.com/obra/tormento--0/>

Equivocal cognates as semantic Anglicisms in present-day Spanish: a few case studies

José L. Oncins

University of Extremadura, Spain

Keywords: *Anglicisms; equivocal cognates; semantic loans*

Abstract

As far as the study of the influence of English on Spanish lexis is concerned, semantic Anglicisms have traditionally received far less scholarly attention than the other major category, *i.e.*, nonadapted loanwords (*e.g.*, *hacker, fair play* or *crack*). The reason for this imbalance is no doubt found in the very different degree of difficulty for identifying them: whereas the foreign outward appearance of non-adapted loanwords allows us to spot them easily (with the aid, for instance, of any unsophisticated spell checker), semantic loans tend to pass unnoticed under the guise of their Spanish cognates. For instance, the Spanish adjective *bizarro* –which is more often used today with the meaning of the English cognate from which it borrows this new sense (*i.e.*, *bizarre* = "very unusual or strange") will not arise any suspicion among Spaniards. Only linguists and translators with a good command of both languages will probably detect the interference. In fact, equivocal cognates like *bizarre* / *bizarro*, popularly known as "false friends", are dealt with in translation studies as translation mistakes, and glossed in dictionaries and glossaries normally aimed at translators and ESL Spanish/ ESL teachers (e.g., Prado).

Another remarkable difference between non-adapted and semantic Anglicisms in Spanish lies in the lexicographic treatment they receive in Spanish dictionaries. While nonadapted loans are duly labelled for their English origin in, for instance, the *DRAE* (the official Dictionary of the Spanish Royal Academy of the Language), semantic loans are never tagged to inform users of the English origin of the semantic extension undergone by the Spanish cognate. Consequently, the English component of a considerable amount of cognate host forms in Spanish goes, so to speak, unnoticed to linguists interested in studying the influence of English on the Spanish lexicon (Dwarkin).

This presentation has a two-fold purpose. In the first place, it proposes a method for spotting this type of "silent" loans, which draws on the abovementioned dictionaries of false friends as its main source of information, and uses the CORPES XXI (*) for tracking the presence of new meaning extensions in Spanish. In the second place, the presentation shows, retrospectively and prospectively, how equivocal cognates have been (still are, and will be) an important source of semantic Anglcisms in Spanish. Thus, retrospectively, it presents some equivocal cognates that became part of the Spanish lexicon a few years ago to form part of the stock of Spanish vocabulary; and, prospectively, it presents some recent examples of semantic Anglicisms that are making their way into Spanish today. This part of the presentation will describe and discuss four case studies of Spanish words that are currently undergoing a semantic extension due to the pushing influence of their corresponding English cognates. Even though these extensions are not registered in the DRAE yet, their presence in the CORPES XXI allows us to predict their inclusion in one of the forthcoming editions of the dictionary (updated every ten years or so).

References:

Dworkin, Steven N. 2012. A History of the Spanish Lexicon. A Linguistic Perspective. Oxford: Oxford University Press.

Prado, Marcial. 2001. Diccionario de falsos amigos inglés-español. Madrid: Gredos.

*CORPES XXI (Official corpus of the Spanish Royal Academy), available at: http://www.rae.es/

Spoken Corpus of KarelMakoň

Jan Oldřich Krůza

Charles University, Czechia.

***Keywords**: spoken corpora, speech transcription, topic identification*

Abstract

We present the spoken corpus of Karel Makoň, a corpus of spontaneous speech by a single speaker in Czech language within the domain of Christian mystic, with partly automatic, partly manual transcription obtained through a custom-tailored web application.

Makoň's corpus is a collection of talks given in a circle of friends in the course of late 60's or early 70's till 1991. The recordings have been kept on magnetophone tapes until their digitization between 2010 and 2012. A complete transcription was obtained using a dedicated ASR system and the work of the community around K.M.'s legacy. The corpus is about 1000 hours in total length, of which about 66 have been transcribed manually.

The author of the talks, Mr. Karel Makoň (Hájek 2007) *1912 †1993, was giving talks in a strife to share his awareness of the proverbial meaning of life and to give a manual to eternal life after his release from a concentration camp during the WWII.

The actual content of the corpus with respect to topics, references and statements, is on one hand quite well known because the the domain stays consistent across the whole set. On the other hand, it is only known vaguely and a systematic effort to analyze it is yet to be carried out.

In general, the whole corpus deals with a single topic: a howto for entering the eternal life before the physical death. On a finer-grained look, we can identify recurring sub-topics, like

- interpretation of the New Testament, notably the parable of the prodigal son, the parable of the talents, and Our Father,
- milestone personal experiences, like the one in the concentration camp,
- explaining symbolism, like the apostles as symbols for human abilities,
- references to specific people, like St. Teresa of Ávila or Padre Pio.

One of our aims is creating a list of recurring topics and their identification in the sound files. There are similarities to the work of Skorkovska (2011). We can take the current

transcription as a source for a baseline. Topic is inherently a vague concept but for the baseline evaluation, we can take a look at some easily defined cases like named entities. We can look for their specific word form and assume the transcription is covering it wherever we have search hits.

We have carried out an experiment where we have selected five example topics (4 of that named entities), performed a stemmed full-text search and manually verified the correctness of the search results. The precision was promisingly over 80%. We cannot directly measure recall but we can assume that it is much lower than the precision because of lower density of search hits in acoustically defect files as well as a higher density of hits in the manually transcribed part.

Another approach is to assume that the topics in the talks correspond to topics of the book the author was writing at the time. A search for the name of St. Teresa exhibits a peak in the year when K.M. was translating her Castillo Interior, which supports the assumption. I plan to investigate possible ways to raise the recall mainly by improving the ASR, as well as to enable manual topic annotation in the web interface.

References

[Hájek2007] Jurik Hájek. 2007. Český mystik Karel Makoň. Dingir, 2007/4:142–143.

[Hannun et al.2014] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

[Maekawa et al.2000] Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of japanese. In LREC. Citeseer.

[Skorkovská et al.2011] Lucie Skorkovská, Pavel Ircing, Aleš Pražák, and Jan Lehečka. 2011. Automatic topic identification for large scale language modeling data filtering. In International Conference on Text, Speech and Dialogue, pages 64–71. Springer.

A Collostructional Analysis of non-relational verbs

Aroa Orrequia-Barea

English Studies, University of Jaén, Spain

Keywords: *Semantics; corpus linguistics; relationality; direct speech, collostructional analysis*

Abstract

In English, there are some embedding verbs which behave differently from the rest semantically and syntactically. Unlike most embedding verbs in English, these verbs do not fulfil the two main requirements of the relational analysis, for their complement clause does not denote a proposition and the verb does not express a relation between the subject

and the complement clause of the verb (Moltmann 2003). This is the reason why they are called non-relational verbs (as opposed to those that establish a relation).

Regarding the first condition for a relational analysis, non-relational verbs cannot replace their complement clause with a complex nominal (e.g. the proposition that S) without a change or loss in meaning. This means that this kind of verbs does not have a

propositional denotation. This property correlates with several grammatical phenomena: parenthetical constructions, raising passive, 'so' anaphor, complementizer omission or adjunct extraction (Orrequia-Barea 2017).

However, there is a grammatical construction that can properly replace the complement clause of non-relational verbs: direct speech. Our hypothesis in this paper is that embedding verbs split into two groups differentiated by their ability to co-occur with socalled direct quotation. We will further claim that this division is due to the different reference to discourse conveyed by embedding verbs in one and the other group.j

To prove this hypothesis, we have queried members of the most representative taxonomies of embedding verbs (Hooper & Thompson 1973, Hooper 1975, Cattell 1978) in the British National Corpus (BNC henceforth) to check whether they can occur with direct speech complements. Apparently, the results seem to confirm our hypothesis: the verb that have a non-relational behaviour do in fact appear with direct speech in the corpus. Although BNC provides the frequency of occurrence, we are going to use a technique of statistical significance, the Collostructional Analysis (Gries & Steefanowitsh 2003) to measure the attraction between lexemes – complement verbs in this case – and grammatical patterns – direct speech in this particular case (Schmid & Küchenoff 2013). Thus, collostructional analysis will measure the strength between the direct speech complements and the complement verbs accounting for the semantic differences between two apparently synonymous constructions, complement verbs which can take that clauses as complements.

References

Cattell, R. (1978). On the source of interrogative adverbs. Language, 54, 61-77.

Gries, S. & Stefanowitsh, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. International Journal of Corpus Linguistics, 9, 95129.

Hooper, J. 1975. On assertive predicates. In Kimball, J. (Ed.), Syntax and Semantics, 4 (pp. 91-124), Nueva York: Academic Press.

Hooper, J. & Thompson, S. 1973. On the applicability of root transformations. Linguistic Inquiry, 4, 465-497.

Moltmann, F. 2003. Propositional attitudes without propositions. Synthese, 135(1), 77118.

Orrequia-Barea, A. 2017. Semántica de los verbos proposicionales del inglés actual. Tesis doctoral inédita. Cádiz: Unversidad de Cádiz.

Schmid, H.-J. & Küchenhoff, H. 2013. Collostructional analysis and other ways of measuring lexico-grammatical attraction. Theoretical premises, practical problems and cognitive underpinnings. Cognitive Linguistics, 24(3), 532-377.

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by

Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http//www.natcopr.ox.ac.uk/ (last access 05/12/2018).

El trasvase interlingüístico (español-inglés) de la estructura retórica de las fichas descriptivas de embutidos: Un estudio basado en corpus[3]

María Teresa Ortego Antón

CITTAC, Universidad de Valladolid, Spain

Keywords: *traducción, inglés, español, fichas descriptivas de embutidos, corpus, estructura retórica.*

Abstract

La internacionalización de la industria agroalimentaria en España va aparejada de un aumento exponencial de los servicios de redacción y de traducción del español al inglés. En el marco de esta necesidad pretendemos comprobar si las fichas descriptivas de embutidos en español y en inglés tienen la misma estructura retórica o si, por el contrario, existen patrones diferenciados en las mencionadas lenguas, puesto que los traductores necesitan conocer este aspecto para adaptarse a las necesidades de los consumidores anglófonos. Para ello, hemos diseñado, compilado y anotado retóricamente C-GEFEM, un corpus virtual comparable en español y en inglés compuesto por 100 fichas descriptivas de embutidos en cada lengua. Para el diseño y la compilación de C-GEFEM hemos aplicado el protocolo propuesto por Seghiri (2017) y Ortego Antón y Fernández Nistal (2019), que asegura el equilibrio, así como la representatividad cualitativa y cuantitativa, calculada esta última con ReCor (Corpas Pastor y Seghiri, 2007a, 2007b). Una vez compilado el corpus, hemos analizado los movimientos y pasos (Biber et al., 2007: 23-24) que componen la estructura retórica de este género lingüístico para detectar posibles patrones de comportamiento. Para ello, a partir del análisis de una muestra del corpus (10 textos en cada lengua) hemos establecido una serie de etiquetas retóricas, las hemos introducido en el Constructor de etiquetadores® desarrollado por el Grupo Interuniversitario ACTRES[4], hemos anotado C-GEFEM con el Etiquetador de movimientos retóricos®[5] y con el Visor de corpus comparables bilingües®[6] hemos establecido la información retórica típica en cada lengua siguiendo una metodología similar a la utilizada por Ortego Antón (en prensa). Los resultados que se desprenden de dicho análisis nos han permitido vislumbrar patrones de comportamiento típicos de este género textual en cada una de las lenguas. Por ejemplo, el movimiento «descripción del producto» ocupa diferente posición en la estructura en cada una de las lenguas y, a su vez,

[3] El presente trabajo ha sido realizado en el seno del proyecto "Producción textual bilingüe semiautomática inglés-español con lenguajes controlados: parametrización del conocimiento experto para su desarrollo en aplicaciones web 2.0. y 3.0." (FFI2016-75672-R).

[4] https://actres.unileon.es/wordpress/?lang=es (Fecha de consulta: 14/03/2019).

[5] http://contraste2.unileon.es/web/es/tagger.html (Fecha de consulta: 14/03/2019).

[6] http://contraste2.unileon.es/web/es/browser.html (Fecha de consulta: 14/03/2019).

en español se compone de seis pasos, a saber, «peso», «ingredientes», «alérgenos», «información nutricional», «conservación» y «utilización», en tanto que en lengua inglesa está formado por siete pasos: «marca», «conservación», «origen», «utilización», «envasado», «reciclaje» y «otra información». Por lo que respecta a las frecuencias de uso, estas son dispares, puesto que «conservación» se emplea con una frecuencia de 85 % en español y de 59 % en inglés. Por tanto, los traductores cuando trasvasan las fichas descriptivas de embutidos del español al inglés tendrán que ser conscientes de las diferencias existentes y adaptar la traducción a las características textuales de este género en lengua inglesa. Asimismo, con los resultados que se desprenden del análisis hemos establecido un prototipo de estructura retórica en español y en inglés para, en posteriores etapas, diseñar GEFEM, un generador automático de fichas descriptivas de embutidos basado en el procesamiento del lenguaje natural y destinado a satisfacer las necesidades de comunicación especializada del español al inglés y a asistir a los traductores y redactores del sector agroalimentario.

References

Biber, D., Connor, U. y T. A. Upton (eds.) (2007). Discourse on the Move. Using Corpus Analysis to Describe Discourse Structure. Amsterdam: John Benjamins. DOI:10.1075/scl.28

Corpas Pastor, G. & Seghiri, M. (2007a). Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness. Translation Journal, 11(3). URL: http://translationjournal.net/journal/41corpus.htm [18/06/18].

Corpas Pastor, G. & Seghiri, M. (2007b). Determinación del umbral de representatividad de un corpus mediante el algoritmo N- Cor. SEPLN: Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, 39, 165-172.

Ortego Antón, M. T. (En prensa). Las fichas descriptivas de embutidos en español y en inglés: un análisis contrastivo de la estructura retórica basado en corpus. Revista Signos, 52(101).

Ortego Antón, M. T. & Fernández Nistal, P. (2019). Estudio contrastivo de la terminología de embutidos en inglés y en español con ParaConc y tlCorpus a partir del corpus paralelo P-GEFEM y del comparable C-GEFEM. In M. Seghiri & G. Corpas Pastor (Eds.), El uso de los corpus lingüísticos como herramienta pedagógica para la enseñanza-aprendizaje de lenguas, traducción e interpretación. Berna: Peter Lang.

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. DOI: 10.1075/babel.63.1.04seg

Los corpus de aprendices y la enseñanza de lenguas: Aplicaciones y posibilidades del CAES “Corpus de Aprendices de Español” para la enseñanza del español

Ignacio M. Palacios Martínez

Departmento de Filología Inglesa y Alemana, Universidade de Santiago de Compostela, Spain

Keywords: *español segunda lengua; corpus de aprendices; enseñanza y aprendizaje*

Abstract

Una de las aportaciones de la Lingüística de Corpus es la de facilitar a los estudiosos del lenguaje el acceso a ejemplos reales de la lengua (Aijmer & Atenberg 1991; Baker 2009, McEnery & Hardie 2012). En esta línea, en los últimos años han proliferado un abundante número de corpus de distintos tipos: orales, escritos, multimodales, generales o especializados, centrados en una variedad estándar o en varias, etc. De entre estos destacan los corpus de aprendices que recogen muestras orales y/o escritas de estudiantes de una segunda lengua. Las posibilidades de estos corpus son inmensas, pues de ellos se puede extraer información de gran interés para la investigación en la adquisición y enseñanza de lenguas, para el diseño y confección de materiales didácticos, así como para la propia metodología de aula, o incluso para el proceso de evaluación (Granger, Hung & Petch-Tyson 2002, O'Keefe, McCarthy & Carter 2007, Aijmer 2009).

En este contexto surgió el CAES "Corpus de Aprendices de Español", financiado por el Instituto Cervantes y llevado a cabo en su primera fase por un grupo de investigadores de la Universidade de Santiago entre 2012 y 2015. Consiste en una colección de textos de alumnado de español de distintos niveles. En su versión actual contiene un total de 570.000 palabras pertenecientes a 1.423 estudiantes de 6 lenguas maternas diferentes. En una primera etapa se diseñó y creó una aplicación informática con el fin de que los propios participantes pudieran introducir directamente sus datos personales y realizar las tareas escritas propuestas. Previamente a esto, tenían que dar su consentimiento para el uso de estos datos con fines de investigación.

Los textos que conforman el CAES adoptan el formato XML. La cabecera muestra todos los datos necesarios para la identificación de los valores en cada una de las tareas realizadas, así como aquellos que corresponden a las diversas variables utilizadas para su clasificación; el texto escrito, por su parte, ocupa la parte central del documento en cada caso. De acuerdo con lo anterior, todos estos documentos se procesan y almacenan en una base de datos de la que es posible extraer ejemplos de una palabra o grupo de palabras, aplicando filtros, según uno o más parámetros seleccionados.

Todo el corpus está etiquetado de acuerdo con un total de 702 categorías y la herramienta de búsqueda desarrollada permite recuperar información estadística y ejemplos textuales de elementos, lemas, clase de palabras en función de los filtros introducidos. La primera línea de la recuperación de datos proporciona información estadística. Así, es posible, por ejemplo, obtener la frecuencia general de cualquier lema, elemento o subcategoría gramatical que, a su vez, puede corresponder a un número de parámetros (una L1 en particular, sexo, nivel, edad, etc.) o incluso a todos ellos al mismo tiempo. La segunda línea presenta los textos o ejemplos en concreto. Una vez descrito este corpus en detalle, se presentarán algunas posibilidades de sus posibles aplicaciones, por ejemplo, en el estudio de cuestiones gramaticales de especial dificultad para los estudiantes de español como lengua extranjera (diferencias entre *ser* y *estar*, preposiciones, uso del verbo *gustar*, etc.) y también otras de carácter léxico como los falsos amigos. La presentación concluye con algunas reflexiones y cuestiones para investigaciones futuras.

Referencias

Aijmer, K. (ed.). (2009). Corpora and language teaching. Amsterdam: John Benjamins.

Aijmer, K. & Atenberg, B. (eds.) (1991). English corpus linguistics. Studies in honour of Jan Svartvik. New York/London: Routledge.

Baker, P. (2009). Contemporary corpus linguistics. New York: Continuum International Publishing Company.

CAES "Corpus de aprendices de español": //http://galvan.usc.es/caes/search//

Granger, S., Hung J. & Petch-Tyson, S. (eds.) (2002). Computer learner corpora, Second language acquisition and foreign language teaching. Amsterdam: John Benjamins.

McEnery, T. & Hardie, A. (2002). Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press.

O'Keeffe, Anne, McCarthy M. & Carter, R. (2007). From corpus to classroom: Language use and language teaching. Cambridge: Cambridge University Press

Feminist activism and social media: The language of sexual assault and survival on Twitter

Patricia Palomino-Manjón

Programa de Doctorado en Lenguas, Literaturas y Culturas, y sus Aplicaciones, Universitat de València, Spain

***Keywords**: Twitter; computer-mediated communication; feminist activism; sexual violence*

Abstract

Twitter is a microblogging site which allows its members to share short messages with other Internet users (Zappavigna, 2017). The affordances of Twitter enable people to communicate through hashtags: keywords used by Twitter users to coordinate their tweets around specific areas of interest and communities (Zappavigna & Martin, 2018). These communities are enacted through ambient affiliation (Zappavigna, 2011), since their members create groups around a common concern but probably do not know each other and do not interact with the other members directly.

The use of hashtags is also related to the reaction of Twitter users to global events and sociopolitical movements (Giaxoglou, 2018). Hashtags help create visibility to topics of relevance in a specific moment. Social media have enabled ordinary citizens to share their opinions about social issues which were previously limited to mass media (Bou-Franch, 2013). Moreover, Twitter users feel free to share their opinions without them affecting their image in their real lives since Twitter gives its users the opportunity to anonymize their accounts (Hardaker & McGlashan, 2016).

In September 16th, 2018, a letter written by Dr. Christine Blasey Ford in which she accused the Supreme Court nominee Brett Kavanaugh of sexually assaulting her in high school was made public. Following its publication, she was called to a hearing in the

Senate. Days before the hearing, the President of the United States, Donald Trump, published a series of tweets giving his support to the nominee and questioning the victim's decision of not reporting sexual assault earlier. This situation sparked the response of feminist communities on Twitter, who created different hashtags in support of survivors of sexual assault. The hashtag *#WhyIDidntReport* gained popularity among Twitter users as a response to the president's allegations and mainly featured stories written by victims of sexual assault. However, it was also used by other Twitter users who did not support the ideas of feminists.

The aims of this study are a) to identify the different topics of conversation found in the hashtag, and b) to examine the different discourses and ideologies surrounding sexual assault and victims. This study goes in line with previous research on social media and gender which supports the idea of the need to create online communities to fight cyberviolence against women (Bou-Franch, 2016; Jane, 2016, 2017).

To this end, a corpus of tweets containing the hashtag *#WhyIDidntReport* was analyzed. The data were collected through the Google Sheets extension 'Twitter Archiver', which triggered a total of 92,884 tweets (3,192,523 words). The corpus includes tweets published in a span of 18 days: from the moment that Trump published his tweet (September 21st) to the day Kavanaugh was confirmed to US supreme court (October 8th). This study takes a corpus-assisted discourse studies approach (Baker et al., 2008), which combines corpus linguistics and critical discourse analysis methodologies. Corpus linguistics methods (frequency lists, collocations and concordances) were used to identify the most common discourses and linguistic patterns related to sexual assault. The results obtained were further scrutinized to examine the discourse and ideology of Twitter users in the hashtag. Preliminary results suggest that sexual assault and victims were mainly constructed around discourses of trauma and empowerment.

References

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T. & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & Society, 19(3), 273–306. doi: 10.1177/0957926508088962.

Bou-Franch, P. (2013). Domestic violence and public participation in the media: The case of citizen journalism. Gender and Language, 7(3), 275–302. doi: 10.1558/genl.v7i3.275.

Bou-Franch, P. (2016). 'Did he really rape these bitches?' Aggression, Women, Language. In: P. Bou-Franch (Ed.), Exploring Language Aggression against Women (pp. 1–14). Amsterdam: Benjamins.

Giaxoglou, K. (2018). #JeSuisCharlie? hashtags as narrative resources in contexts of ecstatic sharing. Discourse, Context & Media, 22, 13–20. doi: 10.1016/j.dcm.2017.07.006.

Hardaker, C., & McGlashan, M. (2016). "Real men don't hate women": Twitter rape threats and group identity. Journal of Pragmatics, 91, 80–93. doi: 10.1016/j.pragma.2015.11.005.

Jane, E. A. (2016). Online misogyny and feminist digilantism. Continuum, 30(3), 284–

297. doi: 10.1080/10304312.2016.1166560.

Jane, E. A. (2017). Feminist digilante responses to a slut-shaming on Facebook. Social Media + Society, 1–10. doi: 10.1177/2056305117705996.

Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on twitter. New Media and Society, 13(5), 788–806. doi: 10.1177/1461444810385097.

Zappavigna, M. (2017). Twitter. In C. Hoffmann, & W. Bublitz (Eds.), Pragmatics of social media (pp. 201–224) De Gruyter Mouton. doi: 10.1515/9783110431070008.

Zappavigna, M., & Martin, J. R. (2018). #Communing affiliation: Social tagging as a resource for aligning around values in social media. Discourse, Context & Media, 22, 4-12. doi: 10.1016/j.dcm.2017.08.001.

Image of the "Self" and "Other" in Political Speech Using Word Embeddings

Ioannis Papachimonas

Department of Linguistics, National and Kapodistrian University of Athens, Greece

***Keywords**: political corpora; word2vec; word embedding; Greek politics; image; computational linguistics*

Abstract

Image is really important in politics. The image of the "self" and of the "other" are constantly used by politicians and political parties to influence the people and swing voters on their side. Usually the "self" is positive and the "other" negative. The aim of this contribution is to use word embeddings (Word2Vec with Python) to detect semantic relations between words in large political corpora and extract images of the "self/us" and the "other/enemy".

Two Greek political parties were chosen for this research. The center-right conservative New Democracy (ND) and the left-wing SYRIZA party. They have both been in power during the Greek Debt Crisis and have both implemented austerity policies requested by the IMF and the EU. However, though implementing the same economic policies, the two parties present themselves as radically different to the electorate. They are trying to present to the public a positive image of themselves and at the same time present a negative image of their main electoral opponent.

References

Bird, S., Loper, E. & Klein, E. (2009). Natural Language Processing with Python. Sebastopol, CA: O'Reilly Media.

Boukala, S. (2014). Waiting for democracy: Political crisis and the discursive

(re)invention of the 'national enemy' in times of 'Grecovery'. Discourse & Society 25 (4), 483-499.

Chaidas, D. (2018). Legitimization Strategies in the Greek paradigm: A comparative analysis of Syriza and New Democracy. Language & Communication 60, 136149.

Goldberg, Y. (2015). A primer on neural network models for natural language processing. arXiv:1510.00726v1 [cs.CL] 2 Oct 2015.

Hall, S. (1997). The spectacle of the other. In S. Hall (ed) Representation: Cultural Representations and Signifying Practices. London: Sage, 223-290.

Lyons, J. (1977). Semantics. Cambridge: Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781v3 [cs.CL] 7 Sep 2013.

Reyes, A. (2011). Strategies of Legitimization in political discourse: From words to actions. Discourse and Society 22 (6), 781-807.

Vuorinen, M. (2012). Introduction: Enemy Images as Inversions of the Self. In M. Vuorinen (ed) Enemy Images in War Propaganda. Newcastle upon Tyne: Cambridge Scholars Publishing, 1-14.

Vowel devoicing affecting moraic duration in Japanese: A preliminary study based on the Corpus of Spontaneous Japanese

Giuseppe Pappalardo

Department of Asian and North African Studies, Ca' Foscari University of Venice, Italy

Keywords: *Japanese language; spoken corpora; moraic isochrony; vowel devoicing; durational compensation*

Abstract

Pike (1945) classified the world languages into two types of rhythmic/prosodic patterns: stress- timed and syllable-timed. According to this classification, stress-timed languages, like English and German, tend to have isochronous interstress intervals, while syllable-timed languages, like Italian and Spanish, tend to have equal syllable duration. Ladefoged (1975) added the mora-timed type, in which isochrony is maintained at the level of the mora, a sub-syllabic constituent that includes either onset and nucleus, or a coda. Japanese is often referred to as a mora-timed language (Otake 2015): the mora is without any doubt the psychological prosodic unit in spoken language, and it is the metric unit of traditional poetry (Bloch 1950). However, it is clear that moras are not strictly isochronous units (Beckman 1982). Thus, experimental studies have focused on detecting compensation effects that make average mora durations more nearly equal through the modulation of the inherent duration of the segments involved (Han 1962, Port et al. 1980, Homma 1981, Hoequist 1983, Warner and Arai 2001). Kawahara (2017) used the Corpus of Spontaneous Japanese (henceforth CSJ) in order to verify whether the durational compensation effect within a CV mora occurs in natural speech, in addition to read-speech in the lab. He observed a statistically significant compensation effect of CV moras, in which vowel duration tends to vary in response to the duration of the preceding consonant. However, as the same author pointed out, the compensation is not absolute

because there are several linguistic factors that potentially affect segments' duration profiles.

The aim of the current study is to expand the results of Kawahara (2017)'s study analysing the effect of linguistic factors, such as vowel devoicing and pitch accent, on moraic duration using the CSJ, a richly annotated corpus of spontaneous Japanese containing more than 650 hours of spoken language. The data analysed in the present study is the so-called Core, an extensively annotated portion of the corpus including about 45 hours of speech (more than 1,000,000 segmental intervals). The use of this large corpus is extremely efficient for this kind of research because it allows us to perform various types of analysis setting many search parameters simultaneously. The main empirical analysis will focus on vowel devoicing, a salient phenomenon of the Japanese language, which involves the complete disappearance of the sonority of close vowels (/i/ and /u/) when they occur between voiceless consonants or between a voiceless consonant and a pause (Fujimoto 2015). Although from a phonetic point of view the mora CV in which the vowel loses its sonority is a segment that comprises only a consonant, the status of the mora is maintained, since native speakers still "hear" the vowel (Beckman 1982). The results of the study suggest that the negative correlation between consonant and vowel within a CV mora occurs consistently with devoiced vowels: the consonant tends to be longer than its inherent duration when the adjacent vowel gets shorter due to the devoicing effect. However, no positive correlation has been observed between CV moras' duration within the same word, showing that the compensation effect occurs only at the level of the mora.

References

Beckman, M. (1982). "Segment Duration and the 'Mora' in Japanese". Phonetica 39, pp. 113-135.

CSJ (The Corpus of Spontaneous Japanese) (2004). National Institute for Japanese Language and Linguistics and National Institute of Information and Communications Technology. http://www.ninjal.ac.jp/english/products/csj/.

Fujimoto, M. (2015). "Vowel devoicing". In Kubozono, H. (Ed.), Handbook of Japanese Phonetics and Phonology. Berlin: De Gruyter Mouton, pp. 167-214.

Han, M. (1962). "The feature of duration in Japanese". Onsei no kenkyuu 10, pp. 65-80.

Honma, Y. (1981). "Durational relationship between Japanese stops and vowels". Journal of Phonetics 9, pp. 273-281.

Hoequist, C. Jr. (1983). "Durational correlates of linguistic rhythm categories". Phonetica 40, pp. 1931.

Kawahara, S. (2017). "Durational compensation within a CV mora in spontaneous Japanese: Evidence form the Corpus of Spontaneous Japanese". The Journal of the Acoustical Society of America, 142 (https://asa.scitation.org/doi/10.1121/1.4994674).

Ladefoged, P. (1975). A course in phonetics. Los Angeles: University of California.

Otake, T. (2015). "Mora and mora-timing". In Haruo K. (Ed.), The handbook of Japanese phonetics and phonology. Berlin: De Gruyter Mouton.

Pike, K. L. (1945). The intonation of American English. Ann Arbor, Michigan University Press.

Port, R. F., Al-Ani, S., Maeda, S. (1980). "Temporal Compensation and Universal Phonetics". Phonetica 37, pp. 235-252.

Warner, N., Aarai, T. (2000). "Japanese mora-timing: A review". Phonetica 58, pp. 1-25.

El acuerdo entre anotadores en la segmentación discursiva: una aproximación desde el modelo Val.Es.Co.

Elena Pascual Aliaga

Universitat de València, Grupo Val.Es.Co.

***Keywords**: segmentación; unidades discursivas; acuerdo entre anotadores; modelo Val.Es.Co.; conversación coloquial*

Abstract

El modelo de unidades del grupo Val.Es.Co. (Briz & Grupo Val.Es.Co. 2014, 2003) surge como respuesta a una problemática común en el análisis del discurso oral: el establecimiento de unidades de análisis que vayan más allá de la oración y que permitan acotar y analizar adecuadamente los fenómenos discursivos (Hidalgo & Pérez 2004, Narbona 1991, Pons 2014). El modelo Val.Es.Co. propone un total de 8 unidades (discurso, diálogo, intercambio, alternancia de turnos, turno, intervención, acto y subacto) que permiten segmentar una conversación coloquial. En la actualidad, se está diseñando un protocolo de anotación manual de corpus basado en el modelo

Val.Es.Co. (ver Pons 2016) con el fin de llevar a cabo la segmentación y anotación de unidades conversacionales en el *Corpus Val.Es.Co. 2.0.* (Cabedo & Pons, en línea). El presente estudio se enmarca en este proyecto y tiene como **objetivo** testar la aplicabilidad y reproducibilidad del protocolo de anotación de unidades discursivas, esto es, determinar 1) si el protocolo permite llevar a cabo la segmentación y anotación de unidades en una conversación sin residuo y 2) si puede aplicarse de forma que se obtengan los mismos resultados de anotación independientemente de la persona que lleve a cabo la anotación.

Para determinar el grado de replicabilidad del protocolo de anotación se ha empleado el **método** del acuerdo entre anotadores (Artstein 2017, Artstein & Poesio 2008), que consiste en comparar las anotaciones de una misma porción del corpus llevadas a cabo por dos o más anotadores que siguen un mismo protocolo. Puede calcularse cualitativamente o a partir de coeficientes estadísticos tales como la kappa de Cohen (1960), la kappa de Fleiss (1971) o el alfa de Krippendorff (2013). Según Artstein (2017), un coeficiente de acuerdo bajo o estadísticamente no significativo puede apuntar a la necesidad de revisar el protocolo de anotación y de modificar aquellos aspectos que generan el desacuerdo entre los anotadores, mientras que un coeficiente de acuerdo significativamente elevado puede indicar que el protocolo es replicable. Aun con todo, la bibliografía admite que las tareas complejas de anotación lingüística de corpus suelen

asumir coeficientes de acuerdo bajos, dada la presencia de categorías de análisis interpretativas (Van Enschot et al. en prensa).

Se ha seleccionado una conversación del *Corpus Val.Es.Co. 2.0* y tres anotadores han realizado de forma independiente la segmentación en unidades discursivas de la conversación. Se ha realizado una comparación cualitativa de los resultados de las tres anotaciones y se han aplicado dos pruebas estadísticas: la kappa de Fleiss, que ha permitido medir el acuerdo en aquellas unidades cuya delimitación coincide exactamente en las tres anotaciones, y el alfa de Krippendorff, que posibilita la medición del acuerdo en la delimitación de segmentos desiguales. Los **resultados** de la investigación apuntan, por un lado, a conclusiones positivas en cuanto a la aplicabilidad del protocolo, puesto que se ha llevado a cabo la segmentación íntegra de una conversación en unidades; y, por otro lado, muestran los aspectos menos replicables de la anotación en relación con los bajos niveles de acuerdo en el análisis de algunos fenómenos, como son las pausas llenas, las vacilaciones o los truncamientos.

References

Artstein, R. (2017). Inter-annotator Agreement. In N. Ide & J. Pustejovsky (Eds.), Handbook of Linguistic Annotation (pp. 297-314). Dordrecht: Springer.

Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4), 555-596.

Briz, A. & Grupo Val.Es.Co. (2014). Las unidades del discurso oral. La propuesta Val.Es.Co. de segmentación de la conversación (coloquial). Estudios de Lingüística del Español, 35(1), 11-71.

Briz, A. & Grupo Val.Es.Co. (2003). Un sistema de unidades para el estudio del lenguaje coloquial. Oralia, 6, 7-61.

Cabedo, A. & Pons, S. (Eds.): Corpus Val.Es.Co. 2.0 [online: www.valesco.es].

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), 378-382.

Hidalgo, A., & Pérez, M. (2004). De la sintaxis a la pragmasintaxis: problemas del análisis sintáctico en el discurso oral espontáneo. Cauce: Revista de filología y su didáctica, 27, 221-245.

Krippendorff, K. (2013). Content Analysis. An Introduction to Its Methodology. California: Sage Publications.

Narbona, A (1991). Sintaxis coloquial y análisis del discurso. Revista Española de Lingüística, 21(2), 189-190.

Pons, S. (2016). Cómo divider una conversación en actos y subactos. In A. M. Bañón Hernández, M.ª M. Espejo Muriel, B. Herrero Muñoz-Cobo & J. L. López Cruces (Eds.), Oralidad y análisis del discurso. Homenaje a Luis Cortés Rodríguez (pp. 545- 566). Almería: Editorial Universidad de Almería.

Pons, S. (2014). Discourse segmentation in Romance languages: an overview. In S. Pons (Ed.), Discourse Segmentation in Romance Languages (pp. 1-9).

Amsterdam/Philadelphia: John Benjamins.

Van Enschot, R. Spooren, W., Van den Bosch, A., Burgers, C., Degand, L., EversVermeul, J., Kunneman, F., Liebrecht, C., Linders, Y. & Maes, A. (en prensa). Taming our wild data: on intercoder reliability in discourse research. Dialogue & Discourse.

El discurso político y su aplicación docente a través del debate en el aula. Los marcos mentales en la competencia comunicativa oral en inglés.

Beatriz Pastor García

Departamento de Lingüística Aplicada, Universitat Politècnica de València

Keywords: *Lakoff; análisis de corpus*; *debate; marcos; comunicación*

Abstract

El objeto de este este trabajo es presentar las posibilidades didácticas del debate en el desarrollo de la competencia comunicativa en el aula, en particular en la enseñanza y aprendizaje del inglés como lengua extranjera para fines específicos, en un nivel B2 de acuerdo con los descriptores definidos dentro del Marco Europeo de Referencia de Lenguas (CEFR) y en el ámbito del Inglés para los Negocios.

El presente trabajo surge de una investigación previa que consiste en la aplicación de los supuestos de las tesis lingüístico-cognitivas de Lakoff (2002, 2006, 2007, 2008a, 2008b 2016) al ámbito de los discursos políticos. De acuerdo con este estudio, podemos concluir que los discursos políticos pueden eficazmente ser entendidos como construcciones textuales y representaciones del mundo, construidas a través de una serie de marcos; estos marcos son utilizados por los partidos políticos de cualquier signo como un recurso específico de la oratoria.

Lakoff (2002, 2006, 2007, 2008a, 2008b, 2016) considera que la construcción de la realidad llevada a cabo en la política tiene que afrontarse desde la selección de unas ideas claves que, por medio de un determinado lenguaje, conforman unos relatos o marcos, en lo que se aglutinan los supuestos del conjunto de la acción social y política de los ciudadanos. Siguiendo a Lakoff (2002, 2004, 2006, 2007, 2008a, 2008b, 2016), podemos distinguir el marco "Nación-Familia", dentro del cual, previamente, hemos diseñado unas tablas según el concepto de "Padre Estricto" y "Familia Protectora" con el lenguaje propio de cada una de estas visiones; estas nociones son las que proponemos aplicar a modelos específicos de dirección de empresa.

La aplicación docente de esta metodología se centra en el análisis del discurso empresarial de acuerdo con el método propuesto, permitiendo al alumnado analizar y comparar los distintos modelos para, posteriormente, crear un debate en el aula aplicando esos mismos marcos. Esta metodología no sólo sirve para la mejora de la competencia comunicativa sino, además, para el desarrollo de habilidades como el pensamiento crítico o el razonamiento e interpretación de la información.

La conclusión y resultado de la propuesta es que la actividad resulta altamente positiva no sólo en cuanto al uso y aplicación del inglés y el vocabulario específico, sino también en el grado de satisfacción mostrado por los estudiantes en los que se observa un interés por la actividad a lo largo del todo el proceso y un claro aumento de la fluidez comunicativa en lengua inglesa. Podemos concluir afirmando que uso del debate en la práctica docente ayuda a fomentar el respeto, las competencias interculturales y a reflexionar sobre del poder de la palabra.

References

Arroyo, J. L. B. (2015). "Pragmatics of Political Discourse". In The Encyclopedia of Applied Linguistics (pp. 1–7). https://doi.org/10.1002/9781405198431.WBEAL1464

Brown, Z. (2015). The use of in-class debates as a teaching strategy in increasing students' critical thinking and collaborative learning skills in higher education. In Educational futures. Vol. 7(1), pp.39-55.

Lakoff, G. (2002). Moral politics: how liberals and conservative think. Chicago: University of Chicago Press.

— (2006). Thinking points: communicating our American values and vision. Berkeley CA: California Press.

— (2007). No pienses en un elefante. Madrid: Editorial Complutense.

— (2008a). Puntos de reflexión. Manual del progresista. Barcelona: Península.

— (2008b). The political mind. New York: Viking.

Lakoff, G. & Wehling, E. (2016). Your brain's politics. How the science of mind explains the political divide. UK: Andrews UK Limited.

Temple, M. (1997). In-class debates: Fertile ground for active learning and the cultivation of critical thinking and oral communication skills. International Journal of Teaching and Learning in Higher Education. 19(2), pp.183-90.

El *lenguaje soez* del TO, *¿tabú* en el TM?

María Eugenia Pastor Ramos

Dpto. Filología Inglesa, Universidad de Valladolid, España

Keywords: *lenguaje soez; traducción audiovisual; doblaje; subtitulación; corpus*

Abstract

La traducción de lenguaje soez es un problema que surge en muchas traducciones y que no es siempre fácil de solventar. Al mismo tiempo, no siempre se le da el lugar que merece al estudio de este tipo de palabras y expresiones debido, muy probablemente, a la delicadeza que supone su trasfondo.

Este tipo de expresiones son cada vez más habituales en nuestro lenguaje cotidiano y, por tanto, se reflejan tanto en el cine, como en la literatura o en otros ámbitos como los propios medios de comunicación. Un gran número de obras actuales, aspirando a conseguir un lenguaje lo más realista posible, utilizan este tipo de palabras o expresiones como algo cada vez más cotidiano. Al ser cada vez más frecuentes, el problema para el traductor es cada vez más común e, incluso, cada vez más personajes son caracterizados no solo de una manera física o en su comportamiento, sino también por su manera de hablar, incluyendo este tipo de lenguaje como un rasgo característico del personaje en cuestión.

Es por ello que resulta cada vez más importante traspasar estos elementos lingüísticos del texto origen al texto meta conservando su esencia original. Este tipo de expresiones no siempre son traducidas de una manera correcta y numerosas veces se tiende a neutralizar el contexto original como veremos en algunos de los ejemplos obtenidos en nuestro corpus de estudio. *Son of a bitch*, *fuck you* o *bully* son algunas expresiones que no siempre encontramos en el texto meta a pesar de aparecer de una manera recurrente en el texto origen.

Esta investigación, basada en corpus, pretende identificar algunas de las principales estrategias utilizadas en el doblaje y la subtitulación en la traducción audiovisual (TAV) del lenguaje soez del inglés al español en diversas series americanas desde los años 90 hasta nuestros días. Se pretende así analizar y comparar ambas modalidades de TAV en la traducción de este tipo de lenguaje.

Por un lado, se ha realizado una clasificación del lenguaje tabú analizado en el corpus en cuatro categorías (sexual, escatológico, religioso, familiar e insultante) y, por otro lado, se ha procedido a catalogar las diferentes estrategias traductológicas utilizadas en su traducción, tanto en las versiones dobladas como subtituladas. De esta manera hemos comprobado si existe verdaderamente esa *censura encubierta,* tanto en una como en otra modalidad de TAV, debida, principalmente, a ese componente cultural que presentan determinadas traducciones en España.

Con el término *censura encubierta* aludimos a una posible realidad que se enmarca dentro de nuestra cultura, es decir, el lenguaje soez es común a las diferentes culturas pero difieren a la hora de tratarlo tanto cultural como lingüísticamente. Cuando hablamos de Traducción y Censura no existe un gran número de trabajos que se centren en analizar los *procedimientos puramente lingüísticos que se han llevado a cabo para que una traducción afectada por la censura presente los rasgos que la caracterizan como tal* (García Aguiar y García Jiménez, 2013: 136).

En este estudio analizamos y comparamos dichos procedimientos lingüísticos considerando las diferencias existentes en la traducción inglés-español en el contexto de la traducción audiovisual en varias series de televisión americanas.

References

García Agiar, L. C. & García Jiménez, R. (2013). Estrategias de atenuación del lenguaje soez: algunos procedimientos lingüísticos en el doblaje para Hispanoamérica de la película Death Proof. Estudios de Traducción, 3, 135-148. doi: http://dx.doi.org/10.5209/rev_ESTR.2013.v3.41995

Matillo, E. (2003). Difficulty of Slang Translation. Amsterdam: Rodopi.

Martí Ferriol, J. L. (2010). Cine independiente y traducción. Valencia: Tirant lo Blanch.

Tamayo Masero, A. (2012). La traducción del slang en Jackie Brown (Tarantino, 1997). Fòrum de Recerca, 17, 791-811.

A facework study of teasing among adolescents during telecollaboration

Barry Pennock Speck[a]; Begoña Clavel Arroitia[a]

[a]*Universitat de València*

Keywords: *Facework; Teasing; Telecollaboration; Language Learning; Secondary Education*

Abstract

In this study, we analyse the use of teasing (Keltner et al., 1998, 2001; Hay, 2000; Tragesser and Lippman, 2000) by adolescents in the context of a telecollaborative task carried out in the European project TILA (Telecollaboration for Intercultural Language Acquisition). We recorded all the exchanges that took place between English, Spanish and German secondary school students, yielding fourteen viable recordings, which make up our corpus. There were twelve exchanges between English and Spanish students and two between Spanish and German students. After our first analysis of these recorded sessions, we found that one of the exchanges between Spanish and German students stood out from the rest since it contained what we judged to be playful but competitive discourse between the interactants.

To ascertain whether our tentative analysis could be backed up by independent informants we asked twelve British university students who were proficient in Spanish to rate the fourteen exchanges for aggressiveness, competitiveness and humour (Pawluk, 1989, La Gaipa, 1977, Gorman and Jordan, 2015) employing a Likert-scale questionnaire. The video we had previously judged as playfully aggressive obtained the highest scores for the three parameters mentioned above. Thus we conducted a detailed qualitative analysis of the discourse in that particular exchange, paying special attention to the episodes of teasing from a facework perspective (Goffman, 1967; Craig, Tracy and Spisak, 1986; Penman, 1990, Locher and Watts, 2005, Culpeper, 2011, Haugh, 2017).

Our main research question was how and why the subjects in our study, who had never met previously, gradually started to employ teasing, which is, prima facie, a face-threatening strategy. We can state that this is due, in part, to the mixed genders of the participants –the Spanish contingent was made up of two boys while the German cohort consisted of two girls– and the nature of some remarks made during the exchange (Whitty, 2003; Lampert and Tripp, 2006). Similarly, we assume that the presence of an audience (Kowalski, 2004), that is, several male classmates in the Spanish school and two female peer onlookers in the German one, exerted some influence on the tone of the discourse. Teasing was found to be pervasive in the exchange under scrutiny –we found 12 episodes, two of which were quite lengthy.

Given the playful nature of teasing, students seem to indulge in it to make the task less tedious. Teasing is also employed to elevate the teasers' standing vis-à-vis their peers as successful ribbing can make the person who is doing it look both clever and fun, thus

bringing about face enhancement. Our study leads us to the conclusion that, in spite of the presence of apparently aggressive verbal behaviour, successful teasing can ultimately enhance not only the face of the teaser but also the butt of the tease if the latter plays along and thus teasing builds up rapport among all the participants –as is the case in our study.

References

Craig, Robert T., Tracy, Karen, Spisak, Frances (1986) The discourse of requests: Assessment of a politeness approach. Human Communication Research 12 (4), 437-468.

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

Goffman, Erving 1967. Interaction Ritual: Essays on Face to Face Behaviour., Garden City, New York.

Gorman, Glen and Jordan, Christian H. 2015. "I know you're kidding": Relationship closeness enhances positive perceptions of teasing. Personal Relationships, (22), 173–187.

Hay, Jennifer 2000. Functions of humor in the conversations of men and women. Journal of Pragmatics 32, 709-742.

Haugh, Michael 2017. Teasing. In: Attardo, S. (Ed.), Handbook of Language and Humour. Routledge, London, pp. 204-218.

Keltner, Dacher, Young, Randall C., Heerey, Erin E. Oemig, Carmen, Monarch, Natalia D. 1988. Teasing in hierarchical and intimate relations. Journal of Personality and Social Psychology 75 (5), 1231-1247.

Keltner, Dacher, Capps, Lisa, Kring, Ann M., Young, Randall C. and Heerey, Erin A. 2001. Just Teasing: A Conceptual Analysis and Empirical Review. Psychological Bulletin, 127 (2), 229-248.

Kowalski, Robin M. 2004. Proneness to, perceptions of, and responses to teasing: The influence of both intrapersonal and interpersonal factors. European Journal of Personality, (18), 331–349.

La Gaipa, John J. 1977. The Effects of Humour on the Flow of Social Conversation. In: Chapman, A. J., Foot, H. C. (Eds.), It's a Funny Thing, Humour. Pergamon Press, Oxford, pp. 421-427.

Lampert, Martin D. and Ervin-Tripp, Susan M. 2006. Risky laughter: Teasing and self-directed joking among male and female friends. Journal of Pragmatics, (38), 51–72.

Locher, Miriam A., Watts, Richard J., 2005. Politeness theory and relational work. Journal of Politeness Research (1/1), 9-33.Pawluk Cheryl J. 1989. Social construction of teasing. Journal for the Theory of Social Behaviour (19), 145167.

Pawluk Cheryl J. 1989. Social construction of teasing. Journal for the Theory of Social Behaviour (19), 145-167.

Penman, Robyn 1990. Facework and politeness: Multiple goals in courtroom discourse. Journal of Language and Social Psychology, (9/1), 15-38.

Kowalski, Robin M. (2004) Proneness to, Perceptions of, and Responses to Teasing: The Influence of Both Intrapersonal and Interpersonal Factors. European Journal of Personality, (18), 331–349.

Tragesser, Sarah L., Lippman, Louis G. 2005. Teasing: for superiority or solidarity? The Journal of General Psychology (132/3), 255-266.

Whitty, Monica T. 2003. Cyber-flirting: playing at love on the Internet. Theory & Psychology, (13/3), 339-357.

Análisis contrastivo (inglés-español) del vocabulario sensorial utilizado en la venta online de manzanas: la descripción del gusto y sus matices

Leonor Pérez Ruiz; Teresa Ortego Antón

[a]Filología Inglesa, University of Valladolid, Spain; [b]Lengua Española (Área de Traducción e Interpretación), University of Valladolid, Spain

***Keywords**: Análisis del corpus; descripción productos hortofrutícolas; traducción; terminología sensorial*

Abstract

Las fichas descriptivas de las manzanas en los sitios web de las empresas hortofrutícolas incluyen información detallada de estos productos, haciendo especial referencia a sus aspectos sensoriales, esto es, a la mezcla de sensaciones visuales, táctiles, olfativas y gustativas que contribuyen a valorar su calidad.

Existe una amplia producción bibliográfica dedicada al estudio y análisis de estas experiencias sensoriales polifacéticas, desde campos tan variados como la horticultura (Stone & Sidel, 2004; Daillant-Spinnler, MacFie, Beyts, & Hedderley, 1996), el marketing (Krishna, 2012), o incluso en áreas del ámbito de las humanidades como la psicología (Abdi, 2002), filosofía (Clark,1993) o la antropología (Classen, 1997). Sin embargo, a este tema se le ha prestado muy poca atención desde el terreno de la lingüística y, más concretamente, en los campos del análisis del discurso y la traducción.

En este trabajo hemos llevado a cabo una análisis comparativo inglés-español de las descripciones de las manzanas en estos sitios web, centrándonos en la forma en que se hace referencia a los aspectos gustativos, en el tipo y características del vocabulario que se utiliza y en cuáles son las colocaciones más frecuentes en ambos idiomas. Para ello hemos compilado un corpus comparable ad-hoc de descripciones de manzanas procedentes de páginas web de negocios de fruta fresca.

Ambos idiomas se caracterizan por contar con una amplia variedad y densidad terminológica, principalmente de adjetivos descriptivos y sustantivos específicos, para describir con gran precisión las distintas sensaciones que el comer una manzana provoca en el paladar. Nuestro corpus refleja que la descripción del gusto hace referencia a tres aspectos fundamentales: al nivel de dulzor, al perfume y a la evocación a distintas frutas, aunque en ocasiones las referencias en uno y otro idioma varían notablemente, debido en gran medida a que las descripciones de las sensaciones gustativas son claramente subjetivas. Por otro lado, puesto que aromas y sabores se perciben por igual durante la

ingesta de una manzana, a menudo los términos descriptivos referidos a estas sensaciones también se mezclan en ambos idiomas (*a mild, rich, high quince flavor and aroma; tiene un sabor muy refrescante, agridulce y aromático*). Además, en inglés hay una muy alta frecuencia de cadenas de adjetivos e intensificadores que premodifican al sustantivo (*a delicate, sweet, almost syrup like taste*) mientras que en español es mucho más habitual la postmodificación (*tendencia marcada a la acidez, y ligeramente azucarada y perfumada*).

Dado que muchas empresas españolas están interesadas en internacionalizar sus servicios, y puesto que el mercado online no tiene fronteras, este trabajo será de utilidad como ayuda tanto para la redacción y traducción de estas páginas como de otros documentos comerciales relativos al tema.

References

Abdi, H. (2002). What can cognitive psychology and sensory evaluation learn from each other? Food Quality and Preference, 13(7-8), 445-451.

Clark, A. (1993). Sensory qualities. Oxford: Oxford University Press.

Classen, C. (1997). Foundations for an Anthropology of the Senses. International Social Science Journal 49(3): 401–12.

Daillant-Spinnler, B., MacFie, H. J. H., Beyts, P. K. & Hedderley, D. (1996). Relationships between perceived sensory properties and major preference directions of 12 varieties of apples from the southern hemisphere. Food quality and preference, 7(2), 113-126.

Krishna, A. (2012). An integrative review of sensory marketing: Engaging the senses to affect perception, judgment and behavior. Journal of consumer psychology, 22(3), 332-351.

Stone H. and Sidel, J.L. (2004). Sensory Evaluation Practices. San Diego, CA: Elsevier Academic Press.

Accounting for the role of individual collocates on the choice between near-synonyms: A diachronic approach

Daniela Pettersson-Traba

Department of English and German, University of Santiago de Compostela, Spain

***Keywords**: near-synonymy; collocation; semantic vector spaces; collocational networks; diachrony*

Abstract

The importance of collocations, first emphasized by Firth (1957) and Sinclair (1966), has been one of the central tenets in the field of lexical semantics. Over the last 50 years, advancements in corpus linguistics have led to the emergence of rigorous investigations into the role of individual collocates on various linguistic phenomena. One such phenomenon is near-synonymy, i.e., words which despite denoting the same concepts display different collocational preferences. Although several studies have analyzed the

collocational preferences of particular sets of near-synonyms from a synchronic viewpoint (e.g., Gries, 2003; Liu, 2010; Liu & Espino, 2012) to quantify their semantic (dis)similarity, the diachronic development of near-synonyms has generally been disregarded. Against this backdrop, the aim of this paper is to analyze the collocational behavior of the near-synonymous adjectives *fragrant*, *perfumed*, and *scented*, which denote the concept SWEET-SMELLING, from 1810 to 2009 in American English.

To this purpose, instances of the near-synonyms and their L5-R5 collocates were extracted from *COHA* (Davies 2010–), and the time span covered by this corpus was divided into four 50-year periods: 1810-1859 (P1), 1860-1909 (P2), 1910-1959 (P3), and 1960-2009 (P4). The role of individual collocates on the choice between the adjectives was then assessed by employing two different methods. First, the data was analyzed by means of semantic vector space modeling, which quantifies semantic (dis)similarity by weighting the co-occurrence frequencies between target words, here the near-synonyms, and their collocates (Levshina, 2015, pp. 323-324). Second, a more fine-grained collocational analysis was conducted, following the approach proposed by Brezina et al. (2015) and later extended to diachronic analysis in Baker (2017, pp. 96-101). This technique consists in plotting the most significant collocates of lexical items in networks in which collocates and target words are connected by means of arrows. For a collocate to be included, it had to exhibit a minimum frequency of co-occurrence with the nearsynonyms of 5 and a Mutual Information score of 3 or higher. Four collocational networks were built, one per period, thus enabling us to determine which collocates the nearsynonyms share at different points in time and to identify potential variations in the relation between the near-synonyms.

Results indicate that *scented* and *perfumed* are semantically more similar to one another than they are to *fragrant*, since their collocational profiles are more alike, that is, they have more collocates in common. Concerning the historical evolution of the nearsynonyms, the data show a significant increase in the number of collocates of *scented* and a decrease of *fragrant* and *perfumed* over time. In some cases, this increase of *scented* takes place at the expense of the other two adjectives, with some collocates that were strongly associated with either *fragrant* or *perfumed* in P1 becoming more tightly connected to *scented* in P4. This points to the conclusion that *scented* is slowly but steadily encroaching upon the territory of *perfumed* and *fragrant*, as it seems to be taking over some of their functions. To conclude, the present study uncovers subtle distinctions in meaning between the three near-synonyms and offers valuable insights into how the concept SWEET-SMELLING develops over time as well as how these changes affect the relationship between the members of the synonym set.

References

Baker, P. (2017). American and British English: Divided by a common language? Cambridge: Cambridge University Press.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. International Journal of Corpus Linguistics, 20(2), 139-173.

COHA = Davies, M. (2010–). The Corpus of Historical American English (COHA): 400 million word, 1810-2009. Available at: https://corpus.byu.edu/coha/

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In F. R. Palmer (Ed.),

Selected papers of J. R. Firth 1952–1959 (pp. 168-205). London: Longman.

Gries, S. Th. (2003). Testing the sub-test: A collocational-overlap analysis of English -ic and -ical adjectives. International Journal of Corpus Linguistics, 8(1), 31-61.

Levshina, N. (2015). How to do linguistics with R: Data exploration and statistical analysis. Amsterdam & Philadelphia: John Benjamins.

Liu, D. (2010). Is it a chief, main, major, primary or principal concern? A corpus-based behavioral profile study of the near-synonyms. International Journal of Corpus Linguistics, 15(1), 56-87.

Liu, D., & Espino M. (2012). Actually, genuinely, really, and truly. A corpus-based behavioral profile study of the near-synonymous adverbs. International Journal of Corpus Linguistics. 17(2), 198-228.

Sinclair, J. (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M.A.K. Halliday, & R. H. Robins (Eds.), In memory of J.R Firth (pp. 410-430). Harlow: Longman

Using sem-grams to study metaphors in travel journalism

Antonio Pinna; David Brett

Dipartimento di Scienze Umanistiche e Sociali, Università degli Studi di Sassari

***Keywords**: travel journalism; semantic tagging; WMatrix; metaphors; sem-gram*

Abstract

Travel journalism is an important site for the study of the connections between the media and contemporary society. As a form of institutional social practice, it attracts and goes a long way to maintain a considerable readership worldwide, thus providing mediation between the providers of tourism services and its potential clients. Its main discursive functions are those of informing, guiding and entertaining, which can be studied from a linguistic point of view to highlight its more important textual strategies and their ideological load.

The current paper will introduce a corpus-driven technique that, to the best of our knowledge, has not been described in the literature so far. The technique in question is that of the "sem-gram", on analogy with the *n*-gram (Biber & Barbieri, 2007) and the PoS-gram (Brett & Pinna, 2015; Pinna & Brett, 2018). It bears particular similarity to the latter, in that it is essentially a concatenation of tags. But while the PoS-gram is a string of Part-of-Speech tags, the sem-gram is a string of semantic tags.

The usefulness of the sem-gram technique will be illustrated by way of the analysis of a 4M token corpus of travel journalism that was tagged semantically using the WMatrix tool (Rayson, 2008), which divides the lexicon into 21 macro semantic areas, each of which then has a varying number of sub-categories. Using tailormade perl scripts, concatenations of six tags were formed starting from each token in the corpus. The types of these concatenations were then quantified and subsequent analysis focussed on the

most frequent.

This procedure allowed the automatic identification of a number of phraseologies, which on further analysis turned out to be metaphors from the source domain of the human body. In the results and discussion section, particular attention will be paid to the semgrams II AT B1 IO AT W3 and II AT B1 IO AT M7c, tokens of which are *into the heart of the mountain* and *under the skin of the city*, respectively. The key elements of the semgrams are B1 (ANATOMY AND PHYSIOLOGY) and W3 (GEOGRAPHICAL TERMS)/M7c (PLACES) and the considerable lexical variation observed in both slots testifies to the great productivity of the conceptual metaphor place is human body in the language of travel journalism. For example in the II AT B1 IO AT W3 semgram the B1 slot, with 54 instances, was occupied by no fewer than 12 types, and apart from the predictable *heart*, *head* and *foot*, these included *bowels*, *cusp*, *brow* and *lip*. Comparison with the BNC showed that all of the phrases the *heart/head/foot/mouth of the* are present in the travel corpus at higher frequencies and use of the Chi-squared test demonstrated that these differences are statistically significant. Further comparison with a corpus of newspaper articles collected from different sections of the Guardian revealed that all of these phrases are overused in the Travel section, with the exception of *the head of the*, which was found to be statistically significant for the other sections, *head* being frequently used as synonym for director.

In conclusion, this as yet unexplored corpus-driven technique facilitates the retrieval of loose phraseologies that would otherwise go undetected, while at the same time producing more manageable results than the similar PoS-gram technique. With respect to travel journalism the phraseologies identified may be connected to textual strategies aiming at guiding potential tourists to envisage places in terms of familiar corporeal entities rather than via more accurate but less engaging descriptive language, thus possibly improving the texts' overall persuasive potential.

References

Biber, D. & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26(3):263-286

Brett, D. & Pinna, A. (2015). Patterns, Fixedness and Variability: Using PoS-grams to Find Phraseologies in the Language of Travel Journalism, in *Procedia - Social and Behavioral Sciences (198), Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015).*

Charteris-Black, J. (2004). *Corpus Approaches to Critical Metaphor Analysis*. Basingstoke: Palgrave.

Francis, G. (1993). A corpus-driven approach to grammar – principles, methods and examples. In *Text and Technology. In Honour of John Sinclair*, Mona Baker, Gill Francis & Elena Tognini- Bonelli (eds), 137–156. Amsterdam: John Benjamins.

Fürsich, E. & Kavoori, A. P. (2001). Mapping a critical framework for the study of travel journalism. *International Journal of Cultural Studies,* 4(2), 149-171.

Greaves, C. (2009). *ConcGram 1.0. A phraseological search engine,* Amsterdam: John Benjamins. Hanusch, F. (2013). *Lifestyle Journalism*. London: Routledge.

Hanusch, F. and Fürsich, E. (2014a). *Travel Journalism*. Basingstoke: Palgrave.

Hanusch, F. and Fürsich, E. (2014b). On the relevance of travel journalism. In Hanusch, F. and Fürsich, E. (eds.) *Travel Journalism. Exploring Production, Impact and Culture*. Houndmills: Palgrave Macmillan, 1-17.

Kövecses, Z. (2015). *Where Metaphors Come from*. Oxford: OUP.

Pinna, A. & Brett, D. (2018), Constance and variability: using PoS-grams to find phraseologies in the language of newspapers, in Kopaczyk, J. and Tyrkkö J. (eds.), *Applications of Pattern-driven Methods in Corpus Linguistics*, John Benjamins Publishing

Pirolli, B. (2019). *Travel Journalism. Informing Tourists in the Digital Age.* Abingdon, Routledge. Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519-549.

Stubbs, M. (2007). An example of frequent English phraseology: Distributions, structures and functions. In *Corpus Linguistics 25 Years on*, Roberta Facchinetti (ed.), 89–105. Amsterdam: Rodopi.

Contrasting personal reference in two languages: referring to 'oneself' in Spanish and Portuguese sociolinguistic interview corpora

Pekka Posio

Department of Languages, University of Helsinki, Finland

***Keywords**: pronouns; reference; intersubjectivity; Spanish; Portuguese*

Abstract

Deictic personal pronouns can be considered as relevant items for contrastive studies since their primary meanings are cross-linguistically same or, at least, very similar. However, the ways in which speakers use these pronouns and what referential strategies are chosen to carry out similar tasks, such as referring to 'oneself' either as an individual or a member of a group, may differ considerably even between two closely related languages. Thus, in a contrastive study, one can ask either (1) how a particular personal pronoun is used in two different languages or (2) what personal pronouns or other referential strategies are used in two languages for the same function. In my presentation, I approach the latter question by examining two closely related languages, Peninsular Spanish (PS) and European Portuguese (EP). Using 'reference to oneself, either as an individual or as a member of a group' as a comparative concept (Haspelmath 2010), I examine (1) what pronouns, verb forms and other referential strategies speakers use for this function in sociolinguistic interview corpora and (2) what pragmatic effects the use of different referential devices have in the discourse.

The data, totaling ca. 160,000 words and 30 informants, consist of sociolinguistic interviews of informants from Salamanca, Spain (Fernández 2005) and Porto, Portugal (Author, in progress), being interviewed mainly on topics related to their studies and work. Thus, the data can be considered comparable, given that they have been collected using the same methodology and virtually same interview questions, prompting answers in the first-person singular but allowing for other person forms such as the first-person

plural, the generic second-person singular and impersonal pronouns such as Spanish *uno* 'one' and grammaticalizing referential devices like Portuguese *a pessoa* 'the person' (Author 2017) to be used.

Although the default form used by the informants in the interviews is the first-person singular, they also regularly employ other referring expressions. The use of different person forms and other referential devices in the answers creates different types of intersubjectivities in the discourse (Nuyts 2006), presenting the speakers' opinions and experiences as pertaining only to themselves or shared by others, possibly including the addressee. Speakers' choice among the referential strategies varies depending on the language, but also correlates with their age, gender, and relationship with the interviewer. For example, in the Spanish corpus there is a positive correlation between the use of collective reference forms and the age of the informants, while in the Portuguese corpus all informants use more collective reference forms than in the Spanish corpus, irrespectively of their age (Author 2016, 2017).

To the extent that the differences revealed between the two languages are systematic, they may be attributed to a change in progress in the use of person forms, or to sociocultural differences such the predominance of positive or negative politeness in the speaker community (Carreira 2005, Hickey 2005). In my talk, I discuss the results in the light of different explanatory hypotheses.

References:

Author. 2016. You and we: Impersonal second person singular and other referential devices in Spanish sociolinguistic interviews. Journal of Pragmatics 99, 1-16.

Author. 2017. Entre lo impersonal y lo individual. Estrategias de impersonalización individualizadoras en el español y portugués europeos. Spanish in Context, 14:2, 209–229.

Author. In progress. Português falado no Porto. Sociolinguistic interview corpus of European Portuguese.

Carreira, Maria Helena Araújo. 2005. "Politeness in Portugal: How to Address Others?" In Hickey, Leo & Stewart, Miranda (eds.), Politeness in Europe, p. 306-316. Clevedon/Buffalo/Toronto: Multilingual matters.

Fernández Juncal, Carmen. 2005. Corpus de habla culta de Salamanca. Burgos, Instituto Castellano y Leonés de la Lengua.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. Language 86(3), 663-687.

Hickey, Leo. 2005. "Politeness in Spain: Thanks but No 'Thanks'. In Hickey, Leo & Stewart, Miranda (eds.), Politeness in Europe, p. 317-333. Clevedon/Buffalo/Toronto: Multilingual matters.

Nuyts, Jan. 2006. Modality: Overview and linguistic issues. In: W. Frawley (ed.), The expression of modality, 1-26. Berlin: Mouton De Gruyter.

Obligatory inversion in scientific texts

Carlos Prado-Alonso

Departamento de Filología Inglesa, Francesa y Alemana, University of Oviedo, Spain

Keywords: *Word-order; Linguistic variation; Obligatory inversion; Scientific texts*

Abstract

Most studies on English inverted constructions base their classifications on the preliminary distinction between two main types of inversion: full-verb inversion, as in "Here is the bus", and subject-operator inversion, as in "By no means should he approach the suspects". These two types of inversion have been further classified into *non-obligatory* and *obligatory* inversion. Non-obligatory full or subject-operator inversions are constructions in which the addressor can opt for either the inverted wordorder or its canonical counterpart, as shown in (1)-(2). By contrast, obligatory full or subject-operator inversions are obligatorily triggered by certain fixed preverbal constituents, and are not replaceable in context by a comparable clause with SVX canonical word-order because this is grammatically unavailable or conveys a different meaning, as in (3)-(4).

1) a. Equally important **is** *the fact that households may be composed of unrelated individuals or contain an extended family of two or more generations*.

b. The fact that households may be composed of unrelated individuals or contain an extended family of two or more generations is equally important.

2) a. He has passed more exams than **have** *his* classmates (**passed**).

b. He has passed more exams than his classmates have.

3) a. Here **is** *the nose*, here the chin.

b. The nose is here, the chin here.

4) a. Undoubtedly the well-to-do of Athens; in imperial Rome as in early modern times the rich supplied funds by the back door to large-scale traders. Nor **did** *aristocrats* totally **surrender** the field.

b. Undoubtedly the well-to-do of Athens; in imperial Rome as in early modern times the rich supplied funds by the back door to large-scale traders. And the aristocrats did not totally surrender.

The analysis of obligatory or formulaic full and subject-operator inverted structures has been neglected in the literature since it has been argued that there seems to be no linguistic variation involved in the use of these types of syntactic constructions. Here, however, I defend the view that obligatory full and subject-operator inverted structures are productive, highly structured constructions, which are worthy of serious linguistic investigation. In my presentation, I will provide an in-depth corpus-based analysis of the factors that drive the distribution and pragmatic use of obligatory full and subject-operator inversions in scientific texts.

The data are taken from four computerized corpora of British and American Present-day English, namely the FLOB, FROWN, BE06, and AmE06 corpora, comprising scientific texts from the 1960s, 1990s and 2000s.

On the basis of the corpus-based results, the different obligatory inversions instances in the scientific texts analysed in the investigation are grouped into five main types:

1) obligatory inverted structures triggered by deictic adverbs,

2) obligatory inverted structures triggered by enumerative listing conjuncts,

3) obligatory inverted structures triggered by a pro-form,

4) obligatory inverted structures triggered by additive adverbs,

5) and obligatory inverted structures triggered by negative or restrictive adverb or prepositional phrases.

The analysis will show that despite their formal and functional dissimilarities, these different obligatory inversions types still relate to one another in systematic and predictable ways and I will show that in fact they are grouped in relation to a unit which is naturally most salient —the prototype— and form with it a family of nodes —extensions from the prototype— in the system.

References

Baker, Paul. 2009. Contemporary Corpus Linguistics. London: Continuum.

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

Chen, Rong. 2003. English inversion: A Ground-before-figure Construction. Berlin: Mouton de Gruyter.

Langacker, Ronald. 1987. Foundations of Cognitive Grammar. Vol. 1: Theoretical prerequi¬sites. Stanford: Stanford University Press.

Langacker, Ronald. 1991. Foundations of Cognitive Grammar. Vol. 2: Descriptive application. Stanford: Stanford University Press.

Patten, Amanda L. 2017. "Well-formed lists: Specificational copular sentences as predicative inversion constructions". English Language and Linguistics. Cambridge: Cambridge University Press.

Prado-Alonso, Carlos. 2016. "A constructional analysis of obligatory XVS structures". *Studia Anglica Posnaniensia* 51(1), 51-82.

Webelhuth, Gert. 2011. "Motivating non-canonicality in construction grammar: the case of locative inversion". Cognitive Linguistic, 22(1): 81-105.

Explorations of television language from a corpus-assisted discourse studies perspective

Joaquín Primo-Pacheco

Doctoral Programme in Language, Literature, Culture, and its Applications, Universitat de València

***Keywords**: television; drama; series; CADS*

Abstract

With the rise of instant access to media content on over-the-top streaming providers such as Netflix, HBO, Hulu or Amazon Prime Video, it is an undeniable fact that the world is witnessing yet another golden age of television. In turn, this overwhelming and widespread availability of thousands of media products just a click away necessarily requires, more so than ever, a profoundly critical reflection regarding the many founding elements that constitute them, one of which, of course, is language.

As an academic discipline, the study of television language is definitely undergoing a parallel golden age too, as demonstrated by recent book-length studies (Bednarek 2018) and many other works (Quaglio 2009; Bednarek 2010, 2015, 2017; Richardson 2010; Queen 2015; Beers Fägersten 2016; Gregori-Signes 2017) which have been published during the last decade.

In keeping with the above, the purpose of this paper is to present the latest advancements of a work currently in progress which takes a CADS (Corpus-Assisted Discourse Studies) approach (Baker et al. 2008; Partington, Duguid and Taylor 2013) to a corpus containing transcripts of a selection of North-American television drama series. Thus, it shall build on and further tentative results from my previous research on the linguistic representation of sexual identities in television fiction, namely the variation in use of sexual gay slang and sexual identity labels in up to three North-American gay-themed television series: *Will & Grace* (NBC, 1998-2006, 2017-), *Queer as Folk* (Showtime, 2000-2005) and *Looking* (HBO, 2014-2015).

In so doing, this paper will also address universal matters pertaining to corpus-linguistic research and more specific questions concerning the study of television language, such as data sources, corpus design and compilation, transcription and annotation issues, as well as the need and feasibility of building a reference corpus, in hopes of contributing to the fulfillment of Bednarek's (2018: 5) recent affirmation that "in the midst of a 'golden' age of television, we need a comprehensive investigation of language use in televisual narratives."

References

Baker, P., Gabrelatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T. & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse and Society, 19(3), 273-306.

Bednarek, M. (2010). The language of fictional television: Drama and identity. London/New York: Continuum.

Bednarek, M. (2015). Corpus-Assisted Multimodal Discourse Analysis of Television and Film Narratives. In Baker, P. & McEnery, T. (Eds.), Corpora and discourse studies: Integrating discourse and corpora (pp. 63-87). Houndmills: Palgrave Macmillan.

Bednarek, M. (2017). The role of dialogue in fiction. In Locher, M. A. & Jucker, A. H. (Eds.), Pragmatics of fiction (pp. 129-158). Berlin: Walter de Gruyter.

Bednarek, M. (2018). Language and television series: A linguistic approach to TV dialogue. Cambridge: Cambridge University Press.

Beers Fägersten, K. (Ed.) (2016). Watching TV with a linguist. Syracuse: Syracuse University Press.

Gregori-Signes, C. (2017). 'Apparently, women don't know how to operate doors': A corpus-based analysis of women stereotypes in the TV series 3rd Rock from the Sun. International Journal of English Studies, 17(2): 21-43.

Partington, A., Duguid, A. & Taylor, C. (2013). Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS). Amsterdam: John Benjamins.

Quaglio, P. (2009). Television dialogue: The sitcom 'Friends' vs. natural conversation. Amsterdam/Philadelphia: John Benjamins.

Queen, R. (2015). Vox popular: The surprising life of language in the media. Chichester, UK: Wiley-Blackwell.

Richardson, K. (2010). Television dramatic dialogue: A sociolinguistic study. Oxford:

Oxford University Press.

First steps towards a new corpus of pseudo-scientific language

Luis Puente-Castelo

Facultade de Filoloxía, Universidade da Coruña, Spain

***Keywords**: corpus design; pseudo-science; scientific register; pseudoscientific language*

Abstract

From the anti-vaccines movement to the rejection of telecommunications antennae, from climate change denialism to "holistic" pseudo-philosophical approaches, and from homeopathy to paid-for articles trying to provide excuses for corporate decisions; the use of non-scientific discourses trying to pose as real science has become a major problematic issue in our society.

These practices have been commonly referred to as "pseudo-science", a term which, according to the *Oxford English Dictionary*, means "a spurious or pretended science;" which is either "a branch of knowledge or a system of beliefs mistakenly regarded as based on scientific method or having the status of scientific truth" or a "study or research that is claimed as scientific but is not generally accepted as such."

This spurious science is characterised by the fact that it tries to pose as real science, and the use of scientific-sounding language plays a crucial role in these attempts. The preoccupation with the use of pseudo-scientific language is not new, as in 1871, the British physicist George Maxwell stated: "Such indeed is the respect paid to science, that the most absurd opinions may become current, provided they are expressed in language, the sound of which recalls some well-known scientific phrase." This preoccupation has continued until our days, and the denounces of pseudo-science commonly emphasize its intention to deceive through the use of a language that purports to be scientific. In the light of this, it seems important to be able to characterize pseudo-scientific discourse from a linguistic point of view, and to distinguish it from genuine scientific discourse.

The aim of this paper is to present the first steps towards the design and future compilation of a new corpus of pseudo-scientific language, focusing on the main difficulties of the process of design of such a corpus. In particular, it will try to address the difficulties in the demarcation of science from non-science (Fuller 1985, Mahner 2007) and how these influence the scope of what can be considered "pseudo-science", and, consequently, the design of the corpus.

Moreover, it will discuss other, still unresolved, methodological issues, such as the representativeness and balance of the corpus, its subdivision in multiple subcorpora and the possible necessity of control corpora to distinguish it not only from scientific language but also from other quasi-scientific registers such as popular science or news about science. The ultimate aim is to compile a corpus that may help ascertain whether pseudoscientific language can be considered a unitary register.

References

Fuller, Steve (1985). "The demarcation of science: a problem whose demise has been greatly exaggerated" Pacific Philosophical Quarterly, 66: 329–341.

Mahner, Martin (2007). "Demarcating Science from Non-Science" in Theo Kuipers (Ed.) Handbook of the Philosophy of Science: General Philosophy of Science – Focal Issues (pp.515-575). Amsterdam: Elsevier.

Oxford English Dictionary. Accessed at www.oed.com on 05/12/2018.

Strategies in tracing linguistic variations in the Corpus of Old Irish texts (CorPH)

Fangzhe Qiu

Department of Early Irish, Maynooth University, Ireland

***Keywords**: linguistic variation; workflow; tagging strategies; ChronHib; CorPH*

Abstract

This paper describes the workflow of the *Chronologicon Hibernicum* (*ChronHib*) project, especially on the part of linguistic variation tagging. *ChronHib* studies the development of the Old Irish language from c. 600-950 CE, including changes in phonology, morphology, orthography and syntax. A corpus (*Corpus Palaeohibernicum*, or CorPH) has been built which comprises more than 110,000 tokens from 70 Old Irish texts, and this corpus has been extensively annotated with POS, morphological, lexical and syntactical information. The corpus not only allows more precise descriptions of variations in Old Irish that have been known to scholars, but also identification of yet undiscovered variations. However, how to convert this rich corpus of data into observation of linguistic variation is a challenge, since, unlike other annotations which can be operated independently on single tokens or sentences, variation by definition involves comparison between data that share similar properties but differ in other regards.

Therefore, information about linguistic variations belong to a higher rank in the data structure. Depending on the nature of the query, different strategies are designed to retrieve information about linguistic variations. If the query is concerned with a limited number of variants of token surface forms or of criteria that are already tagged, we can directly search for all relevant tokens using the built-in function of the database, and then evaluate them, e.g. 'find all "U-STEM" "NOUN" in "GEN.SG." that end in "*o" or "*a"'. This strategy is most frequent in current studies of linguistic variation. It applies to cases when we know the existence and scope of a variation and all its possible variants, or when we wish to probe if variation exists under given conditions. However, if the query criteria are too difficult to define, or if untagged properties are involved, the first strategy would be invalid. For instance, it is impossible to write a query for 'hyper-corrected "U" in an unstressed syllable', since, firstly, "unstressed syllable" is not a tagged entity, and secondly, algorithm to predict what vowel should appear in an unstressed syllable is hardly available, although human can make such a judgment based on etymology, grammar and sound-laws.

To tackle such difficulties, we design two further strategies. The first is an 'ad-hoc' approach, viz. to start from possible queries to limit the corpus to relevant data, tag the results with extra entities needed for the final query, and repeat the round for several times if necessary. This works when there is still some ways to limit the pool to a reasonable size for further, selective tagging. The other is a 'bottom-up' approach. We produce a list of variations that are known or can be expected to have happened in Old Irish, and give each variation an ID. Then we go to the dataset under investigation and tag the tokens directly with the entity of 'variation', whenever a variation can possibly happen to that token. The tag consists of the ID of variation and one of these values: "YES", "NO" and "MAYBE". This paper will discuss in detail the design, workflow, standard of evaluation and advantages of this last strategy.

Collocations in English and Spanish Promotional Texts Describing Cheese

Noelia Ramon[a]; Belen Labrador[a]

[a]*Departamento de Filología Moderna, Universidad de León, Spain*

Keywords: *collocation; cheese descriptions; corpus; phraseology*

Abstract

The concept of *collocation*, coined by Firth (1957), is based on the notion that each word in a particular language tends to co-occur with certain words more often than it does with others. This paper is a corpus-based contrastive study of the most common collocations found in English and Spanish online promotional texts describing different types of cheese. The main objective is to provide a detailed account of the main collocational patterns of the 10 most frequent nouns in this particular text type, to determine to what extent these patterns contribute to the persuasive function of the cheese descriptions.

Together with wine, cheese is a type of food industry which is becoming increasingly internationalized, with a wide variety of cheeses from many different countries easily made available worldwide. In Spain the dairy industry has developed greatly in the recent past and a large number of our cheeses have won international awards and are now being

exported all over the world. Because of the great variety of cheeses, the type of language used to describe them is also varied, focusing, in particular, on a wide range of adjectives available for the characterization of the taste or smell of a particular cheese. Collocational patterns involving nouns provide essential positive information about the cheeses which clearly enhance the persuasive function of this text type.

For the purpose of this paper, an *ad hoc* English-Spanish comparable corpus of online cheese descriptions was compiled. The English corpus contains 109,152 words and the Spanish corpus 136,010 words. The corpus was compiled and exploited using Sketch Engine. This tool also provides a POS tagging in the two languages, so it was possible to obtain information relating to grammatical strings. Sketch Engine was used to extract wordlists in both corpora so as to determine the most common nouns employed in both languages. The 10 most common nouns in English are: *cheese, milk, flavour, cheddar, texture, cow, rind, England, vegetarian,* and *blue,* and the 10 most common nouns in the Spanish corpus are: *queso, leche, sabor, corteza, pasta, cabra, vino, maduración, oveja* and *tipo.* These nouns were then isolated and a word sketch was produced for each one in order to obtain a clear picture of the collocational patterns they are involved in. Finally, the two separate descriptions were contrasted to identify similarities and differences in the trends found in English and Spanish to produce persuasive idiomatic strings in cheese descriptions. This enables us to determine if recurrent lexical collocations may be the result of persuasive strategies trying to convince potential customers to buy a particular cheese.

The results of this cross-linguistic phraseological study are expected to provide deeper understanding of the persuasive language used in marketing, in the particular subgenre of cheese descriptions. The potential areas of application include ESP teaching, translation training as well as the direct use of these inventories by non-native writers of this particular text-type, i.e. Spanish cheese sellers who aim at expanding their sales scope to include international markets.

References

Firth, J.R. 1957. Papers in Linguistics 1934–1951. London: Oxford University Press.

Anti-gender movements in contemporary Brazil: an analysis of the speeches given in the Chamber of Deputies, 2000-2017

Daniela Rezende[a]; Luciana Beatriz Ávila[b]

[a]Departamento de Ciências Sociais, Universidade Federal de Viçosa, Brazil; [b]Instituto de Humanidades, Artes e Ciências (IHAC/CSC), Universidade Federal do Sul da Bahia, Brazil

Keywords*: gender ideology; Chamber of Deputies; Brazil*

Abstract

This paper analyses anti-gender movements in Brazil, focusing on its strategies through the legislative branch and political discourse performed by elected representatives in the Brazilian lower chamber. Thus, we identified actors, arguments and time frames of the debates around *"gender ideology"*. According to Kuhar and Paternotte (2017, p. 5), this is "a term initially created to oppose women's and LGBT rights activism as well as the scholarship deconstructing essentialist and naturalistic assumptions about gender and sexuality". In this sense, one of the central elements of these movements is the defense of the "natural family", that is, of the heterosexual union with the objective of guaranteeing the reproduction of the species, and of "human ecology", i.e., biological sex, binary and correspondence between gender, gender and gender identity, as created by God (Kuhar & Paternotte 2017, 8). Methodologically, using GoogleTrends, we verified that, in Brazil, the insurgency of anti-gender demonstrations in the lower chamber took place during the debates about the education plans in 2014, having gained strength also when associated with the project School without Party (Escola sem Partido, PL 867/2015). We surveyed the speeches of the deputies in the Brazilian lower chamber, from 2000 to 2017, related to the theme. We searched for the keywords "gender ideology", "decriminalization of abortion" and "homosexuality". In total, 337 discourses were obtained and analyzed quantitatively and qualitatively. From the descriptive statistics, we found 145 discourses through the keyword "Homosexuality", 68 using the keyword "Decriminalization of abortion" and 57 discourses searching for "Gender ideology". The most frequent years of speeches are 2011 (70), followed by 2015 (54), 2016 (47) and 2017 (39). To this work, we selected and analyzed 57 discourses referring only to "gender ideology". Using software IRAMUTEQ (Ratinaud 2009), we generated 4 classes of discourses, divided into 3 subcorpora (A, B e C). Each of these corresponds to one of the central mobilized arguments: 1. rejection of gender ideology, associated with debates over education plans; 2. defense of biological sex, associated with the strengthening of the "natural" and religious character of the same, in addition to its association with human reproduction; 3. family preservation and children's rights, identified with a traditional (or "natural") family defense and the preservation of children, who would be threatened in different family arrangements. The argumentative strategies against "gender ideology" pervaded the criminalization of gender in education plans because of the supposed threat it might bring to the biologically given definitions of feminine and masculine. Moreover, the discourses are based on arguments of a nationalist, religious and pseudo-scientific nature, given the articulation between "gender ideology", that represented a "coup against the nation", based on its ties to the United Nations and international treaties, such as the Beijing Platform for Action, and "the destruction of the [heterosexual] family", characterizing a moral panic strategy present in movements anti-gender in other countries, such as Hungary, Poland, France, Germany, Mexico, Colombia, just to name a few examples.

References

Kuhar, R. and Paternotte, D. (eds). (2017). Anti-gender campaigns in Europe: Mobilizing against equality. London: Rowman & Littlefield International.

Ratinaud, P. (2009). IRAMUTEQ: Interface de R pour les analyses multidimensionnelles de textes et de questionnaires (computer software) [Internet]. Available from: http://www.iramuteq.org. Last access: 21 mars. 2019.

Contrastive analysis of affective domain in the professional register: Spanish and English doctor-patient online interaction

Alicia Ricart Vayá[a]; Rosa Giménez Moreno[a]

[a]*Department of English and German Philology, Universitat de València*

***Keywords**: register variation; corpus linguistics; translation; appraisal; sentiment*

Abstract

The complexity of the communication between doctor and patient has been studied from multiple points of view (Ong et al., 1995); however, the information that reaches both doctors (e.g. in their university education) and patients (e.g. through social or health media) to facilitate the interaction that occurs in this type of exchanges today is still scarce. Part of the difficulty in their investigation, and subsequently in their teaching/learning (Cate and Haes, 2009), lies in the fact that their degree of success depends, not only on the information accuracy, but also on the affective tactics used by doctors to reply to the patients' requests and/or questions, and vice versa (McNeilis, 2001) The systematic study of these tactics of great emotional and subjective magnitude is nowadays even more complex when taking into account the variety of socio-cultural factors that can converge, the alternation of interpersonal roles that can be adopted and the diversity of contexts in which these consultations can take place (e.g. by telephone, in online forums, mobile applications, etc.).

Following previous research on linguistic appraisal mechanisms (Martin and White, 2005) and professional register variation (Giménez-Moreno, 2006, 2011), the aim of this study is to analyse the interpersonal roles, attitude and affective strategies of doctors and patients in Peninsular Spanish and British English online forums, in order to determine the main linguistic and multimodal factors that influence the emotional effect that this communicative exchange conveys and contrast these factors in both corpora of online medical consultations.

Each corpus consists of a significant number of interactive threads in which patients formulate at least one consultation to real doctors and they respond with at least one full explanation. Parameters and methodological principles of RIB register analysis (Giménez-Moreno and Martínez-Sierra, 2017), appraisal analysis (Whitelaw et al., 2005) and sentiment analysis (e.g. Bobicev and Sokolova, 2018) have been adapted and used to obtain both qualitative and quantitative results on the affective tactics used by the participants in these corpora.

As the results show, doctors in this context use, in addition to the expected politeness and deference conventions (Berland et al. 2001), a range of interpersonal techniques to empathise with patients and mitigate or positivise the impact of taboo issues (Goldsmith, 2000; Ricart-Vayá and Giménez-Moreno, 2017). They also display, on the one hand, a series of affective strategies (Burleson and Goldsmith, 1998) to sensitise, persuade and encourage their patients to follow their indications and instructions; and on the other hand, a set of attenuation tactics (Caffi, 1999) when they warn and alert their patients about more urgent or serious issues.

The type and number of strategies used in each corpus vary considerably so that the results can also be useful for the teaching of medical-patient communication in the case of students of Spanish and English as L2 and for specialists in translation and interpretation of the current medical discourse. Additionally, our results may also help medical doctors improve in terms of understanding and learning about the affective domain both in their own and in the British or Spanish culture.

References

Berland, G. K., Elliot, M. N., Morales, L. S., Algazy, J. I., Kravitz, R. L., Broder, M. S., 2001. Health information on the Internet: Accessibility, quality, and readability in English and Spanish. Journal of the American Medical Association 285: 2612– 2621.

Bobicev, V., Sokolova, M., 2018. Thumbs up and down: sentiment analysis of medical online forums. Proceedings of the 3rd Social Media Mining for Health

Applications (SMM4H) Workshop & Shared Task, pp. 22–26. Brussels 31/10/2018, Association for Computational Linguistics.

Burleson, B. R., Goldsmith, D. J., 1998. How the comforting process works: Emotional distress through conversationally induced reappraisals. In P. A. Andersen and L. K. Guerrero (eds.), Handbook of communication and emotion: Research, theory, applications, and context, pp. 246–281. San Diego: Academic Press.

Caffi, C. 1999. On mitigation. Journal of Pragmatics 31: 881-909.

Giménez-Moreno, R. 2006. A new approach to register variation: the missing link. Ibérica 12: 89-110.

Giménez-Moreno, R. 2011. Register variation in international business correspondence. International Journal of English Studies 11/1: 15-34.

Giménez-Moreno, R., Martínez-Sierra, J.J., 2017. Roles and registers in digital forum interaction: developing a communicative identity-based approach to register variation. RLA: Revista de Lingüística Teórica y Aplicada 55/2: 143-167.

Goldsmith, D. J. 2000. Soliciting advice: The role of sequential placement in mitigating face threat. Communication Monographs 67: 1-19.

Martin, J.R., White, P.R., 2005. The Language of Evaluation: Appraisal in English, London: Palgrave Macmillan.

McNeilis, K. S. 2001. Analyzing communication competence in medical consultations. Journal of Health Communication 13: 5-18.

Ong, L. M., De Haes, J. C., Hoos, A. M., Lammes, F. B., 1995. Doctor-patient communication: A review of the literature. Journal of Social Science & Literature 40/7: 903-918.

Ricart-Vayá, A., Giménez-Moreno, R., 2017. The expression of taboo and corruption in the British academic milieu: a lexical-semantic study of the university press. 1st International Conference on Corpus Analysis in Academic Discourse, Valencia 23/11/2017, Universidad Politécnica de Valencia.

A Corpus-Based Approach to Alliterative Collocations in English

Gustavo A. Rodríguez Martín

Departamento de Filología Inglesa, Universidad de Extremadura, Spain

Keywords: *alliteration; collocations; binomials; corpus stylistics; variation*

Abstract

Alliteration has received a great deal of scholarly attention as part of the stylistic repertoire of many literary authors, from medieval poets to Lewis Carroll or Poe. The same applies to alliteration and other phonetic devices in their role as historical and cultural elements that have shaped the evolution of the English language and a number of folklore genres (Adams; Lawton; Chism; Minkova). However, the study of alliteration outside literary discourse has been for the most part neglected as a marginal phenomenon (Boers et al.), and it is only mentioned among all the other "poetic devices" that certain phraseological genres sometimes include, whether at present or in historical terms (Norrick, 46 et passim; Deskis). This neglect is far from justified in the domain of phraseology, where plenty of common expressions hinge on alliteration— "the belle at the ball," "pay the piper," and "trick or treat?" quickly come to mind. Another aspect that seems to be missing from the relevant literature is that practically all publications on the question fail to follow a replicable methodology, with but few exceptions (author, forthcoming).

Thus, the purpose of this paper is to apply a corpus-based methodology to the study of alliteration in English phraseology outside literature. More specifically, this study seeks to analyze alliterative two-word combinations with data from different online corpora.[7] The concept of alliteration that has been adopted here follows Halmari and, among other relevant notions, it only considers alliteration in initial position. The present study faces, first and foremost, the challenge of automating queries that solve the discrepancies between sound and spelling in the English language. The technical issues that were tackled in order to overcome these constraints and the amount of data available account for the presentation of preliminary results that only focus on plosive bilabials (/p/ and /b/ phonemes). Indeed, these are the phonemes that are most commonly spelled with the same initial letter ('p' and 'b', respectively). As a result, very few instances have had to be eliminated from the sample.

The results here will hopefully shed light on questions such as the frequency, distribution, register, genre, syntax, and stylistic use of these Phraseological Units— results that can will be extrapolated to and contrasted with other manifestations of alliteration in further research. For example, alliteration shows a strong correlation with idiomaticity, because there seems to be little room for systemic variation (plural forms, different verb tenses, even when alliteration is not altered. Also, alliteration seems to work best with certain grammatical categories (adjective-noun) than with others (verb-adverb). Finally, data from corpora of British English and American English will also provide interesting insights on the prevalence of this phenomenon across different varieties of the language.

[7] Corpus data available at https://corpus.byu.edu/.

Here, it is interesting to note that the most striking differences stem from idiosyncratic phenomena such as sports, food, and rituals.

References

Adams, Percy G. (1977). Graces of Harmony: Alliteration, Assonance, and Consonance in Eighteen-Century British Poetry. Athens: The University of Georgia Press.

Rodríguez Martín, Gustavo A. (2019). Mapping and Interpreting Alliteration in Ulysses: A Corpus-Based Study. James Joyce Quarterly (forthcoming), 2019.

Boers, Frank et al. (2012). Are Alliterative Word Combinations Comparatively Easy To Remember For Adult Learners? RELC Journal 43(1), 127–135.

Chism, Christine. (2002). Alliterative Revivals. Philadelphia: University of Pennsylvania Press.

Deskis, Susan E. (2016). Alliterative Proverbs in Medieval England: Language Choice and Literary Meaning. Ohio State University Press.

Halmari, Helena. (2011). Alliteration in Inaugural Addresses: From George Washington to Barack Obama. In Jonathan Roper (ed.), Alliteration in Culture (pp. 45-61). Basingstoke: Palgrave Macmillan.

Lawton, David, ed. (1982). Middle English Alliterative Poetry and Its Literary Background. Cambridge: D.S. Brewer.

Minkova, Donka. (2003). Alliteration and Sound Change in Early English. Cambridge: Cambridge University Press.

Norrick, Neal R. (1985). How Proverbs Mean: Semantic Studies in English Proverbs. Berlin: Mouton.

Teaching the pragmatics of speech acts: a corpus-based proposal.

Manuel Rodríguez-Peñarroja

English Studies Department, Universitat Jaume I.

***Keywords**: Corpus linguistics; corpus pragmatics; fictional TV discourse; audiovisual input, EFL teaching*

Abstract

Corpus pragmatics, and in particular the study of fictional TV discourse, is a thriving field in linguistics, as proved by volumes such as Telecinematic Discourse (2011), Watching TV with a Linguist (2016), and The Pragmatics of Fiction (2017). As for the intersection between Corpus Linguistics (CL) and ELT, numerous publications (see O'Keeffe and McCarthy, 2010; Campoy-Cubillo et al. 2010) proved that EFL students certainly benefit from this approach.

The purpose of this talk is to show the appropriateness of using CL methodologies in order to help EFL students seeking to attain a B2 level learn about the pragmatics involved in using certain speech acts (SA henceforth) (Rühlemann and Aijmer, 2015). SAs are the most basic communicative units; however, their identification is not an easy task. First, because some SAs conventionalised forms (Aijmer, 1996; Deutchman, 2003) may not always correspond with their intended illocutionary force. Secondly, indirect

SAs are difficult to identify using computer-aided searches (McAllister, 2015). Consequently, the present research combines computer-based searches and line-by-nine reading (Garcia, 2007; Santmaría-García, 2007) for the exploration of a multimodal ad hoc corpus which comprises the first season of the TV series "Life on Mars" & "How not to Live your life".

The corpus was first explored quantitatively, using the tools in Antconc, in particular clusters and n-grams. The results obtained facilitated the task of localising and selecting the most frequent SA structures (Boxer, 2010; Chang, 2010; Kondo, 2010; MartínezFlor, 2005; Salazar et al, 2009; Trosborg, 1995), which were afterwards manually explored using Concordance Plot and File View with the intention of checking the SAs' illocutionary force and using them as raw material to design contextualised tasks for the students.

The teaching methodology combines CL methodologies and a Focus on Forms approach (Harmer, 2007), explicit-deductive type of instruction (Halenko and Jones, 2011) including metapragmatic information provision (Birjandi and Derakhshan, 2014; Glaser, 2013) and the use of identification and awareness-raising tasks (Abrams, 2014). The reasons underlying the validity of the study are mainly two. First, as argued by Fuster-Marquez and Gregori-Signes (2018), an important ingredient of any methodology is that it wins the acceptance of learners [….] and that the students are "enthusiastic and willing to accept the proposal as a way of complementing exercises in textbooks". Secondly the results and progress reached by the students proved it to be an adequate teaching tool.

References

Abrams, Z. I. (2014). Using film to provide a context for teaching L2 pragmatics, System, 46, 55–64.

Aijmer, K. (1996). Conversational Routines in English: Convention and Creativity. London: Longman.

Aijmer, K., & Rühlemann, C. (Eds.). (2015). Corpus pragmatics. Cambridge University Press.

Birjandi, P., & Derakhshan, A. (2014). The impact of consciousness-raising videodriven vignettes on the pragmatic development of apology, request, & refusal. Applied Research on English Language, 3, 67-85.

Boxer, D. (2010). Complaints, How to gripe and establish rapport. In A. Martínez-Flor and E. Usó-Juan (eds.), Speech act performance: theoretical, empirical and methodological issues (pp. 163-178). Amsterdam: John Benjamins.

Campoy, M. C., Cubillo, M. C. C., Belles-Fortuno, B., & Gea-Valor, M. L. (Eds.). (2010). Corpus-based approaches to English language teaching. A&C Black.

Chang, Yuh-Fang. (2010). "I no say you say is boring": the development of pragmatic competence in L2 apology. Language Sciences 32, (pp. 408-424).

Deutschmann, M. (2003). Apologising in British English. Umeå Institutionen för moderna språk, Umeå Universitet.

Fägersten, K. B., Richardson, K., Trotta, J., Ann, J., Green, H., & Eitelmann, M. (2016). Watching TV with a Linguist. Syracuse University Press.

Fuster-Márquez, M., & Gregori-Signes, C. (2018). Learning from learners: a nonstandard direct approach to the teaching of writing skills in EFL in a university context. Innovation in Language Learning and Teaching, 12(2), 164-176.

Garcia, P (2007). Pragmatics in academic contexts: A spoken corpus study. In M. C. Campoy and M. J. Luzón (eds), Spoken Corpora in Applied Linguistics, pp. 97–126. Bern: Peter Lang.

Glaser, K. (2013). The Neglected Combination: A Case for Explicit-Inductive Instruction in Teaching Pragmatics in ESL. TESL Canada Journal, 30(7), 150163.

Halenko, N., & Jones, C. (2011). Teaching pragmatic awareness of spoken requests to Chinese EAP learners in the UK: Is explicit instruction effective?. System, 39(2), 240-250.

Harmer, J. (2007). The practice of English language teaching. Harlow: Pearson Longman.

Kondo, S. (2010). Apologies. Raising learners' cross-cultural awareness. In A. Martínez-Flor and E. Usó-Juan (Eds.), Speech act performance. Theoretical, empirical and methodological issues (pp. 145-162). Amsterdam: John Benjamins.

Locher, M. A., & Jucker, A. H. (Eds.). (2017). Pragmatics of fiction (Vol. 12). Walter de Gruyter GmbH & Co KG.

Martínez Flor, A. (2005). A theoretical review of the speech act of suggesting: Towards a taxonomy for its use in FLT. Revista alicantina de estudios ingleses, No. 18 (Nov. 2005); (pp. 167-187).

McAllister, P. G. (2015). Speech acts: a synchronic perspective. Corpus pragmatics: A handbook, 29-51.

O'Keeffe, A., & McCarthy, M. (Eds.). (2010). The Routledge handbook of corpus linguistics. Routledge.

Piazza, R., Bednarek, M., & Rossi, F. (2011). Introduction: Analysing telecinematic discourse. Telecinematic discourse: Approaches to the language of films and television series. Amsterdam: John Benjamins Publishing Company, 1-20.

Salazar, P., Safont, M. P., & Codina, V. (2009). Refusal strategies: A proposal from a sociopragmatic approach. RAEL, Revista Electrónica de Lingüística Aplicada, 8. (pp. 139-150).

Santamaría-García, C. (2007). Preference structure in agreeing and disagreeing responses. In A. M. Hornero, M. J. Luzón and S. Murillo (eds), Corpus Linguistics: Applications for the Study of English, (pp. 113–123). Bern: Peter Lang.

Trosborg, A. (1995). Interlanguage pragmatics. Requests, complaints and apologies.

Berlin: Mouton de Gruyter.

Variation *-ity* and *-ness* in Early Modern English: A cross-genre perspective

Paula Rodríguez-Puente

Departamento de Filología Inglesa, Francesa y Alemana, Universidad de Oviedo

Keywords: *derivation; Early Modern English; register; style*

Abstract

This paper examines stylistic variation in the use of two roughly synonymous suffixes, the Romance *-ity* and the native *-ness*, typically added to adjectives for the creation of abstract nouns (e.g. *curious – curiosity*; *happy – happiness*). In spite of their apparent similitude, the two suffixes differ with regard to the kinds of bases they attach to, the genres they appear in, and, arguably, their semantics (Säily 2014: 1), something which relates to the more learned and prestigious connotations of the borrowed form *-ity*. In examining the Middle English (ME) section of the *Helsinki Corpus*, Dalton-Puffer (1996: 75) notes that native suffixes like *-ness* and *-ung/-ing* experience a rapid decline during the period in favour of borrowed suffixes due to the influence of French and Latin as languages of prestige. Similar results are reported by Palmer (2015), who finds that the frequency of *-ness* was in relative decline in the letters of the *Corpus of Early English Correspondence* (1401-1600) compared to other several borrowed derivatives, something which has been accounted for in terms of derivational rivalry with *-ity* (see Riddle 1985; Cowie 1998; Anderson 2000; Säily & Suomela 2009). Gardner (2014) studies alternations in the productivity of the two suffixes in multi-genre corpora in the ME period, showing that genre variation is crucial in the case of *-ity*, which is particularly productive in learned types of writing, such as sermons, documents, literary texts and translations.

Although research in the alternation of the two suffixes is not new, a thorough cross-genre comparison in their use during the Early Modern English (EModE; 15001700) period has not been carried out yet, to the best of my knowledge. This is a crucial period in the expansion of the English vocabulary, not only due to the introduction of new loan words but also to the highly productive use of word formation processes (Nevalainen 1999: 336-337). My aim here is to compare the usage, productivity and development of the two suffixes in several corpora containing a wide range of formal and informal speech-related and writing-based and writing-purposed text types (Culpeper & Kytö 2010: 17) in order to assess whether stylistic variation is perceived across speech and writing. To my purpose, data have been extracted from a *Corpus of English Dialogues* (1560-1760; Kytö & Culpeper 2006), the *Penn-Helsinki Parsed Corpus of Early Modern English* (1500-1710; Kroch et al. 2004) and the EModE section of the *Corpus of Historical English Law Reports, 1535-1999* (Rodríguez-Puente et al. 2018), thus covering a wide variety of text types distributed along the formal-informal continuum and the speech-written continuum. The initial hypothesis is that *-ity* is more productive in formal written genres, whereas -

ness will predominate in informal speech-related ones. However, during a time in which the standard was being shaped and new genres were emerging, variation at register level can be expected.

References

Anderson, K. (2000). Productivity in English nominal and adjectival derivation, 1100–2000. PhD dissertation, University of Western Australia.

Cowie, C. (1998). Diachronic word-formation: A corpus-based study of derived nominalizations in the history of English. PhD dissertation, University of Cambridge.

Culpeper, J. & Kytö, M. (2010). Early Modern English dialogues. Spoken interaction as writing. Cambridge: Cambridge University Press.

Dalton-Puffer, C. (1996). The French influence on Middle English morphology: A corpus-based study of derivation. Berlin: Mouton de Gruyter.

Gardner, A.C. (2014). Derivation in Middle English: Regional and text type variation. Mémoires de la Société Néophilologique de Helsinki XCII. Helsinki: Société Néophilologique.

Kroch, A., Santorini, B. & Delfs, L. (2004). Penn-Helsinki Parsed Corpus of Early Modern English.

Kytö, M. & Culpeper, J. (2006). A Corpus of English Dialogues 1560-1760.

Nevalainen, T. (1999). Early Modern English lexis and semantics. In R. Lass (Ed.), The Cambridge history of the English language, Vol. III: 1476-1776 (pp. 332-458). Cambridge: Cambridge University Press.

Palmer, C.C. (2015). Measuring productivity diachronically: Nominal suffixes in English letters, 1400-1600. English Language and Linguistics 19(1): 107-29.

Riddle, E.M. (1985). A historical perspective on the productivity of the suffixes -ness and -ity. In J. Fisiak (Ed.), Historical semantics, historical word-formation (pp. 435461). Berlin: Mouton de Gruyter.

Rodríguez-Puente, P., Fanego, T., López-Couso, M.J., Méndez-Naya, B., Núñez-Pertejo, P., Blanco-García, C. & Tamaredo, I. (2018). Corpus of Historical English Law Reports 1535-1999 (CHELAR), v.2. Santiago de Compostela: Research Unit for Variation, Linguistic Change and Grammaticalization, University of Santiago de Compostela.

Säily, T. (2014). Sociolinguistic variation in English derivational productivity. Studies and methods in diachronic corpus linguistics. Mémoires de la Société Néophilologique de Helsinki XCIV. Helsinki: Société Néophilologique.

Säily, T. & Suomela, J. 2009. Comparing type counts: The case of women, men and -ity in early English letters. In A. Renouf & A. Kehoe (Eds.), Corpus linguistics: Refinements and reassessments (pp. 87-109). Amsterdam & New York: Rodopi.

Corpus of Historical English Law Reports 1535-1999: TEI-XML

Annotated Version

Paula Rodríguez-Puente[a], Teresa Fanego[b], Cristina Blanco-García[b]; Iván Tamaredo[b]

Departamento de Filología Inglesa, Francesa y Alemana, Universidad de Oviedo;
[b]Facultad de Filología Inglesa y Alemana, Universidad de Santiago de Compostela

Keywords: *corpus annotation; XML; law reports*

Abstract

A Corpus of Historical English Law Reports (CHELAR; Rodríguez-Puente et al. 2016) is a specialised corpus consisting of law reports, records of judicial decisions which are "cited by lawyers and judges for their use as precedent in subsequent cases" (*Encyclopædia Britannica Online* s.v. *law report*). Law reports typically contain an account of all the facts of the case, the arguments of the judge, his reasoning, the judgment he arrives at and the kind of authority and evidence he uses. The corpus contains approximately half a million words and is structured into nine periods of 50 years each, running from 1535 to 1999 (Rodríguez-Puente 2011; Fanego et al. 2017). CHELAR, which was made available in 2016 as plain text and with POS annotation (CLAWS C7; see Garside 1987), has been now continued by CHELAR v.2, an extensively revised and enhanced second version of the corpus (Rodríguez-Puente et al. 2018). CHELAR v.2 has also served as the basis for the creation of the newly released TEI-XML edition of CHELAR, which follows the *Text Encoding Initiative P5 Guidelines for Electronic Text Encoding and Interchange* developed by the Text Encoding Initiative Consortium (Bray et al. 2008). Multiple forms of annotation are currently being used to add value to primary data; among these, TEI-XML is a language used for storing and transporting data based on its inherent structure (see Carletta et al. 2004). Elements in a body of data are marked with customizable tags which can be further defined using attributes (Gries & Berez 2017). XML encoding has become the standard practice adopted in digitally based humanities research for present-day English and diachronic corpora, such as the *Helsinki Corpus* or the *Late Modern English Medical Texts 1700–1800* (Taavitsainen et al. 2014).

In this poster we focus on the particular structure and contents of law reports, the specific XML tags used for our purposes, and the problem-solving process followed during the annotation of the corpus. Although the annotation possibilities of the TEIXML schema are infinite, we have given priority to a type of annotation that satisfies the needs of our texts, yet at the same time facilitates a varied range of corpus analyses. In general, we advocate for a modest XML tagging which includes some renditional (e.g. italics), structural (paragraphs, line breaks, page breaks, etc.) and conceptual (foreign words, proper names, names of cases, etc.) features of the texts.

References

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E. & Yergeau, F. (Eds.). (2008). Extensible Markup Language (XML) 1.0. Fifth edition. W3C Recommendation 26 November 2008. <https://www.w3.org/XML/>

Carletta, J., McKelvie, D., Isard, A., Mengel, A., Klein, M. & Baun Møller, M. (2004). A generic approach to software support for linguistic annotation using XML. In G. Sampson & D. McCarthy (Eds.), Corpus linguistics: Readings in a widening discipline (pp. 449-459). London & New York: Continuum.

Encyclopædia Britannica Online. Available at <http://www.britannica.com>

Fanego, T., Rodríguez-Puente, P., López-Couso, M.J., Méndez-Naya, B., Núñez-Pertejo, P., Blanco-García, C. & Tamaredo, I. (2017). The Corpus of Historical English Law Reports 1535-1999 (CHELAR): A resource for analysing the development of English legal discourse. ICAME Journal 41: 53-82.

Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech & G. Sampson (Eds.), The computational analysis of English: A corpus-based approach (pp. 30-41). London: Longman.

Gries, S.T. & A.L. Berez. (2017). Linguistic annotation in/for corpus linguistics. In N. Ide & J. Pustejovsky (Eds.), Handbook of linguistic annotation (pp. 379-409). Berlin & New York: Springer.

A. Honkapohja, Kaislaniemi, S., Kauhanen, H., Kilpiö, M., Marttila, V., Nevalainen, T., Nurmi, A., Rissanen, M. & Tyrkkö, J. (2011). Helsinki Corpus TEI XML Edition. First edition. Implemented by H. Kauhanen & V. Marttila. Based on The Helsinki Corpus of English Texts (1991). Helsinki: The Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki.

Rodríguez-Puente, P. (2011). Introducing the Corpus of Historical English Law Reports: Structure and compilation techniques. Revista de Lenguas para Fines Específicos 17: 99-120.

Rodríguez-Puente, P., Fanego, T., López-Couso, M.J., Méndez-Naya, B., Núñez-Pertejo, P. (2016). Corpus of Historical English Law Reports 1535-1999 (CHELAR). Santiago de Compostela: Research Unit for Variation, Linguistic Change and Grammaticalization, University of Santiago de Compostela.

Rodríguez-Puente, P., Fanego, T., López-Couso, M.J., Méndez-Naya, B., Núñez-Pertejo, P., Blanco-García, C. & Tamaredo, I. (2018). Corpus of Historical English Law Reports 1535-1999 (CHELAR), v.2. Santiago de Compostela: Research Unit for Variation, Linguistic Change and Grammaticalization, University of Santiago de Compostela.

Taavitsainen, I., Hiltunen, T., Lehto, A., Marttila, V., Pahta, P., Ratia, M., Suhr, C. & Tyrkkö, J. (2014). Late Modern English Medical Texts 1700-1800: A corpus for analysing eighteenth-century medical English. ICAME Journal 38: 137-153.

Extraction of Terms Associated with Named Beaches from a Specialized Corpus

Juan Rojas-Garcia

Departamento de Traducción e Interpretación, Universidad de Granada

***Keywords**: named beach; conceptual information extraction; terminology; terminological knowledge base; text mining*

Abstract

EcoLexicon (http://ecolexicon.ugr.es) is an electronic, multilingual, terminological knowledge base on environmental sciences, whose flexible design permits the contextualization of data so that they are more relevant to specific subdomains and geographic areas (León-Araúz, Reimerink & Faber, 2013). However, to ease the geographic contextualization of concepts such as those belonging to the semantic category of landform, it is necessary to know what terms are semantically related to each type of landform according to the research papers published by experts, and how those terms are related to each other.

This paper describes a semi-automatic method for extracting knowledge about terms related to named beaches as a type of landform, from a specialized corpus on Coastal

Engineering comprising 7 million words. A GeoNames database dump (http://www.geonames.org/export/) was first applied to automatically match the sequences of words in the corpus which are the proper names of beaches (*Sound Beach, Barcovan Beach, Black River Beach,* etc).

For all the named beaches recognized in the corpus, their respective geographic coordinates, i.e. longitude and latitude, were automatically retrieved from the GeoNames database dump, and then visualized on top of a map, with rectangles that further informed about the occurrence frequency of each named beach in the corpus – the darker the rectangle, the larger the frequency of occurrence (see Figure 1). This type of visualization accounted for the representativeness of the corpus in reference to the location of beaches and the number of times that they were mentioned. Moreover, a hierarchical clustering technique was deployed in order to group the named beaches, based on their latitude and longitude. This allowed us to automatically annotate each beach with the geographical area (California, Brazil, etc) it belongs to.

Subsequently, the corpus was lemmatized for reducing data sparsity and yielding better performance in the Distributional Semantic Model (DSM) used to extract conceptual information (Kiela & Klark, 2014). The multi-word terms (i.e., a sequence of words that function as a single specialized term, such as *sediment transport process* or *beach profile change*) recorded in EcoLexicon were automatically matched in the lemmatized corpus and joined with underscores.

Then, a count-based DSM was selected to obtain term vectors since this type of DSM outperforms prediction-based ones on small-sized corpora (Ars, Willits & Jones, 2016; Sahlgren & Lenci, 2016). To group beaches with the same associated terms by means of a clustering technique, it was necessary to previously select the terms that best described the beaches. This term selection procedure consisted of five statistical criteria suggested by Moisl (2011), and Moisl (2015, pp. 71-92).

The preliminary results show that there is a slight association between the geographical areas of the named beaches and the processes mentioned by researchers affecting them. Once these experimental results were validated by Coastal Engineering experts, the knowledge extracted with this method facilitates the geographical contextualization of EcoLexicon with regard to beaches, in the sense that a specific named beach can be linked to its more highly associated terms dealt with in the corpus data.

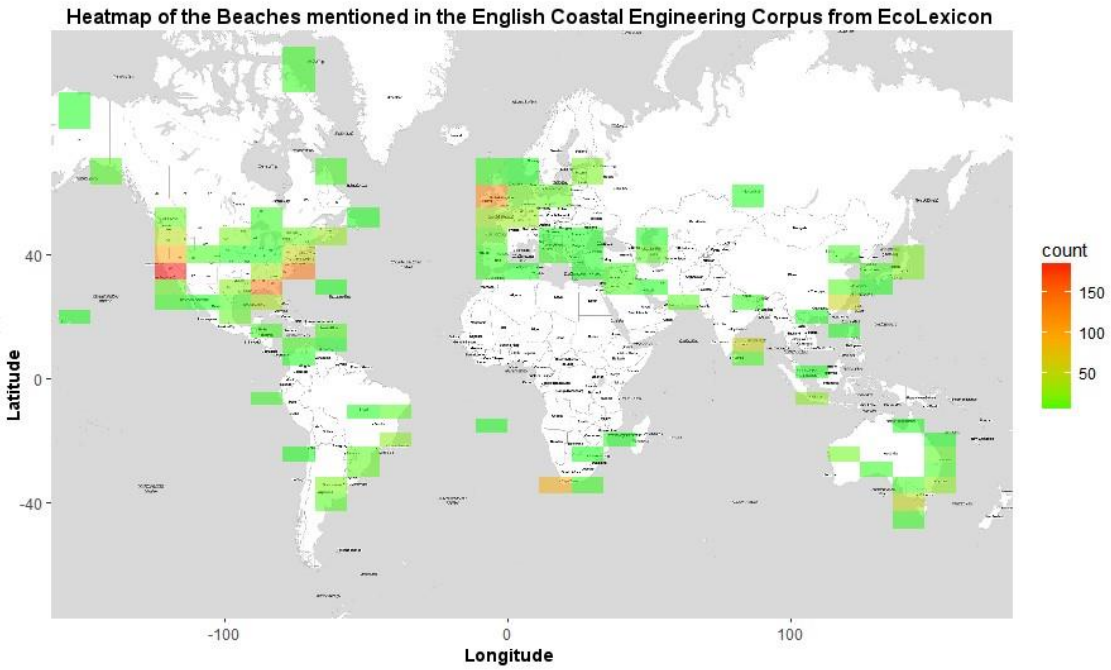


Figure 1: Heatmap (location and counts) of the beaches mentioned in the English Coastal Engineering Corpus from Ecolexicon Database.

References

Asr, F.; Willits, J., & Jones, M. (2016). Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained on Child-directed Speech. In A. Papafragou, D. J. Grodner, D. Mirman & J. Trueswell (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 1092-1097). Philadelphia, Pennsylvania: CogSci.

Kiela, D., & Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. In: Proceedings of the 2nd Workshop on Continuous Vector Space

Models and their Compositionality (CVSC) (pp. 21-30). Gothenburg, Sweden: EACL.

León-Araúz, P.; Reimerink, A., & Faber, P. (2013). Multidimensional and Multimodal Information in EcoLexicon. In A. Przepiórkowski, M. Piasecki, K. Jassem & P. Fuglewicz (Eds.), Computational Linguistics (pp. 143-161). Berlin, Heidelberg: Springer, Studies in Computational Intelligence, 458.

Moisl, H. (2011). Finding the Minimum Document Length for Reliable Clustering of Multi-Document Natural Language Corpora. Journal of Quantitative Linguistics, 18, 23-52. doi.org/10.1080/09296174.2011.533588.

Moisl, H. (2015). Cluster Analysis for Corpus Linguistics. Berlin: Walter de Gruyter.

Sahlgren, M., & Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 975-980). Austin, Texas: ACL.

Lengua y género: los corpus al servicio del desarrollo de estrategias discursivas no discriminatorias en aprendientes de L2/LE

Esperanza Román-Mendoza

Modern and Classical Languages, George Mason University, United States of America

Keywords: *lenguaje inclusivo; género; aprendizaje de L2/LE; corpus del español*

Abstract

En los últimos treinta años se han producido numerosos avances en la investigación sobre los usos sexistas del lenguaje, incluido el español, y las ramificaciones que estos tienen en la vida personal, académica y profesional, sobre todo de mujeres y colectivos LGTB (Bengoechea Bartolomé, 2015). Una de estas consecuencias es la discriminación por el sexo, definida por esta misma autora como "un tratamiento menos favorable hacia una persona de un sexo en comparación con el que recibiría una persona del otro sexo que se encontrara en similares circunstancias; puede ser directa (el trato desigual se produciría meramente por tratarse de alguien de ese sexo) o indirecta (el trato desigual se produce por la aplicación de un criterio que deja en desventaja ese sexo frente al otro)". (Bengoechea Bartolomé, 2014, p. 13)

Para paliar las consecuencias de dicha discriminación a través de estrategias lingüísticas que fomenten un uso más inclusivo del español, tanto actores gubernamentales a nivel supranacional, nacional y local, como organizaciones de toda índole (sin ánimo de lucro, académicas, empresariales, de los medios de comunicación de masas) han elaborado sus propias guías para el uso no sexista del lenguaje (p. ej., Ayuntamiento de Granada, 2008; Generalitat Valenciana, 2009; Congreso de Intendentes de Uruguay, 2010; Consejo Nacional de la Cultura y las Artes de Chile, 2016). Sin embargo, el impacto que han tenido estas medidas oficiales e iniciativas culturales puede resultar desconocido o parecer poco relevante a los ojos de un grupo de la población cuyo contacto con el español (1) se produce en contextos académicos en los que imperan ideologías lingüísticas de tipo tradicionalista/normativista o (2) se nutre de la interacción en espacios que no siempre son políticamente correctos, como las redes sociales o los medios audiovisuales. Nos referimos a los estudiantes de español como segunda lengua o lengua extranjera (L2/LE) en contextos de no inmersión.

Así surgió la iniciativa de crear un curso sobre lenguaje inclusivo en el mundo hispanohablante dirigido a estudiantes avanzados de español a nivel de grado y de máster en el que se estudiara las relaciones entre lenguaje y género desde una nueva perspectiva con la ayuda de los corpus de español disponibles en la red. Los objetivos principales del curso fueron: (1) familiarizarse e identificar las medidas de carácter lingüístico que se pueden tomar para favorecer el trato no discriminatorio hacia una persona por razón de su sexo y (2) localizar mediante búsquedas en corpus los ámbitos y las regiones en las que estas medidas han tenido más impacto y analizar su difusión en los últimos 10 años.

Esta ponencia presenta en primer lugar el programa del curso y su enfoque desde la pedagogía crítica, esto es, como práctica educativa que "modifica la manera de estar en el mundo, de verlo, de enfocar las relaciones, de decidir qué es importante y qué no lo es". (Lledó Conill, 2016, p. 73). A continuación, se discuten las aportaciones de las guías

de lenguaje inclusivo publicadas en diversos países hispanohablantes y qué estrategias lingüísticas fueron susceptibles de ser estudiadas en este curso a través de los corpus. Además, se analizan los resultados del pre-test y post-test sobre lenguaje inclusivo realizado con el alumnado del curso. La presentación finaliza con una serie de recomendaciones surgidas durante las discusiones, como la necesidad de utilizar y crear otras fuentes de datos, además de los corpus tradicionales.

References

Ayuntamiento de Granada (2008). Propuestas para un uso no sexista del lenguaje administrativo. https://www.granada.org/inet/wmujer8.nsf/4809dddc96a1f632c12573f00040f584/603 ee0447179b3f6c125747a002dbd0f/$FILE/Manual%20lenguaje%20no%20sexista.PDF

Bengoechea Bartolomé, M. (2014). Las buenas alumnas ante los TFG atrapadas entre la cultura comunicativa femenina y el androcentrismo. En Unidad de Igualdade (ed). III Xornada de Innovación Educativa en Xénero: docencia e investigación, pp. 9-35. Vigo: Universidade de Vigo.

Bengoechea Bartolomé, M. (2015). Lengua y género. Madrid: Síntesis. Congreso de Intendentes/Comisión Interdepartamental de Género (2010): Guía de lenguaje inclusivo. http://www.inmujeres.gub.uy/innovaportal/file/21489/1/6guia_lenguaje_inclusivo_im m.pdf

Consejo Nacional de la Cultura y las Artes de Chile. (2016). Guía del lenguaje inclusivo de género. https://www.cultura.gob.cl/wp-content/uploads/2017/01/guia-lenguajehttps://www.cultura.gob.cl/wp-content/uploads/2017/01/guia-lenguaje-inclusivo-genero.pdfinclusivo-genero.pdf

Generalitat Valenciana (2010). Igualdad, Lenguaje y Administración: Propuestas para un uso no sexista del lenguaje. http://www.inclusio.gva.es/documents/610706/162187124/Igualdad%2C%20lenguaje+y+administraci%C3%B3n/9d54e5c2-6527-499c-b8ab-4154d8bc4c79

Lledó Cunill, E. (2016). Lengua, literatura y coeducación. Filanderas: Revista Interdisciplinar de Estudios Feministas, 1, 61-76.

Interactuación comunicativa multimodal frente una estimulación audiovisual simulada

Asier Romero Andonegi[a]; Irati de Pablo Delgado[b]

[a]Departmento de Didáctica de la Lengua y la Literatura, Universidad del País Vasco/Euskal Herriko Unibertsitatea, España; [b]Departmento de Didáctica de la Lengua y la Literatura, Universidad del País Vasco/Euskal Herriko Unibertsitatea, España

***Keywords**: Información visual; Percepción del habla; Desarrollo del lenguaje; Gesto; Mirada; Vocalización*

Abstract

En este trabajo se subraya la importancia del patrón multimodal comunicativo como rasgo fundamental de la condición humana. Como señala Breckinridge et al. (2017) en el lenguaje gestual humano las modalidades discursivas son coordinadas no sólo en el nivel fonológico sino también en los niveles semánticos y pragmáticos. Los niños/as comienzan a usar combinaciones intencionadas de gesto-habla-mirada cerca del final del primer año, precediendo a la etapa lingüística de la primera palabra (Borrás et al. 2013; Esteve-Gibert et al., 2016; Kita et. al. 2017). La presencia de estos comportamientos combinados ha sido interpretada como un indicador de comunicación intencional; y aunque existe un acuerdo bastante generalizado en la literatura sobre la edad de aparición de la conducta intencional, se observan discrepancias a la hora de invocar las causas subyacentes que posibilitan la aparición de dicha conducta, y del papel que puede desarrollar la percepción visual (Hoehl et al. 2012; Ouni et al. 2006; Snow, 2017)

Por lo tanto, teniendo en cuenta todo lo que se ha mencionado el objetivo principal es precisar la interactuación de los elementos presentes en la comunicación multimodal (mirada, gesto y vocalización) mientras se visualiza un cuento modificado en su código lingüístico (IDS-ADS), y definir cómo influyen las variables de género y lengua en niños y niñas al comienzo del segundo año de vida y en dos idiomas diferentes: español y euskera.

En el estudio han participado 10 niños y niñas de entre 13 y 14 meses residentes en las provincias de Madrid y de Bizkaia. Los niños/as presentan una diferente lengua materna (L1) atendiendo a su procedencia geográfica (Madrid con L1 español y Bizkaia con L1 euskera) y todos ellos están escolarizados en el aula de un año de Educación Infantil.

Metodológicamente se adaptó el cuento de Caperucita Roja (versión de los hermanos Grimm) para su visualización a través de diferentes láminas que representaban distintos pasajes del cuento. El estímulo auditivo fue grabado por una profesora bilingüe (vasco-español) de Educación Infantil en los dos idiomas y en dos versiones diferentes: (1) mediante el código lingüístico estándar para el lenguaje adulto y (2) mediante el código lingüístico *motherese*. Procedimentalmente, los niños/as tuvieron que visualizar los dos vídeos en su lengua materna, entre una visualización y otra se dejó un corto espacio temporal, y en todo momento el sujeto estuvo acompañado por su profesora.

Entre los resultados caben destacar: (1) el tiempo dedicado a mirar la historia y la frecuencia de la mirada al adulto aumenta en la condición de IDS independientemente del idioma de la narración; (2) en ningún caso la variable L1 resulta significativa para la

elección de un tipo comunicativo u otro, atendiendo a la diferente estimulación audiovisual; (3) los niños produjeron más comportamientos comunicativos multimodales mientras miraban la versión de IDS en español y en euskera. Estos resultados sugieren que IDS aumenta la atención y el compromiso social promoviendo episodios de atención conjunta y, por lo tanto, el desarrollo del lenguaje.

References

Borrás, J., Kaland, C.,, Prieto, P., & Swerts, J. (2013). Audiovisual correlates of

interrogativity: A comparative analysis of Catalan and Dutch. *Journal of Nonverbal Behavior,* 38(1), 53-66.

Breckinridge, R., Alibali, M.W, & Kelly, S.D. (eds.) (2017). *Why Gesture? How the hands function in speaking thinking and communicating*. Amsterdam: John Benjamins.

Esteve-Gibert, N., Liszkowski, U., & Prieto, P. (2016). Prosodic and gestural features distinguish the intention of pointing gestures in child-directed communication. In M. E. Armstrong, N. Henriksen & M. M. Vanrell (Eds.), *Intonational grammar in Ibero-Romance. Approaches across linguistic subfields* (pp. 251-276). Amsterdam: John Benjamins.

Hoehl, S., Wahl, S., Michel, C., & Striano, T. (2012). Effects of eye gaze cues provided by the caregiver compared to a stranger on infants' object processing doi:https://doi.org/10.1016/j.dcn.2011.07.015

Kita, S., Alibali, M. W., & Chu, M. (2017). How Do Gestures Influence Thinking and Speaking? The Gesture-for-Conceptualization Hypothesis. *Psychological Review*, 124(3), 245-266.

Ouni, S., Cohen, M., Ishak, H., & Massaro, M. (2006). Visual contribution to speech perception: Measuring the intelligibility of animated talking heads. *Journal on Audio, Speech, and Music Processing*, 17, 1-12.

Snow, D.P. (2017). Gesture and intonation are "sister system" of infant communication: Evidence from regression patterns of language development. *Language Sciences*, 59, 180-191.

Coordination and Subordination as Markers of Linguistic Complexity in Early Modern English Scientific Prose (1500–1700)

Jesús Romero-Barranco; Paula Rodríguez-Abruñeiras

Departament de Filologia Anglesa i Alemanya-IULMA, Universitat de València

***Keywords**: Linguistic Complexity; Coordination; Subordination; Early Modern English; Scientific Writing*

Abstract

Complexity was first defined by Simon as hierarchies of different elements originating from simplicity (1962: 468). In Linguistics, Givon (2009) has analysed syntactic complexity from the point of view of language typology; Nichols (2009) has assessed grammatical complexity cross-linguistically; and Maas (2009), among others, has studied the different levels of complexity in spoken and written registers. Furthermore, Lehto (2015) elaborated a diachronic analysis of the levels of complexity among different text types in early Modern English legal material, based on Biber's works on linguistic complexity (see Biber 1988; 1992). Biber analysed the distribution of 67 linguistic features across 481 texts belonging to 23 spoken and written registers and, by means of an exploratory factor analysis, five different dimensions of variation were identified among registers in English (1988: 170–198). In Biber (1992), 33 out of those 67 linguistic

features were labelled as potential markers of linguistic complexity, some of them indicating reduced complexity (i.e. *that* deletions, contractions or clause coordination, among others) and others pointing to increased complexity (i.e. nominalizations, subordination or passive constructions, among others). It is important to consider that what is being measured by way of the frequency of these linguistic features is the product rather than the process, that is, the representation of the surface structure of language. In other words, linguistic features such as that-deletions, contractions or the use of the anaphoric pronoun *it* would imply a less complex surface structure but, at the same time, these utterances would require a greater effort on the part of the listener/reader to decipher the message (for a similar discussion, see Pérez-Guerra 2007: 132). According to Biber (1992), coordination implies reduced linguistic complexity whereas subordination is identified as an indicator of increased linguistic complexity. Following this thread, Kohnen analysed coordination and subordination in Middle English and early Modern English sermons finding that the spread of subordination is correlated with a drop of coordination, as the levels of literacy increased in society and complex mechanisms (i.e. subordination) progressively outnumbered coordination instances that were characteristic of the spoken register (2007: 294). However, the competition between coordination and subordination has not been hitherto evaluated in the history of scientific English (i.e. Early Modern English), especially considering its different text types. In the light of this, the present paper analyses the levels of linguistic complexity through the use of coordination and subordination in three different text types belonging to the field of scientific writing: theoretical texts, surgical texts and remedies. The objectives are the following: 1) to analyse the use and distribution of coordinated and subordinated constructions in Early Modern English scientific prose; 2) to assess the occurrence of the phenomenon in the different text types; and 3) to classify the instances of coordination and subordination in order to see whether the different text types feature any likely preference. The data used as source of evidence come from the *Early Modern English Medical Texts* (EMEMT).

References

Biber, D. (1988). Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. Discourse Processes 15, 133-163.

Givon, T. (2009). The Genesis of Syntactic Complexity. Amsterdam: John Benjamins.

Kohnen, T. (2007). 'Connective profiles' in the history of English texts. In U. Lenker & A. MeurmanSolin (Eds.), Connectives in the History of English (pp. 289-308). Amsterdam: John Benjamins.

Maas, U. (2009). Orality versus literacy as a dimension of complexity. In G. Sampson, D. Gil & P. Trudgill (Eds.), Language Complexity as an Evolving Variable (pp. 164-177). Oxford: Oxford University Press.

Lehto, A. (2015). The Genre of Early Modern English Statutes: Complexity in Historical Legal Language. PhD Dissertation. University of Helsinki.

Nichols, J. (2009). Linguistic complexity: A comprehensive definition and survey. In G. Sampson, D. Gil & P. Trudgill (Eds.), Language Complexity as an Evolving Variable (pp. 110-125). Oxford: Oxford University Press.

Pahta, P. & I. Taavitsainen. (2011). An interdisciplinary approach to medical writing in Early Modern English. In I. Taavitsainen & P. Pahta (Eds.), Medical Writing in Early Modern English (pp. 1-8). Cambridge: Cambridge University Press.

Pérez-Guerra, J. (2007). 'Am I more complex when I speak or when I write?' A corpusbased study on linguistic complexity in spoken and written Present-day English". In M.C. Campoy & M.J. Luzón (Eds.), Spoken Corpora in Applied Linguistics (pp. 127-146). Bern: Peter Lang.

Simon, H.A. (1962). The architecture of complexity. Proceedings of the American Philosophical Society 106(6), 467-482.

A cognitive follow up on phrasal verbs in English for the Police: the case of OUT, UP and OFF

Andreea Rosca

Departament de Filologia Anglesa y Alemanya, Universitat de València

***Keywords**: phrasal verbs; Cognitive Linguistics; frequency; English for the Police; teaching materials*

Abstract

As Liu (2011) pointed out, the frequency of phrasal verbs (PVs) is register specific and as such, English for Specific Purposes (ESP) learners should become familiar with the most productive PVs in their own field of study. Since our ultimate goal is to create teaching materials for L2 trainee police officers, the present research aims to determine the frequency and meaning extensions of PVs with the particles *out*, *up* and *off* in the context of crime and police investigative work. For this reason, we have compiled our own corpus composed of spoken dialogues extracted from the script of the American TV series *Castle* (seasons 1 to 4). The corpus was cleaned of stage directions, character names, and all incidental languages, leaving a total of 504,124 words. As far as the choice of the TV series is concerned, this is motivated by its growing viewer popularity both in the US and Europe (approximately 10 to 12 million viewers).

Building on McCarthy and O'Dell's (2004) work on crime-related PVs, our study tries to broaden the scope from purely criminal activities (e.g. *break out of jail*, *beat sb up, tip sb off*) to PVs describing the actions carried out by the police in their investigation, the processing of evidence, and the reactions of crime victims. After the preparation of the corpus, searches were conducted using the AntConc concordance tool to obtain all the combinations with the particles *out*, *up*, and *off*. We sifted through the PVs to divide them into two categories, viz. those related or unrelated to a criminal context. This classification left us with a total of 1786 and 1817 tokens respectively. For example, verbs like *ask sb out* (e.g. *Do you think he's asked her out yet?* [S02E20]), or *grow up* (e.g. *Roger was my friend. We grew up together* [S02E24]) were considered as unrelated to the context of crime. Out of the1786 tokens, a total of 266 lexical verb-types were encountered for all three particles, more specifically 88 lexical verb-types for *out*, 127 for *up* and 51 for *off* respectively.

It is noteworthy to mention that our PVs are distributed differently on the cline of prototypicality. Thus, PVs like *alibi out* (e.g. *Anyway, they alibied out. They were in New Paltz when Zack was killed*. [S03E21]) or *lawyer up* (e.g. *The second you lawyer up and leave, you become the focus of a major investigation*. [S03E06]) are undoubtedly more prototypical than *pull out* (e.g. *And then the kindly Mexican gentlemen […] suddenly pull out guns and try to kill me*. [S04E21]) or *open up* (e.g. *Charlie Coleman! NYPD! Open up !*[S04E22]). This can be accounted for by the fact that the semantic meaning of the former allows for an automatic association with a criminal context, i.e. it is normally law offenders that need a solid alibi or the services of a lawyer to prove their innocence. In the case of the latter it is the surrounding context that triggers the connection to crime (e.g. using a firearm to shoot someone is a felony and suspects can be requested to provide access to their homes if the police want to search premises).

References

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. Tesol Quarterly, 45(4), 661-688.

McCarthy, M. & O'Dell, F. (2004). English phrasal verbs in use. Cambridge: Cambridge University Press.

El uso de recursos elativos en hablantes de Valencia

Jorge Roselló

Departament d'Estudis Hispànics, Universitat de València

Keywords: *recursos elativos; corpus; PRESEEA; sociolingüística; valenciano*

Abstract

Podemos definir los recursos elativos como aquellos elementos lingüísticos que expresan el grado extremo de una cualidad. La gramática tradicional los denominaba superlativos absolutos, y para la *Nueva gramática de la lengua española* (2009: 921), los adjetivos elativos, como *atroz* o *fabuloso*, se caracterizan por "no admitir los recursos morfológicos que se suelen emplear para expresar el grado elevado o extremo". En la presente comunicación, llevamos a cabo un estudio centrado en la aparición de elativos morfológicos, esto es, los prefijos de grado extremo como *re-*, *requete-* o *super-* y los sufijos que expresan esa misma noción (*-ísimo*, *-érrimo*), y los elativos léxicos, es decir, los que denotan tales propiedades en función de su naturaleza léxica, encontrados en el corpus PRESEVAL. Dicho corpus, integrado dentro del proyecto PRESEEA, constituye una de las publicaciones más completas para el estudio de los fenómenos lingüísticos del español hablado en Valencia. La *NGLE* establece un total de 36 adjetivos elativos, entre los que destacamos como más usados en nuestro corpus *enorme, horroroso, increíble, magnífico, maravilloso, mínimo, precioso, terrible, tremendo*, etc.

Mediante estos recursos podemos ver el grado de subjetividad del hablante respecto a lo que está expresando, ya que todos los individuos recurrimos, en mayor o menos medida, a ellos. Desde el punto de vista sociopragmático, como indica Paredes (2017), el empleo

de estas unidades afecta a la imagen personal que transmite el hablante, dado el carácter extremo del contenido semántico que aportan.

En este estudio que estamos realizando y cuyos resultados provisionales presentamos aquí, se han analizado los recursos elativos mencionados anteriormente en relación con cuatro variables: la edad, con tres franjas: primera generación (20-34 años), segunda generación (35-55 años), y tercera generación (más de 55 años); el sexo (hombre y mujer), el nivel sociocultural (bajo, medio y alto) y la lengua habitual, ya que, aunque todos los informantes utilizan el castellano, muchos tienen como lengua materna el valenciano. Este dato nos interesa especialmente para caracterizar el español hablado en Valencia.

Igualmente, hemos comparado los resultados con otras comunidades de habla (Madrid) que han realizado estudios similares basados en el corpus PRESEEA.

References

Paredes García, F. (2017). "Uso y adquisición de recursos elativos como indicadores de integración sociolingüística en Madrid", en M.C. Ainciburu y C. Fernández (editores). La adquisición de la lengua española: aprendizaje, enseñanza, evaluación, Buenos Aires, Autores de Argentina.

RAE y ASALE (2009). Nueva gramática de la lengua española, Madrid, Espasa.

Towards an annotated corpus of reporting clauses in 19th-century English fictional narratives: some notes on reporting verbs

Pablo Ruano

Departamento de Filología Inglesa, Universidad de Extremadura

***Keywords**: 19th-century English fiction; annotated corpus; reporting clause; reporting verbs*

Abstract

In this presentation I discuss how I am compiling and annotating a corpus of reporting clauses introducing stretches of direct speech (and also thought and writing) presentation in 19th-century English fiction. To do so, I have used the *Dnov* (*Dickens's novels*) (c. 4 million words) and the *19Cor* (*19th-century reference corpus*) (c. 4.5 million words) from *CLiC* (Corpus Linguistics in Cheshire) (see Mahlberg et al., 2016) as a starting point. *DNov* is a corpus of Dickens's novels, whereas *19Cor* is a corpus of 19th-century English novels to which *DNov* is frequently compared. Both corpora have been marked up to distinguish between several subsets. The main distinction is that of quoted and nonquoted text, which distinguishes between quoted words (characters' words) and the rest of the text (narration). I have used the non-quoted text subset as the basis on which to build my corpus of reporting clauses.

Specifically, in this presentation I will focus on the procedure I have followed to identify reporting clauses and on the labels I have used to annotate my corpus. Specifically, each reporting clause is marked with six labels, which indicate: (i) the position of the reporting

clause (initial, final or suspended reporting clause) with regard to the reported speech, (ii) the reporting verb that introduces the reported discourse (*say*, *answer*, *ask*, *reply*, etc) and (iii) its type following Caldas-Coulthard's (1987) taxonomy (neutral, structuring, illocutionary, descriptive, discourse and signaling reporting verbs), (iv) the type of discourse (speech, thought & writing presentation), (v) the sex of the character whose discourse is being introduced (male or female), and (vi) the type of installment (weekly or monthly) in which the novel that contains the reporting clause was published.

Although the annotation of the corpus is still underway, a brief survey of some findings will be shown. Specifically, I will focus on the reporting verbs so far identified. As will be shown, the annotation process makes it possible to identify reporting verbs introducing characters' words (and thoughts) in every novel, which has resulted in a more comprehensive catalogue of this type of verbs than previous analyses of this element have shown, such as Wierzbicka's (1987) and Levin's (1993). Due to time constraints, I will mainly concentrate on those verbs which are normally not included in other catalogues, highlighting their functions from a stylistic point of view. Thanks to the different labels used to annotate the corpus, I will also discuss novelists' use of reporting verbs from different dimensions than those usually scrutinized in traditional approaches to these verbs. For example, I will discuss novelists' use of reporting verbs according to aspects such as the sex of the character whose discourse is being introduced, the position of the reporting clause, etc.

References

Caldas-Coulthard, C. R. (1987). Reported speech in written narrative texts. In R. M. Coulthard (Ed.), Discussing Discourse (pp.149–167). Birmingham: University of Birmingham.

Levin, B. (1993). English Verb Classes and Classifications. Chicago: The University of Chicago Press.

Mahlberg, M., Stockwell, P., de Joode, J., Smith, C. & O'Donnell, M.B. (2016). CLiC Dickens: novel uses of concordances for the integration of corpus stylistics and cognitive poetics. Corpora, 11(3), 433–463.

Wierzbicka, A. (1987). English Speech Act Verbs: A Semantic Dictionary. Sydney: Academic Press.

En torno a un corpus de noticias radiofónicas amazónico-peruanas para el análisis de la intensificación

Magali Ruiz Barrera

Departamento de Lingüística Aplicada, Universitat Politècnica de València, España

***Keywords**: intensificación; español amazónico peruano; noticias radiofónicas*

Abstract

Se han hecho numerosos estudios sobre las variedades del español peruano, sobre todo del español andino y costeño o de litoral (Rivarola, 1990; Cerrón Palomino, 2003; Escobar, 1978, entre otros). Por mucho tiempo el español amazónico peruano despertó

poco interés en los investigadores, sin embargo, a lo largo de poco más de una década, al parecer, esta situación está siendo revertida, está floreciendo con serias intenciones de cambiar el panorama pues hoy en día podemos encontrar estudios que describen la situación actual, de modo más o menos general, del español amazónico peruano (Ramírez, 2003; Marticorena, 2010; Vallejos, 2014; Chirif, 2016; Ruiz, 2016; Valenzuela y Jara, 2015). Como resultado de la mayoría de estos estudios incluyen la elaboración de diccionarios, glosarios amazónicos tomando en cuenta corpus de conversaciones coloquiales, literatura amazónica, hasta artes gramaticales, otros estudios se enfocan en explicar las características propias de la entonación de esta variedad de español, la cual se puede decir que es bastante llamativa.

Nosotros, por nuestra parte, queremos ser punto de partida para un tipo de estudio nuevo, por lo que centramos esta investigación en el análisis de la intensificación de la noticias radiofónicas de Loreto en Perú utilizando un corpus de noticias de radio. Los objetivos que persiguen nuestra investigación son, por un lado, identificar los mecanismos de intensificación que se utilizan en la presentación de noticias radiofónicas en Loreto en Perú en dos géneros (noticia explicada que expresa opinión y crónica) y en dos temáticas (política y sucesos) para luego comparar entre un género y otro. Es de nuestro interés, también, identificar la existencia o no de regionalismos en el uso de la intensificación.

Para este estudio, hemos recolectado un corpus de 10 horas de noticias radiofónicas, para la transcripción de los audios nos valdremos de Servicios Web de Amazon (AWS) denominado Amazon Transcribe, un servicio de reconocimiento de voz automático y para su edición se tendrá en cuenta la transcripción ortográfica. El corpus será analizado con la herramienta METOOL de la que se extraerán las formas intensificadoras para responder a nuestros objetivos y poder llegar a las conclusiones del caso.

Para nuestro análisis tendremos en cuenta trabajos que nos servirán de base como los de Albelda (2004, 2007, 2008), Briz (2011), Mur Dueñas (2011) y trabajos más actuales como el de Carrió y Moya (2018). La intensificación es considerada como una categoría pragmática que maximiza mayor fuerza a las acciones y puntos de vista (Briz, 2008), así que creemos que nuestro trabajo puede contribuir con las investigaciones del español hablado, puede proporcionar una explicación del uso y, probablemente, una aclaración sobre la formación del español amazónico peruano gracias al corpus inédito con el que contamos.

References

Albelda, M. (2004). La intensificación en el español coloquial. Universitat de València: Valencia.

Albelda, M. (2007). La intensificación como categoría pragmática: revisión y propuesta. Frankfurt: Peter Lang

Briz, A. (2011). La atenuación como categoría pragmática. Actos del VII Congreso Internacional de la Asociación Asiáticas de Hispanistas: 3-21.

Carrió-Pastor, M. y Moya, P. (2018). Estrategias de intensificación en los comentarios digitales sobre noticias en español: Un análisis de la variación entre España y Chile. Spanish in Context 15(3): 369-391.

Chirif, A. (2016). Diccionario amazónico voces del castellano en la selva peruana. Lima: Lluvia Editores.

Cerrón-Palomino, R. (2003). Castellano Andino. Aspectos sociolingüísticos, pedagógicos y gramaticales. Lima: Fondo Editorial Pontificia Universidad Católica del Perú, Cooperación Técnica Alemana GTZ.

Escobar, A. (1978). Variaciones sociolingüísticas del castellano en el Perú. Lima: IEP ediciones.

Marticorena, M. (2010). El castellano amazónico del Perú. Instituto de Investigaciones Educativas e Históricas de la Amazonía Peruana.

Mur-Dueñas, P. (2011). An intercultural analysis of metadiscourse features in research articles written in English and in Spanish. Journal of Pragmatics 43 (12), 30683079.

Ramírez, H. (2003). El español amazónico hablado en el Perú: hacia una sistematización de este dialecto. Lima: Juan Gutemberg.

Rivarola, J. (1990). La formación lingüística de Hispanoamérica. Lima: Fondo Editorial Pontificia Universidad Católica del Perú.

Valenzuela, P. y Jara, M. (2015). Usos del marcador discursivo "ya vuelta" en el castellano amazonico peruano. En A. Ezcurra (presidencia), First International Symposium on Amazonian Spanish. Simposio llevado a cabo en Lima, Perú.

Vallejos, R. (2014). Peruvian Amazonian Spanish: uncovering variation and deconstructing stereotypes. Spanish in Context, 11(3), 425-453.

Estudio de “deber (de) + infinitivo” en el CORPES, un corpus oral y escrito del español actual

Alicia San Mateo Valdehíta[a], Cecilia Criado de Diego[a]

[a]*Departamento de Lengua Española y Lingüística General, Universidad Nacional de Educación a Distancia, España*

***Keywords**: lingüística de corpus; perífrasis modales; deber (de) + infinitivo; modalidad; variación dialectal*

Abstract

En español las perífrasis verbales se clasifican según los valores aspectuales y modales que el verbo auxiliar aporta a la construcción. La Real Academia Española y la Asociación de Academias de la Lengua Española diferencian entre *modalidad radical* o *deóntica* y *modalidad epistémica*, y recomiendan la utilización de la perífrasis *deber* + infinitivo para expresar obligación (modalidad deóntica) y la variante preposicional *deber de* + infinitivo para indicar posibilidad o conjetura (modalidad epistémica) (RAE, 1973; RAE y ASALE, 2005, 2013). No obstante, en la *Nueva Gramática* (RAE y ASALE, 2009, p. 2144) indican que el empleo de la variante no preposicional como expresión de conjetura no es incorrecto. Las Academias de la Lengua (RAE, 1973, p. 448; RAE y ASALE, 2009, p. 2144) y lingüistas como Blas Arroyo (2013) o Gómez Molina (2013) constatan que los hablantes no respetan siempre las recomendaciones de uso de estas dos perífrasis.

Nuestro objetivo es estudiar la distribución de las 3364 ocurrencias de *deber (de)* + infinitivo en el Corpus del Español del Siglo XXI (CORPES) con el fin de analizar de qué manera los hablantes siguen las indicaciones académicas y si el factor de la modalidad, el tipo de discurso (oral y escrito) y el área dialectal condicionan la elección de las variantes perifrásticas no recomendadas[8].

Los resultados revelan que es mayoritaria la variante no preposicional (87,4%), tendencia detectada por Blas Arroyo (2011, p. 10). Para expresar conjetura, los hablantes prefieren significativamente *deber* (61,3%) —no recomendado, pero no incorrecto—. Pero, para indicar obligación, se decantan por la perífrasis aconsejada: solo en el 3,5% de los casos aparece *deber de*. Por lo tanto, la realidad de la lengua es que los usos incorrectos son los menos numerosos y la elección de la variante no preposicional es significativamente mayor, desde el punto de vista estadístico, en ambas modalidades.

Según el tipo de discurso, la oralidad favorece las variantes no recomendadas: la no preposicional en el 70,2% de los casos en los que se expresa conjetura; y la preposicional en el 13,3%, cuando se indica obligación. Sin embargo, hay que tener presente que las ocurrencias en textos orales representan solo el 21,1% del total de casos.

En cuanto a la influencia de la variable diatópica, en la modalidad epistémica hay una clara diferencia entre España y el resto de las áreas: los casos de *deber* están por encima del 70% en Chile, Antillas, Andes, Río de la Plata, Caribe Continental, y México y Centroamérica, mientras que en España representan el 33,6%. Por lo tanto, el uso aconsejado es el preferido únicamente en España y en las demás zonas la variante no preposicional se ha generalizado. En la modalidad deóntica, el mayor número de ocurrencias de *deber de* se registra en España y representa el 6,1%. En definitiva, el uso no recomendado en esta modalidad es minoritario.

Estadísticamente la procedencia dialectal es un factor significativo de incidencia en la elección de la variante de esta perífrasis en ambas modalidades. Los resultados obtenidos confirman las conclusiones de Blas Arroyo y Vellón Lahoz (2014) sobre la revitalización de *deber de* en España desde finales del siglo XX y principios del XXI; y la tendencia a su desaparición en América desde el XVIII.

References

Blas Arroyo, J. L. (2011). Deber (de) + infinitivo: ¿un caso de variación libre en español? Factores determinantes en un fenómeno de alternancia sintáctica. Revista de Filología Española, 91(1), 9-42.

Blas Arroyo, J. L. (2013). Norma y uso en un fenómeno de variación sintáctica. Nuevos datos a propósito de la oposición modal epistémico-deóntica y la variable deber/deber de + infinitivo. In E. Casanova & C. Calvo (Eds.), Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas (pp. 519-529). Berlin, Germany: E. de Gruyter.

Blas Arroyo, J. L., & Vellón Lahoz, J. (2014). La supervivencia de deber de + infinitivo en el español moderno (ss. XVIII-XX). Revista de Filología Española, 94(1),

[8] Véase el análisis de otros posibles factores condicionantes (tiempo verbal y tipo de texto) en Criado de Diego y San Mateo Valdehíta (en prensa). Los datos completos están disponibles en GREC-UNED, en LYNEAL (Letras y Números en Análisis Lingüísticos): http://shimoda.lllf.uam.es/ueda/lyneal/grecuned.htm.

938.

Criado de Diego, C., & San Mateo Valdehíta, A. (en prensa). Las perífrasis modales con deber en el español actual. Estudio de corpus. Spanish in Context (SiC).

Gómez Molina, J. R. (2013). Norma y usos de las perífrasis "deber + infinitivo"/"deber de + infinitivo. In J. R. Gómez Molina (Coord.), El español de Valencia: estudio sociolingüístico (pp. 71-108). Bern, Switzerland: Peter Lang.

Real Academia Española (RAE). Corpus del Español del Siglo XXI

(CORPES). Retrieved from http://www.rae.es

Real Academia de la Lengua Española (RAE) (1973). Esbozo de una nueva gramática de la lengua española. Madrid, Spain: Espasa.

Real Academia de la Lengua Española (RAE) & Asociación de Academias de la Lengua Española (ASELE) (2005). Diccionario panhispánico de dudas. Madrid, Spain: Santillana.

Real Academia de la Lengua Española (RAE) & Asociación de Academias de la Lengua Española (ASELE) (2009). Nueva Gramática de la lengua española. Madrid, Spain: Espasa Calpe.

Real Academia de la Lengua Española (RAE) & Asociación de Academias de la Lengua Española (ASELE) (2013). El buen uso del español. Madrid, Spain: Espasa.

La terminología de la discapacidad en las leyes españolas y británicas (1978-2016): un estudio basado en corpus

Jaime Sánchez Carnicer

Departamento de Lengua Española, Universidad de Valladolid, España

Palabras clave: *terminología; discapacidad; corpus; fraseología; inglés; español*

Abstract

Durante las últimas décadas del siglo pasado y todo este siglo se ha comprobado que los corpus (monolingües, bilingües, multilingües) son una herramienta importante para ayudar en las múltiples actividades que engloba el proceso de la traducción (Sinclair, 1998; Corpas Pastor,2001). Sin embargo, como indica Seghiri (2011), no siempre están en la red disponibles corpus de determinados campos del saber o que satisfagan todas las necesidades documentales del traductor y, por ello, la única solución es compilarlos. Asimismo, al mismo tiempo que los corpus adquirían esta importancia, se han ido elaborando leyes en nuestro país y en Reino Unido destinadas a procurar una vida mejor a las personas con discapacidad. Por tanto, en este trabajo pretendemos comprobar la evolución que ha sufrido la terminología empleada para denominar a las personas con algún tipo de discapacidad (ej. *disabled person, handicapped, deficiente, minusválido*), así como a las afecciones (ej. *blind, ceguera, discapacidad intelectual*). Para ello, extraeremos los resultados de un corpus virtual comparable que compilaremos siguiendo las cuatro fases presentadas por Seghiri (2017): (1) búsqueda y acceso a la

documentación, para lo que utilizaremos Internet para la búsqueda de leyes nacionales relativas a la discapacidad en España y Reino Unido promulgadas durante el periodo que abarca nuestra investigación (1978-2016); (2) descarga de las leyes; (3) normalización de los documentos descargados (transformación de todos los documentos a formato TXT); (4) almacenamiento de los archivos en carpetas y subcarpertas (por lenguas) y denominación de los mismos mediante un código alfanumérico propio, diseñado para este corpus, basado en la clasificación UNESCO. Para la extracción de los términos presentes en los documentos que compondrán nuestro corpus y que serán aquellos que analizaremos, nos basaremos en los criterios marcados por L'Homme (2004): pertenencia al campo de especialidad que se estudie, frecuencia de aparición en los textos que componen el corpus y su distribución en los mismos. Los resultados que se desprendan de este estudio nos permitirán avanzar en este campo del saber, tanto en la lengua inglesa como en la española; así como constatar si se ha producido evolución en esta terminología y se ha avanzado hacia un lenguaje más inclusivo en ambas lenguas o si, por el contrario, se ha producido una mayor evolución y unos cambios más significativos en una lengua u otra; así como, aunque se traten en ambas lenguas de textos especializados, observar en cuál de los dos la presencia de terminología de este campo del saber es mayor. Además, podremos estudiar la fraseología y las colocaciones en los documentos analizados de los aspectos estudiados en este trabajo (afecciones y denominación de las personas con discapacidad) y observar en cuál de las dos lenguas la distribución léxica de este campo del saber es mayor.

References

Corpas Pastor, G. (2001). Compilación de un corps ad hoc para la enseñanza de la traducción inversa especializada. TRANS. Revista de Traductología, 5, 155-184.

L'Homme M.C. (2004). Sélection des termes dans un corpus d'informatique: comparaison de corpus et critères lexico-sémantiques. En Euralex International Congress. Proceedings. Lorient (Francia), 583-593.

Seghiri, M. (2011). Metodología protocolizada de compilación de un corpus de seguros de viajes: aspectos de diseño y representatividad. RLA, Revista de Lingüística Teórica y Aplicada, 49/2, 13-30.

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.

Sinclair, J. (1998). Korpustypologie. Ein Klassifikationsrahmen. En W. Teubert (Ed.) Neologie und Korpus (pp. 111-128). Tübingen, Alemania: Günter Narr.

How does violence-motivated online discourse differ from its nonviolent counterpart? Insights from a corpus-assisted discourse analysis

Alfonso Sánchez-Moya

Departamento de Filología y Lingüística Inglesas, Universidad Complutense de Madrid, España | Universidad Nebrija, España

Keywords: *online discourse; corpus-assisted discourse analysis; Intimate Partner Violence; LIWC*

Abstract

It is undeniable that violence in general and Intimate Partner Violence (IPV) in particular can be said to entail worrying challenges in most contemporary societies (WHO, 2013). Despite recent and more visible attempts to better understand this complex social phenomenon are gradually gaining attention, the majority of studies are still related to the health, psychological and legal sciences. Although discourse-motivated research concerned with IPV has recently blossomed (Bou-Franch, 2014; Santaemilia & Maruenda, 2014), there is a significant shortage of attempts to investigate the discourses of key social actors within abusive relationships from the perspective of (critical) discourse studies (AUTHOR, 2018).

Thus, and deeply influenced by the epistemological tenets put forward by the Corpus-Assisted Discourse Studies (Partington et al., 2013), one of the main objectives of this paper is to feed the above-mentioned niche by providing a corpus-assisted discourse analysis of the language employed by women experiencing this type of violence in their lives. To this end, this paper relies on Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), a text-analysis software that has been widely employed to offer interesting insights into discourse types pertaining to personal accounts of intimate and social relationships or health issues (Holmes et al., 2007).

This study is mostly based upon a corpus of 120,000 words collected from a publicly accessible, anonymised, online forum ascribed to a British charity devoted to the improvement of women undergoing IPV. In the attempt to elucidate the specific discursive features that characterise the language used in this IPV-motivated forum, this experimental corpus is contrasted with an *ad hoc* referential corpus compiled from similar digital spaces that are not, however, related to abuse. The discursive output of both corpora, which amounts the same number of words to guarantee a balanced examination, is therefore analysed in the light of some of the most relevant categories provided by LIWC.

By and large, findings are presented according to the main analytical axes developed throughout the study. Thus, LIWC results obtained from both corpora are approached from several perspectives. Nonetheless, for the purposes of this paper, emphasis is placed on the discursive expression of emotionality. Apart from drawing on how both positive and negative emotions are discursively represented in both corpora, the nuanced expression of negativity (namely anxiety, anger and sadness) is also taken into account. All in all, the results obtained from this empirical examination reveal comprehensive accounts of how women survivors of IPV express their concern in contrast to women engaged in different digitally-mediated communicative practices.

References

AUTHOR, A. (2018). Exploring digital discourses on Intimate Partner Violence: A socio-cognitive approach. Amsterdam: ProefschriftMaken.

Bou-Franch, P. (2014). An Introduction to Language Aggression against Women.

Journal of Language Aggression and Conflict, 2(2), 177-181.

Holmes, D., Alpers, G. W., Ismailji, T., Classen, C., Wales, T., Cheasty, V., & Koopman, C. (2007). Cognitive and emotional processing in narratives of women abused by intimate partners. Violence Against Women, 13(11), 1192-1205.

Partington, A., Duguid, A. & Taylor, C. (2013). Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS). Amsterdam: John Benjamins.

Pennebaker, J.W., Boyd, R.L., Jordan, K. & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin.

Santaemilia, J. & Maruenda, S. (2014). The linguistic representation of gender violence in (written) media discourse. Journal of Language Aggression and Conflict, 2(2), 249-273.

World Health Organization. (2013). Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence. Geneva, Switzerland: Department of Reproductive Health and Research, World Health Organization.

El tratamiento de los adverbios oracionales alemanes de significado evaluativo derivados con el sufijo /*weise*/ en la traducción DE<>ES. Un estudio basado en el corpus *PaGeS*

María Teresa Sánchez-Nieto

Facultad de Traducción e Interpretación, Universidad de Valladolid, España

Keywords: *adverbios oracionales; Modalwörter; corpus PaGeS; técnicas de traducción; traducción automática neurona; Appraisal Theory*

Abstract

En este trabajo se centra en la investigación del tratamiento de un grupo de adverbios oracionales alemanes –también conocidos como *Modalwörter* o *Satzadverbien*–, concretamente aquellos formados con el esquema "adjetivo + infijo /*er*/ + sufijo /*weise*/" (p. ej. *dankenswerterweise, dummerweise, unverzeihlicherweise*, etc.) en la traducción del alemán al español y del español al alemán. Estos adverbios oracionales a) no han sido estudiados sistemáticamente desde el punto de vista traductológico en el par de lenguas mencionado; b) no tienen en todos los casos un equivalente léxico directo en español y, por lo tanto, no suelen estar catalogados en repertorios lexicográficos tradicionales; c) contienen significado evaluativo explícito, razón por la cual constituyen un reto tanto para la traducción humana como para la traducción automática.

El objetivo de este trabajo es la descripción de las técnicas empleadas en la traducción de los adverbios oracionales alemanes formados con el sufijo /*weise*/, observando específicamente el tratamiento de su significado evaluativo y prestando atención a los datos de ambas direcciones de traducción (DE>ES y ES>DE). Para ello, nos serviremos

de un corpus secundario de ejemplos (xxxx, 2018) formado por bitextos extraídos del *Parallel Corpus of German and Spanish* (*PaGeS*, https://www.corpuspages.eu), un corpus disponible en línea y de acceso abierto compilado en la Universidad de Santiago de Compostela (Doval et al. 2018). Este corpus contiene testimonios de textos –principalmente literarios– originales alemanes y españoles y sus respectivas traducciones a la otra lengua.

En primer lugar, se expone la problemática específica de los adverbios oracionales objeto de estudio: su atribución por parte de autores como Helbig y Buscha (1998) a una clase de palabras específica diferente de la de los adverbios y su clasificación interna en función de su significado. En segundo lugar, se describe el tipo de significado evaluativo que poseen estos adverbios empleando las herramientas de análisis semántico-discursivo que proporciona la Teoría de la Valoración o *Appraisal Theory*. En tercer lugar, se describe el tratamiento de las palabras modales a) en la traducción humana (a partir de los ejemplos de *PaGeS*) y b) en la traducción automática, a partir de los resultados que motores de traducción automática neuronal como Deepl y Google Translate devuelven de una serie de palabras modales de nuestro corpus en contexto. Para describir la actuación traductora recurrimos tanto a conceptos del marco traductológico conocido como *voices framework* (Alvstad et al. 2017), como a una adaptación a nuestro objeto de estudio del catálogo de técnicas de traducción disponible en Hurtado (2001).

References

Alvstad, Cecilia, Annjo K Greenall, Hanne Jansen y Kristiina Taivalkoski-Shilov. 2017. Textual and Contextual Voices of Translation. Ámsterdam/Filadelfia: Benjamins.

Doval, Irene, Santiago Fernández Lanza, Tomás Jiménez Juliá, Elsa Liste Lamas, y Barbara Lübke. 2018. “Corpus PaGeS: A Multifunctional Resource for Language

Learning, Translation and Cross-Linguistic Research.” En Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications. Ámsterdam: Benjamins (=Studies in Corpus Linguistics; 90).

Helbig, Gerhard, y Joachim Buscha. 1998. Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht. Berlín: Langenscheidt.

Hurtado Albir, Amparo. 2001. Traducción y Traductología: introducción a la traductología. Madrid: Cátedra.

Martin, James R. y Peter R. R. White. 2005. The Language of Evaluation: Appraisal in English. Londres: Palgrave.

(xxxx, 2018)

Creating a Dataset for Domain Bilingual Semantic Annotation based on the USAS Framework

Hugo Sanjurjo-González[a], Rosa Rabadán[b], César Gutiérrez-Pérez[b]

[a]*Department of English, Linguistics and History, University of Huddersfield, UK;*
[b]*Departamento de Lenguas Modernas, Universidad de León, España*

Keywords: *semantic annotation; semi-supervised annotation; USAS; MWE; gold standard dataset*

Abstract

Semantic annotation is crucial for a significant number of linguistic and NLP tasks, including information extraction (Acosta 2011), text mining (Rayson 2010), language learning (Brooke et al. 2015), lexicography and lexicology (Torner and Bernal 2017) or designing controlled natural languages (Davis et al 2009), among others. This paper reports on the construction of a dataset to be used for semi-supervised semantic annotation in English and Spanish in the domain 'Food and Drink' (USAS F1 and F2). The aim is to create an error-free lexicon about Food and Drink that complements Spanish USAS (Jiménez-Yáñez et al. 2017) and a corpus to serve as training set of a future model based on fastText (Bojanowski et al. 2017).

Starting from a small comparable corpus (842,516 words) comprising six subcorpora corresponding to different genres in each of the languages, we extracted frequency lists which were scored manually for domain relevant terms. These were manually annotated after the USAS model and revised by all participants simultaneously to minimize possible inter-coder discrepancies. This initial master lexicon contained nearly 2,000 entries between single items and multiword expressions. MWE were a challenge as neither linguistic nor NLP standard approaches (Sag et al. 2002; Baldwin and Kim 2010; Ramisch 2015; Monti et al. 2018) were adequate to account for cross-linguistic MWE semantic patterns. Unlike recent approaches which focus on i) morphologically defined taxonomies (Escartín et al. 2018), ii) exclusively multiword translation equivalents (Clematide et al. 2018), or iii) retrieval and translation systems (Mendoza et al. 2018), we decided to extract recurrent, domain productive semantic patterns (e.g. En F1/L3+ F2 *orange juice*/ F2 + F1/L3 Sp *zumo de naranja*) including those typically associated to culture-bound, opaque meanings (e.g. Z2 +F1 *Blackforest gateau*/ F1 + Z2 *pastel Selva Negra*; Z5+ Z2 *a la gallega/ Z2 + F1Galician octopus*; Z5 + Z5 + O2 *a la cazuela, al horno*/ F1 *casserole*; F1/O4.6 *roast*, etc.)

The comparable corpus has also been used to train a simple model based on word vectors, more precisely on fastText. This model, together with a custom algorithm, has helped us to guess the annotation of the words that are not contained in the master lexicon. Even using such a small corpus (842,516 words) as a training set our results have gone from 52,63% successful detection if using the original USAS Spanish lexicon, to 70,76% when using the USAS Spanish Lexicon and our master lexicon to an encouraging 81% if using our algorithm. These percentages only take into account specific domain (food and drink) terms.

Adjustments in the algorithm and the model together with additions to the master lexicon will hopefully improve our results making the creation of new multilayer corpora/ the enlargement of existing ones easier and quicker. Additional goals and applications are also put forward. The procedure can be replicated for other specific domains.

References

Acosta, O., Villavicencio, A. & Moreira, V. (2011). Identification and treatment of multiword expressions applied to Information Retrieval. In Proceedings of the

workshop on multiword expressions: From parsing and generation to the real world (pp. 101-109). Stroudsburg, PA, USA: Assoc. for Computational Linguistics.

Baldwin, T. & Kim, S. N. (2010). Multiword expressions. In N. Indurkhya & F. J. Damerau (eds.), Handbook of natural language processing, 2nd ed., (pp.267–292). Boca Raton: CRC Press.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

Clematide, S., Lehner, S., Graën, J. & Volk, M. (2018). A multilingual gold standard for translation spotting of German compounds and their corresponding multiword units in English, French, Italian and Spanish. In R. Mitkov, J. Monti, G. Corpas Pastor, & V. Seretan (eds.), Multiword units in machine translation and translation technology (pp. 126-145). Amsterdam: Benjamins. doi.org/10.1075/cilt.341.06cle

Brooke, J., Hammond, A., Jacob, D., Tsang, V., Hirst, G., & Shein, F. (2015). Building a lexicon of formulaic language for language learners, In Proceedings of the 11th workshop on multiword expressions (pp. 96-104). Denver, Colorado, USA.

Davis B., Dantuluri P., Dragan L., Handschuh S., Cunningham H. (2010). On designing controlled natural languages for semantic annotation. In: N.E. Fuchs (eds) Controlled natural language. International workshop on CNL 2009. Lecture notes in computer science, vol. 5972 (pp. 187-205). Berlin/ Heidelberg: Springer. doi.org/10.1007/978-3-642-14418-9_12

Jiménez-Yáñez, R., Sanjurjo-González, H., Rayson, P., & Piao, S. (2017). Building a Spanish lexicon for corpus analysis. In Proceedings of the 35th edition of the international conference of the Spanish Association of Applied Linguistics – Languages at the crossroads: Training, accreditation and context of use, (pp. 227-339). Jaén: Universidad de Jaén

Parra Escartín, C., Nevado Llopis, A. & Sánchez Martínez, E. (2018). Spanish multiword expressions: Looking for a taxonomy. In M. Sailer & S. Markantonatou (eds.), Multiword expressions: Insights from a multi-lingual perspective (pp. 271–323). Berlin: Language Science Press. doi:10.5281/zenodo.1182605 et al. 2018

Mendoza Rivera, O., Mitkov, R., Corpas Pastor, G. (2018). A flexible framework for collocation retrieval and translation from parallel and comparable corpora. In R. Mitkov, J. Monti, G. Corpas Pastor, & V. Seretan (eds.), Multiword units in machine translation and translation technology (pp. 166-180). Amsterdam: Benjamins. doi.org/10.1075/cilt.341.08riv

Monti, J., Seretan V., Corpas Pastor, G. & Mitkov, R. (2018). Multiword units in machine translation and translation technology. In R. Mitkov, J. Monti, G. Corpas Pastor, & V. Seretan (eds.), Multiword units in machine translation and translation technology (pp. 2-37). Amsterdam: Benjamins. doi.org/10.1075/cilt.341.01mon

Ramisch, C. (2015). Multiword Expressions Acquisition: A Generic and Open Framework. Berlin: Springer.

Rayson, P., Piao, S., Sharoff, S., Evert, S., & Moiron, B. V. (2010). Multiword expressions: hard going or plain sailing. Language Resources and Evaluation, 44(1-2), 1-5. doi:10.1007/s10579-009-9105-0

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (ed). Proceedings of the 3rd international conference on intelligent text processing and computational linguistics (CICLing-2002) (pp. 1–15). Berlin: Springer.

Torner, S. & Bernal, E. (eds.). (2017). Collocations and other lexical combinations in Spanish. Theoretical and applied approaches. London: Routledge.

Design and construction of the Guayaquil radio speech corpus

Brygida Sawicka-Stępińska

Institute of Romance Philosophy, Adam Mickiewicz University of Poznań, Poland

Keywords: *Ecuadorian Spanish; /s/-weakening; Spanish phonetics; Spanish dialectology; acoustic phonetics; mass media language*

Abstract

The variety of Spanish spoken in the linguistic area of the Ecuadorian coast has not been the focus of much attention so far and therefore needs some scrutiny. Even though there are some valuable studies on the Ecuadorian Spanish (e.g. Estrella 2009, Strycharczuk et al. 2014), hardly any of them deal with the Guayas variety, especially with its phonetics. In fact, the principal reference, *El español del Ecuador* by Humberto Toscano Mateus, dates back to 1953. It would not be surprising if not the fact that Guayaquil, the main city of the coast, is currently the most populated urban center in the country, slightly surpassing the capital city of Quito. The two centers are parts of different dialectal zones and the distance between them, both linguistic and non-linguistic, is strongly embedded in the social awareness (Flores Mejía 2014).

In this contribution, the design and structure of the Guayaquil radio speech corpus is discussed. The corpus is the first such a database for this under-resourced variety of Spanish. Its creation is prompted principally by the interest in exploring the linguistic and extralinguistic motivations of the /s/-weakening.

The target population for this study is defined as the radiophonic universe of the city of Guayaquil (López 2003). Eight local radio stations were selected as the most popular, based on official reports (INEC 2010) and social media statistics. The weekly time of transmission was calculated for each of the stations. The weekends have been excluded, since their schedule varies from week to week and contains much amount of material that could not enter the corpus (e.g. live transmissions, programmed music, etc.). The time of transmission was calculated proportionally to obtain the total of 24 hours of recordings, which gives an average of three hours per station. The Ecuadorian law (*LOC*, art. 106) obliges all radio channels to tag transmitted programs as at least one of six formats (informative, opinion, entertainment, formative, sports and publicity). Since the principal variables considered for the present research are the type of contents and speech style, the recording time was distributed proportionally following these criteria.

Other predictor variables planned for the research are phonological context, word length, syllabic accent, local speech rate or lexical frequency (File-Muriel & Brown 2011, Flores 2016). Finally, some circumstantial conclusions are expected to be drawn on the social aspect of the phenomenon (age and gender of the speakers).

The speakers had to follow three criteria: be of at least 18 years old, be native Guayaquil residents and have higher education. 143 speakers have been obtained. All speech utterances pronounced by the speakers have been transcribed orthographically and phonetically using Annotation Pro (Klessa et al. 2013), Praat (Boersma & Weenink 2018) and EasyAlign (Goldman 2011), assured satisfying audio quality (programs with background music and noise –13% of the corpus- have been substituted by other that met similar criteria).

The first results proved that, although the sound quality differed between channels, the acoustic parameters of the phoneme in question are consistent and thus can be used to draw solid conclusions. Since the sampling frame is defined by non-linguistic characteristics (ie. program format), the corpus can be employed also for the analyses of other features, e.g. prosodic, morphosyntactic or pragmatic.

This research has been funded by the National Science Center of Poland (2017/25/N/HS2/00006).

References

Boersma, P., Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.29, retrieved 24 May 2017 from http://www.praat.org/

Estrella Santos, A. (2009). El léxico de Pichincha y Guayas: un estudio comparativo. Quito: Centro de Publicaciones PUCE.

File-Muriel, R. J., & Brown, E. K. (2011). The gradient nature of s-lenition in Caleño Spanish. Language Variation and Change, 23(02), 223–243. https://doi.org/10.1017/S0954394511000056

Flores, T. L. (2016). Velar palatalization in Chilean public speech. Glossa: a journal of general linguistics, 1(6), 1–17. https://doi.org/10.5334/gjgl.105 5. Flores Mejía, E. (2014). Actitudes lingüísticas en Ecuador: Una tradición normativa que subsiste. In A. B. Chiquito & M. Á. Quesada Pacheco (Eds.), Actitudes lingüísticas de los hispanohablantes hacia el idioma español y sus variantes (pp. 409–488). Bergen: BeLLS. Retrieved from http://dx.doi.org/10.15845/bells.v5i0.683.

Goldman, J.-Ph. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. Proceedings of InterSpeech, Firenze.

Instituto Nacional de Estadísticas y Censos (2010a). Fascículo provincial Guayas [http://www.ecuadorencifras.gob.ec/wp-content/descargas/Manu-lateral/Resultadosprovinciales/guayas.pdf].

— (2010b). Fascículo provincial Pichincha [http://www.ecuadorencifras.gob.ec/wpcontent/descargas/Manu-lateral/Resultados-provinciales/guayas.pdf].

Klessa, K., Karpiński, M., Wagner, A. (2013). Annotation Pro – a new software tool for

annotation of linguistic and paralinguistic features. In D. Hirst & B. Bigi (Eds.) Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, 51-54.

LOC (Ley Orgánica de Comunicación), Registro Oficial Suplemento 22, 25/06/2013.

López González, A. M. (2003). El lenguaje radiofónico en la ciudad de Almería. Universidad de Almería.

Strycharczuk, P., Van'T Veer, M., Bruil, M., Linke, K. (2014). Phonetic evidence on phonology – morphosyntax interactions: Sibilant voicing in Quito Spanish. Journal of Linguistics, 50(2), 403452.

Toscano Mateus, H. (1953). El español en el Ecuador. Madrid.

¿Cómo elaboramos un corpus oral conversacional? Problemas y soluciones del procesamiento de la oralidad para su estudio lingüístico

Andreu Sentí[a]; Sandra Montserrat[b]

[a]Departament de Filologia Catalana, University of València; [b]Departament de Filologia Catalana, University of Alacant

***Keywords**: Construcción de corpus; corpus oral; recogida de datos; transcripción; oralidad; variación diatópica y diafásica; conversación*

Abstract

La elaboración de un corpus oral presenta distintas complejidades metodológicas, tecnológicas y lingüísticas (cf. Love *et al.* 2017; Lüdeling & Kytö 2008; Wichmann 2008). A diferencia de los corpus escritos, la oralidad conlleva, entre otros, dos retos para la lingüística de corpus: i) la fuente primaria siempre debe ser el sonido acústico de la voz, pero también necesitamos un código escrito que permita procesar los datos y recuperar la información; ii) normalmente no podemos "recopilar" datos preexistentes, sino que hay que crear la situación para "documentarlos". En esta comunicación nos proponemos reflexionar sobre estos dos temas a partir de las soluciones propuestas en el diseño de un corpus oral del catalán.

El catalán oral tiene un déficit de corpus orales, especialmente de tipo conversacional, coloquial y dialectal (cf. Montserrat & Sentí 2018). El corpus *Parlars* es un proyecto en elaboración que tiene los objetivos siguientes:

1) Documentar el catalán coloquial menos interferido posible (Beltran & Segura 2017).

2) Ofrecer a los investigadores materiales adecuados para la descripción y análisis de la variación lingüística.

En el marco del diseño del corpus *Parlars* nos centraremos en tres aspectos: la metodología de recogida de datos, el procesamiento de la oralidad (transcripción y anotación) y el estudio lingüístico que nos puede ofrecer un corpus oral del catalán.

En primer lugar, explicaremos los diferentes tipos de textos orales que hemos elicitado: la conversacional espontánea secreta, la conversación (semi)espontánea no secreta, la

narración de historia oral monológica y el habla informal mediatizada por ordenador. Nos centraremos en el análisis de las similitudes y diferencias entre la conversación espontánea real (*prototípica*, según Briz 2010) y la conversación (semi)espontánea en la cual interviene un participante-entrevistador. Discutiremos los rasgos de coloquialidad que conserva y las diferencias respecto de la conversación prototípica. Además, mostraremos el resultado de otras estrategias metodológicas para la obtención de conversación (semi)espontánea no secreta (la ausencia del participante-entrevistador, la introducción de temas para la conversación, etc. cf. Basanta 2018).

La transcripción y anotación de los textos la realizamos con la herramienta libre y de código abierto ELAN[9] (Wittenburg *et al.* 2006), ampliamente utilizada para la transcripción de documentos multimedia. Entre otras ventajas, ELAN permite definir múltiples *tiras* de transcripción, permitiendo transcripciones alternativas. A diferencia de modelos de transcripción anteriores (Payrató & Alturo 2002; Briz 2002; Hidalgo & Sanmartín 2005; Bladas 2009), veremos las ventajas de optar por un modelo de doble transcripción por hablante: una transcripción fonoortográfica estrecha más cercana a la señal de habla y una segunda transcripción más ancha que usa ortografía estandarizada. Esto nos permitirá recuperar características dialectales en comparación con el estándar y, a su vez, la codificación estandardizada facilitará la tarea de lematización (semi)automática con Apertium (Forcada et al. 2011) (tokenización, lematización, etiquetación morfosintáctica y alineación) (cf. Ide & Pustejovsky).

Finalmente, trataremos algunas construcciones lingüísticas como ejemplos del aprovechamiento del corpus oral, a partir del trabajo de campo y las conversaciones registradas hasta el momento. Con la intención de acercarnos al análisis de las construcciones en la lengua oral (cf. Taylor 2012, Hilpert 2014, 2018), nos ocuparemos de los verbos modales y otras construcciones o verbos que han asumido valores inferenciales o epistémicos:

(1) ho **havien de** prohibir (Benissa)

(2) **tenien que** treballar (Benissa)

(3) ara no sé si en **quedarà** una o dos, si en queden (Benissa)

(4) all **diu que** està molt vell (Benissa)

(5) **Trobe que** va ser… (Benissa)

References

Basanta, N. (2018). “As formas cambiaron porque o mundo cambiou”: construción discursiva e interseccional de identidades de xénero e sexualidade en conversas sobre ligar. PhD dissertation. Universidade de Santiago de Compostela.

Beltran, V. & Segura-Llopes, C. (2017). Els parlars valencians. València: PUV.

Bladas, O. (2009). Manual de transcripció del discurs oral. Materials de treball. Universitat de Barcelona.

Briz, A. & Grupo Val.Es.Co. (2002). Corpus de conversaciones coloquiales, Anejo de la revista Oralia, Madrid, Arco-Libros.

Briz, A. (2010). El registro como centro de la variedad situacional. Esbozo de la

[9] https://tla.mpi.nl/tools/tla-tools/elan/

propuesta del grupo Val.Es.Co. sobre las variedades diafásicas. In Fonte, I.; Rodríguez Alfano, L. (ed.): Perspectivas dialógicas en estudios del lenguaje (pp. 21-56). México: Universidad Autónoma de Nuevo León.

ELAN (Version 5.2) [Computer software]. (2018, April 04). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tlatools/elan/

Forcada, M. L. & Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., PérezOrtiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G. & Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation.

Machine translation 25.2, 127-144.

Hidalgo, A. & Sanmartín, J. (2005). Los sistemas de trascripción de la lengua hablada, Oralia, 8, 13-36.

Hilpert, M. (2014). Construction Grammar and its Application to English. Edimburgh, Edimburgh University Press.

Hilpert, M. (2018). Construction Grammar and the analysis of spoken language. Presentation at LingCor2018. 1st International Workshop on Spoken Corpus Linguistics. València: Universitat de València.

Ide, N. & Pustejovsky, J. (eds.) (2017). Handbook of Linguistic Annotation. Dordrecht: Springer.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014. International Journal of Corpus Linguistics.

Lüdeling, A., & Kytö, M. (eds.) (2008). Corpus Linguistics: An International Handbook. Berlin: Walter de Gruyter.

Montserrat, S. & Sentí, A. (2017). El corpus Parlars. Elaboración de un corpus oral y coloquial del catalán valenciano. Presentación en el X Congreso Internacional de lingüística de corpus. Cáceres.

Payrató, Ll & Alturo, N. (ed.) (2002). Corpus oral de conversa col·loquial. Materials de treball. Barcelona: Publicacions de la Universitat de Barcelona.

Taylor, J. R. (2012). The mental corpus. Oxford: Oxford University Press.

Wichmann, A. (2008). Speech corpora and spoken corpora. In A. Lüdeling & M. Kytö (Eds.), Corpus Linguistics: An International Handbook (pp. 187–206). Berlin: Walter de Gruyter,

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. &Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.

Accesibilidad, traducción audiovisual y normas en la subtitulación online: EMPAC, EuroparlTV Multimedia Parallel Corpus

Iris Serrat Roozen [a]; José Manuel Martínez Martínez [b]

Department de Filologia Anglesa i Alemanya, Universitat de València;

Department of Language Science and Technology, Universität des Saarlandes, Germany

***Keywords**: accesibilidad; traducción audiovisual; subtitulación; corpus; normas*

Abstract

El Parlamento Europeo tiene como uno de sus objetivos establecer puentes entre los ciudadanos y la institución a partir de diferentes estrategias; entre ellas destaca, por la ambición del proyecto, la creación en 2008 del canal de televisión *online* EuroparlTV[10]. Esta plataforma que distribuye material audiovisual subtitulado en 24 lenguas nos permite por una parte, estudiar la traducción audiovisual en el contexto institucional y por otra, analizar el cumplimiento de los requisitos de accesibilidad de los contenidos audiovisuales establecidos por las Pautas de accesibilidad al contenido web (WCAG) 2.0.

Observar las normas que subyacen a la subtitulación *online* generada por profesionales de la traducción en el ámbito institucional y determinar si están alineadas con el compromiso de acceso a la información por igual para todos los ciudadanos europeos supone un reto considerable. Con el objetivo de realizar un estudio de estas características hemos creado el corpus EMPAC (*EuroparlTV Multimedia Parallel Corpus*) que reúne los subtítulos en inglés y en español de los vídeos emitidos en EuroparlTV entre 2009 y 2017. Para cada video se ha recogido el archivo de los subtítulos, se han anotado los metadatos del texto (título, fecha de publicación, tipo de contenido, URL, etc.), la estructura del documento compuesto por subtítulos y líneas, se ha añadido información lingüística (tokenizado, lematizado y análisis morfosintáctico) obtenida con *TreeTagger*[11] (Schmid, 1994; 1995), se ha segmentado el texto en oraciones con *NLTK* (Bird et al. 2009), y, finalmente, se han alineado los subtítulos entre ambos idiomas (inglés y español). El resultado es un corpus compuesto aproximadamente de cuatro mil textos, más de doscientos mil subtítulos, dos millones y medio de palabras y una duración cercana a las 280 horas para cada lengua. La riqueza de los datos, metadatos y anotaciones con el que lo hemos dotado nos ha permitido analizar distintos parámetros inherentes a la subtitulación.

Es este trabajo describiremos, en primer lugar, el corpus EMPAC. En segundo lugar, compararemos las recomendaciones y guías que tanto desde las WCAG 2.0 como desde el ámbito profesional (*BBC subtitle guidelines*, Norma UNE 153010, *DCMP Captioning Key*) y académico (Karamitrouglou, 2000; Díaz Cintas y Remael, 2007; Pedersen, 2017; Gerber-Morón y Szarkowska, 2018) proponen con el anhelo de estandarizar la práctica de la subtitulación. En tercer lugar y, a partir de estas recomendaciones, planteamos un análisis de los aspectos técnicos de la subtitulación (velocidad de lectura, pausa entre subtítulos, número de caracteres por línea, entre otros) y un análisis lingüístico centrado en la segmentación de los subtítulos, tanto entre las líneas de un mismo subtítulo (segmentación *intrasubtítulo*) como entre un subtítulo y otro que continúa con la misma oración (segmentación *intersubtítulo*).

[10] https://multimedia.europarl.europa.eu/es/home?referer=%22www.europarltv.europa.eu%22

[11] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Para la exploración del corpus y el análisis de la segmentación en particular hemos indexado EMPAC con el *IMS Open Corpus Workbench* (Evert and Hardie, 2011) lo que permite utilizar el procesador de consultas *Corpus Query Processor* (Christ, 1994) a través de la interfaz gráfica *CQPWeb* (Hardie, 2012). Todo ello posibilita la realización de búsquedas complejas aprovechando la información contenida tanto en nuestras etiquetas estructurales (metadatos, subtítulos, líneas y oraciones) como en los atributos posicionales de cada palabra (token, lema y categoría morfosintáctica). Puesto que es la primera vez que se lleva a cabo un estudio tan minucioso de la segmentación de los subtítulos utilizando un corpus electrónico de las dimensiones de EMPAC, mostraremos cómo *CQPweb* nos ha permitido extraer los casos de segmentación incorrecta de forma automatizada en las versiones en inglés y en español del corpus.

Por último, presentaremos los resultados del análisis cuantitativo y cualitativo de los parámetros técnicos de la subtitulación en nuestro corpus, datos que nos permitirán determinar tanto, el grado de accesibilidad de los vídeos del Parlamento Europeo como, describir en qué medida las normas detectadas en nuestro corpus están alineadas con aquellas propuestas desde el ámbito profesional y académico.

References

Aenor. (2012). Norma UNE 153010 Subtitulado para personas sordas y personas con discapacidad auditiva. Subtitulado a través del teletexto. Madrid, Spain. AENOR.

BBC. (2018). BBC subtitle guidelines. London, UK. The British Broadcasting Corporation. Retrieved from http://bbc.github.io/subtitle-guidelines

Bird, S., Loper, E. & Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

DCMP. (2017). DCMP Captioning Key. DMCP. Retrieved from http://www.captioningkey.org/quality_captioning.html

Díaz Cintas, J. & Aline R. (2007). Audiovisual Translation: Subtitling. Manchester, United Kingdom. St Jerome.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In Proceedings of the Corpus Linguistics 2011 conference. Birmingham, UK. University of Birmingham.

Gerber-Morón, O. & Szarkowska, A. (2018). Line breaks in subtitling: an eye tracking study on viewer preferences. Journal of Eye Movement Research 11(3):2.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics 17 (3): 380-409.

Karamitroglou, F. (1998). A proposed set of subtitling standards in Europe, Translation Journal 2:2. Retrieved from https://translationjournal.net/journal/04stndrd.htm

Pedersen, J. (2017). The FAR model: assessing quality in interlingual subtitling. The Journal of Specialised Translation, 28, 210-228.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees.

Proceedings of International Conference on New Methods in Language Processing. Manchester, UK.

World Wide Web Consortium. (2008). Web Accessibility Initiative. Web Content Accessibility Guidelines (WCAG) 2.0. Retrieved from https://www.w3.org/TR/WCAG/

Patterns of booster use in Engineering and Linguistics research articles: a comparative study by booster form

Hanna Skorczynska

Department of Applied Linguistics, Universitat Politècnica de València

***Keywords**: metadiscourse; research articles; boosters; Engineerin; Linguistics*

Abstract

The present study looks into the use of metadiscourse markers known as boosters, and in particular, with regard to their part-of-speech forms (verb, noun, adjective and adverb). Boosting is a communicative strategy for expressing commitment to statements formulated by writers in order to persuade readers (and also editors) of their veracity and validity (Holmes, 1984; Hyland, 1998a; Peacock, 2006). Several studies approached cross-disciplinary variations in the use of boosters (Hyland, 1997; 1998a; 2001; 2004; Hyland & Tse, 2004; Peacock, 2006) and reported differences related to their frequencies and forms. However, no study, to my knowledge, has undertaken a comparative study of specific booster part-of-speech forms. Two corpora of around 650,000 words containing research papers from the fields of Engineering and Linguistics were compiled. The corpora were manually annotated for metadiscourse markers, boosters included, by a group of annotators using Metool software (http://meta-marisa.dinel.org.uk/), designed to that end. A predetermined list was used for annotation, but throughout the annotation process the list was modified to better reflect the metadiscourse patterns in the corpora. The normalized frequencies per 1,000 words revealed similarities and differences in the booster use in the corpora studied. The pattern shared by the corpora is that verbs are the most frequent booster form, followed by adverbs, nouns and adjectives. The main difference is that all booster forms are more frequent in Linguistics than in Engineering with the greatest variation detected for nouns (0.762 per 1,000 words in the former as compared to 0.406 in the latter). The number of items identified for each part-of-speech form in the corpora varied slightly for all booster forms, being higher in Engineering in most of the cases. The most frequent items in each group also showed many similarities, except for adverbs. Regarding the verbs, 16 items were identified in Engineering and 15 in Linguistics with ‘show’, ‘determine’ and ‘demonstrate’ being the most frequent. The range of noun boosters was slightly broader in Engineering (eight) than in Linguistics (five) with ‘evidence’ being the most frequent, followed by ‘conclusion’ and ‘fact’. There were eight adjective boosters used in both corpora with ‘significant’ and ‘clear’ as the most frequent, followed by ‘substantial’ in Engineering and ‘evident’ in Linguistics. With reference to adverbs, 26 items were found in Engineering and 24 in Linguistics, with ‘highly’, ‘generally’ and ‘particularly’ as the most frequent items in the former, and ‘generally’, ‘indeed’ and ‘always’ in the latter. As this is work in progress, the study will

also look into the co-text of selected boosters to provide a more detailed view of their use in the discipline of Linguistics and Engineering. The results obtained allow for a deeper understanding of the rhetorical strategies used in these disciplines and are of interest for academic writing teaching applications.

This study has been conducted within the research project "Identificación y análisis de las estrategias metadiscursivas en artículos científicos en español e inglés" of the Spanish Ministry of Economy, Industry and Competitiveness (FFI2016-77941-P).

References

Holmes, J. (1984). Modifying illocutionary force. Journal of Pragmatics, 8, 345-65.

Hyland, K. (1997). Scientific claims and community values: Articulating an academic culture. Language and Communication, 17(1), 19-31.

Hyland, K. (1998a). Boosting, hedging and the negotiation of academic knowledge. Text 18(3), 349-82.

Hyland, K. (1998b). Persuasion and context: The pragmatics of academic metadiscourse. Journal of Pragmatics, 30, 437-55.

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. Applied linguistics, 25(2), 156-177.

Peacock, M. (2006). A cross-disciplinary comparison of boosting in research articles. Corpora, 1(1), 61-84.

The use of a specialized target corpus for translation training

Francisca Suau-Jiménez; Diana González-Pastor

Departament de Filologia Anglesa i Alemanya-IULMA, Universitat de València

***Keywords**: translators training; corpus resources; specialized texts*

Abstract

Corpus resources have largely been used for translation purposes, including the research, the practice and the training of translators. Research based in corpora offer different and interesting possibilities, with an implication for translation studies, practice and training, based on quantitative, but also on qualitative evidence. As Zanettin (2014: 178) puts it, "the use of corpora in translation has provided a basis for empirical description, involving the comparison of two sub-corpora, either as parallel or comparable texts". Gathering parallel or comparable texts in the form of an electronic corpus has numerous benefits, as they can help to find genre or domain regularities, thus mapping a pattern of common linguistic features that can increase translation speed. A useful contribution is that of Bernardini (2016), who proposes the use of corpora as a learning aid to develop thematic, technological and information mining competences, in order to master competences like translation provision, language and intercultural information.

The training of new professionals in specialized translation demands, among other actions, the compilation and exploitation of *ad hoc* corpora. Students can be trained in the practicalities of corpus compilation, so that they are able to address their changing translation needs (Monzó Nebot, 2008). A particularly interesting use of corpora for translation training is that of searching and compiling a *specialized target corpus* (Bowker, 2000) that comprises a parallel or comparable set of texts of a variable length in one or more domains. This allows the handling of examples of real language use, which in turn provides rich varieties of language patterns, in one or several languages. This practice facilitates the analysis and further mastering of lexis and discourse structures that aid to achieve a pragmatic/adequate translation of texts, so that languageculture nuances can be considered for quality purposes (Nord, 1997; Durán Muñoz, 2011; Suau Jiménez, 2015).

This paper reports a practical experience with corpora used for the training of specialized translation students of English into Spanish where four domains had to be covered (science, technology, economy and finance) at university level. Through a small-scale number of texts that constituted a specialized target, and the didactic perspective focusing on the concepts of genre, register and metadiscourse (SuauJiménez, 2010), students translated a variety of texts featuring various prototypical translation problems associated with lexical and discursive features. A post-final exam survey was collected from students, who showed a general compliance and satisfaction with this approach that provided them with a basic expertise to tackle four possible fields of specialization as translators. Furthermore, results show that the exploitation of such a corpus boosted students' involvement in the management of their own learning process. Implications point towards the adequacy of using *ad hoc* internet-based specialized target corpora as a translation resource when time is a constraining variable in the length of translation courses.

References

Bernardini, S. (2016). "Discovery Learning in the Language-For-Translation Classroom: Corpora as Learning Aids", Cad. Trad., Florianópolis, 36 (1), 14-35.

Bowker, L. (2000). "Towards a Methodology for Exploiting Specialized Target Language Corpora as Translation Resources", International Journal of Corpus Linguistics, 5 (1), 17-52.

Durán Muñoz, I. (2012). "Caracterización de la traducción turística: problemas, dificultades y posibles soluciones", Revista de Lingüística y Lenguas Aplicadas, 7, 103-113.

Nord, C. (1997). Translation as a purposeful activity. Functionalist Approaches Explained. Manchester: St. Jerome Publishing.

Monzó Nebot, E. (2008). "Corpus-based activities in Legal Translator Training", The Interpreter and Translator Trainer, 2(2), 221-252.

Suau-Jiménez, F. (2010). La Traducción Especializada en géneros de economía y empresa en inglés y español. Madrid: Arco Libros.

Suau-Jiménez, F. (2015). "Traducción de calidad de las webs hoteleras: discurso interpersonal e implicación del cliente", Onomázein, 32, 152-170.

Zanettin, F. (2014). Translation-Driven Corpora. Corpus Resources for Descriptive

Applied Translation Studies. Hermeneus: Revista de la Facultad de Traducción e Interpretación de Soria, 16, 371-375

Probabilistic grammars across registers and modes: Pronominal subject expression in some varieties of English

Iván Tamaredo

Departamento de Filoloxía Inglesa e Alemá, Universidade de Santiago de Compostela

Keywords: *register; mode; World Englishes; probabilistic grammar; subject pronoun omission; random forest; cluster analysis; multidimensional scaling*

Abstract

The present paper is located at the crossroads of three different but complementary lines of research. First, it deals with variation within the paradigm of World Englishes (e.g., Schneider, 2007; Mesthrie & Bhatt, 2008), that is, the study of the structural characteristics and the sociohistorical background of varieties of English. Second, it focuses on syntactic (dis)similarities between English dialects as manifested in different registers and modes of production (e.g., Szmrecsanyi & Hinrichs, 2008; Grafmiller, 2014; Seoane, 2017). Finally, it approaches cross-varietal and register/mode variation from the perspective of the probabilistic grammar framework (e.g., Bresnan, 2007; Bresnan et al., 2007; Bresnan & Ford, 2010), which assumes that grammatical knowledge is partially stochastic and that multiple constraints operate simultaneously, sometimes with opposite effects, on the alternation between competing syntactic variants.

Against this backdrop, the aim of the present study is to investigate how – that is, subject to which grammatical constraints – speakers of different varieties of English select between syntactic alternatives across registers and modes. As a case study, we explore variation in the constraints determining the choice between overt and omitted pronominal subjects, exemplified in (1) and (2) respectively, in three varieties of English, namely British (BrE), Indian (IndE), and Singapore English (SgE):

(1) The vision$_i$ was not very clear. **It$_i$** was murky or rather uh foggy or misty. (ICEIND:S1B-006)

(2) Oh, be4 I forget, "Chitra$_i$" sends you her love. **Ø$_i$** Has been asking about you since you left. (ICE-SIN:W1B-003)

More specifically, we address the following research questions: Which grammatical constraints do speakers rely on the most when choosing between the two variants? Which constraints are the most stable ones and which the most variable across registers, modes, and varieties? Can differences be explained on the basis of a core pronoun omission grammar with some unique register, mode, and/or variety patterns or should we talk about independent register-, mode-, and/or variety-specific grammars?

To this purpose, relevant observations of the overt and omitted variants were extracted from the British, Indian, and Singaporean components of the *International Corpus of English* and annotated for several language-internal constraints, namely those identified in previous research on subject omission (e.g., Torres Cacoulos & Travis, 2014; Schröter & Kortmann, 2016), as well as for the register (informal versus formal), mode (spoken versus written), and variety (BrE versus IndE versus SgE) from which each example was taken. Probabilistic differences in the way in which speakers omit subject pronouns across

registers, modes, and varieties were computed by means of random forests and subsequently visualized with the aid of cluster analysis and multidimensional scaling (e.g., Baayen, 2008; Tagliamonte & Baayen, 2012; Levshina, 2015).

Results indicate that (i) the most reliable grammatical constraints are also the most stable ones across registers, modes, and varieties; (ii) speakers of BrE, IndE, and SgE omit pronouns differently in informal language but in the same way in formal registers; and (iii) even though speech and writing differ significantly, no clear variety-specific patterns emerge in any of the two modes of production. Therefore, the present study sheds light on how the effects of language-internal constraints are modulated by external factors, particularly register and mode of production.

References

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using R. Cambridge/New York: Cambridge University Press.

Brenan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherson & W. Sternefeld (Eds.), Roots: Linguistics in search of its evidential base (pp. 77–96). Berlin: Mouton de Gruyter.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. Krämer, & J. Zwarts (Eds.), Cognitive foundations of interpretation (pp. 69–94). Amsterdam: Royal Netherland Academy of Science.

Bresnan, J., & Ford, M. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. Language, 86(1), 168–213.

Grafmiller, J. (2014). Variation in English genitives across modality and genres. English Language and Linguistics, 18(3), 471–496.

Levshina, N. (2015). How to do linguistics with R: Data exploration and statistical analysis. Amsterdam/Philadelphia: John Benjamins.

Mesthrie, R., & Bhatt, R. M. (2008). World Englishes: The study of new linguistic varieties. Cambridge/New York: Cambridge University Press.

Schneider, E. W. (2007). Postcolonial English: Varieties around the world. Cambridge/New York: Cambridge University Press.

Schröter, V., & Kortmann, B. (2016). Pronoun deletion in Hong Kong English and Colloquial Singaporean English. World Englishes, 35(2), 221–241.

Seoane, E. (2017). Modelling morphosyntactic variation in World Englishes from a register perspective. Miscelánea, 55, 109–133.

Szmrecsanyi, B., & Hinrichs, L. (2008). Probabilistic determinants of genitive variation in spoken and written English: A multivariate comparison across time, space, and genres. In Nevalainen, T., Taavitsainen, I., Pahta, P., & Korhonen, M. (Eds.), The dynamics of linguistic variation: Corpus evidence on English past and present (pp. 291–209). Amsterdam: Benjamins.

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests and trees of York English: Was/were variation as a case study for statistical practice. Language Variation and Change, 24(2), 135–178.

Torres Cacoullos, R., & Travis, C. E. (2014). Prosody, priming and particular constructions: The patterning of English first-person singular subject expression in conversation. Journal of Pragmatics, 63, 19–34.

Análisis del léxico de un corpus sociolingüístico

Marcial Terradez

Departament d'Estudis Hispànics, Universitat de València

Keywords: *Sociolingüística; Lingüística basada en corpus; Entrevista semidirigida*

Abstract

La Lingüística del corpus busca, mediante el uso de métodos cuantitativos, sacar a la luz los patrones o recurrencias más reseñables de los corpus que estudia, algo que sin la ayuda de los ordenadores y de la cuantificación sería prácticamente imposible de realizar.

En la investigación de la cual hablaremos en nuestra comunicación, se ha analizado el léxico de un corpus de entrevistas semidirigidas utilizando para ello algunas de las técnicas cuantitativas habitualmente empleadas por la lingüística del corpus para el estudio del vocabulario. Partimos de la base de que, mediante el análisis cuantitativo del léxico de nuestro corpus seremos capaces de sacar a la luz algunos de los rasgos más destacables de nuestros textos orales, y ese es nuestro objetivo.

Como objetivo subsidiario, buscamos comprobar si existe un léxico caracterizador de cada una de las secuencias discursivas de que consta nuestro corpus.

Para llevar a cabo nuestro estudio, en primer lugar se ha realizado una comparación de las frecuencias léxicas de nuestro corpus con las palabras más frecuentes del español general, y para ello utilizamos la lista de las 10000 palabras más frecuentes del Corpus de Referencia del Español Actual (CREA).

Por lo que respecta al método utilizado para el estudio de aquellas palabras representativas de cada secuencia discursiva, hemos operado de una forma similar. Tras la oportuna codificación de nuestro corpus, extraemos las diferentes secuencias de cada una de nuestras conversaciones, y después comparamos las palabras más frecuentes de cada secuencia con nuestro corpus general. En este caso, hemos utilizado el programa TXM (http://textometrie.ens-lyon.fr/spip.php?rubrique96&lang=fr) que presenta como una de sus funcionalidades principales precisamente la comparación de parte de un corpus respecto al corpus general del cual está extraída esa parte.

Este mismo programa es el que se ha empleado a fin de realizar el análisis de correspondencias de las palabras clave de cada secuencia discursiva. Para localizar las semánticas léxicas de un corpus podemos utilizar diferentes métodos, pero consideramos que el análisis de correspondencias es el que nos ofrece un mejor punto de partida para esa identificación de las redes léxicas del corpus.

En nuestro estudio hemos realizado el análisis de correspondencias de dos aspectos: las secuencias discursivas, por una parte, y las palabras clave de cada secuencia según sus asociaciones, por otra.

Por último, para la identificación de las frecuencias de las palabras relacionadas con los centros de interés propios de los estudios de disponibilidad léxica, hemos construido una serie de lexicones o campos léxicos con todas las palabras que aparecen en los estudios

recuentes sobre disponibilidad, y posteriormente se ha procedido a un etiquetado automático de nuestro corpus utilizando la plataforma GATE http://gate.ac.uk/.

Los resultados de nuestra investigación muestran, por una parte, que la entrevista semidirigida se encuentra más cerca de la oralidad que de otro género discursivo, si atendemos a su léxico frecuente. Por otra parte, y en relación con las palabras clave de las secuencias discursivas del corpus, los resultados nos aconsejan replantearnos la habitual distinción en tipologías textuales.

Referencias

ÁVILA, Antonio (1999): Léxico de frecuencia del español hablado en la ciudad de Málaga. Málaga: Universidad de Málaga.

DAVIES, Mark (2006): A frequency dictionary of Spanish. London: Routledge.

GRIES, Stefan Th. & Dagmar DIVJAK (2009): "Behavioral Profiles: A corpus-based to cognitive semantic analysis", in V. EVANS & S. POURCEL (eds.), New Directions in Cognitive Linguistics. Amsterdam / Philadelphia: John Benjamins, págs. 57-75.

KILGARRIFF, Adam (2001): "Comparing Corpora" International Journal of Corpus Linguistics 6.1. Amsterdam: John Benjamins: 97-133.

MCENERY, Tony and Andrew WILSON (1996): Corpus linguistics. Edinburgh: Edinburgh University Press.

SCOTT, Mike. (2001): "Mapping key words to problem and solution", Patterns of Text: in honour of Michael Hoey. Eds. Mike SCOTT and Geoff THOMPSON. Amsterdam: John Benjamins, págs. 109-127.

SINCLAIR, John (1991): Corpus, concordance, collocation. Oxford: Oxford University Press.

Lo oculto en la lengua. El caso del macho mexicano

Justyna Tomczak-Boczko

Institute of Linguistics, Adam Mickiewicz University of Poznan, Poland

Keywords: *colocaciones; Análisis Crítico del Discurso; macho; México*

Abstract

El objetivo es presentar la aplicación de los métodos usados por la escuela del Análisis Crítico del Discurso para analizar las entrevistas de índole etnográfico con el fin de definir el término *macho* en la cultura mexicana. El material por analizar consiste en 20 entrevistas no estructuradas, grabadas en Guadalajara, México.

El principal tema de la investigación es la definición del *macho*. El mayor problema lingíüstico que encontramos es que es un término de baja frecuencia de uso. No obstante, nos atreveríamos decir que es una de las palabras claves de la cultura mexicana (Wierzbicka, 2007). Octavio Paz le atribuía a la sociedad mexicana el cáracter defensivo

y buscaba su origen en la figura de macho "un ser hermético, encerrado en sí mismo, capaz de guardarse y guardar lo que se le confía" (Paz, 1975, 1993, p. 34). Mientras que Samuel Ramos llegó a la conclusión de que "La psicología del mexicano es resultante de las reacciones para ocultar un sentimiento de inferioridad" (Ramos, 1938, p. 76). Para el filósofo la figura central era el 'pelado' quien perseguía el ideal del macho.

Teniendo en cuenta que "a veces lo que no está dicho o escrito es más importante de lo que está. [porque] Un discurso hegemónico puede ser más poderoso cuando ni siquiera tiene que ser invocado, porque simplemente se da por sentado" (Baker, 2006, p. 19), hemos decidido utilizar los métodos del ACD, en particular, las colocaciones – los "gatillos/ desdencadenantes", como los llama Paul Baker, de las que pueden surgir "asociaciones inconscientes que son formas en que los discursos se pueden mantener". (Baker, 2006, p. 114). Además "la colocación es un aspecto de la negociabilidad en el significado: no hay colocación inevitable, sino simplemente consagración a través de la frecuencia de uso" (Williams, 1998, p. 156)

Construimos la red de colocaciones, "marcos de referencia" (ibídem, p. 151), de la palabra macho, en la primera etapa a base de las definiciones que los entrevistados formularon y en la segunda examinaremos el resto de las entrevistas. Los resultados ya obtenidos demuestran que los elementos que los entrevistados evitaban al dar la definición del macho, por ejemplo, la violencia, entran en la red de colocaciones construida a base de las entrevistas (compuestas por sus historias de vida y opiniones sobre los temas determinados a base de la literatura sobre el machsimo). El fin es llegar al discurso oculto, entendido como "prácticas que sistemáticamente forman el objeto del que hablan" (Foucalt, 1972, p. 49, en: Baker, 2006, p. 4).

References

Baker, P., 2006. Using Corpora in Discourse Analysis. Londyn, Nowy Jork: Continuum.

Paz, O., 1975, 1993. El laberinto de la soledad. México: Fondo de Cultura Económica.

Ramos, S., 1938. El perfil del hombre y la cultura en Mexico. Meksyk: Editorial Pedro Robredo.

Wierzbicka, A., 2007. Słowa klucze. Różne języki - różne kultury. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.

Williams, G. C., 1998. Collocational Networks: Interlocking Patterns of Lexis ina Corpus of Plant Biology Research Articles. International Journal of Corpus Linguistics, pp. 151-171.

The study of specialized collocations with the European Central Bank parallel corpus (English-Spanish)

Chelo Vargas-Sierra

Departamento de Filología Inglesa, Universidad de Alicante

Keywords: *contrastive studies; specialized texts; collocations; parallel corpus; co-occurrents; semantically-related combinations*

Abstract

The manual compilation of collocation dictionaries from large corpora is a time consuming and cost-intensive procedure. An automation of this procedure is a burning issue in computational lexicography and in terminology. Thus, the automatic identification of semantic (or conceptual) relations between terms in specialized text corpora has recently gained considerable attention among researchers working in terminology. In the last years, an increasing number of parallel corpora of specific purpose has been developed and made available to the public by multilingual international bodies in a translation memory format. Together with the growth of corpora, computer tools to analyse and exploit them have been improved. Nowadays, relevant terminological information, such as terms, combinations, contexts of use, definitions, etc. is semiautomatically extracted.

This paper describes the most frequent combinations of several groups of semanticallyrelated terms extracted from the European Central Bank corpus (henceforth ECB), a specialized multilingual parallel corpus freely available in Translation Memory eXchange format that contains financial vocabulary. We will also study the specialized collocations that appear in this corpus and the terminological value of specialized combinations as items that convey specialized knowledge. To attain these goals, we analyse contrastively the word sketches of the most frequent nominal terms semantically related and its co-occurrents. Our working hypothesis assumes that when the bases or nodes of a term combination belong to the same semantic field, then it is very likely that these terms will share a high percentage of co-occurrents. We show that, besides the valuable use of parallel corpus freely available in a TMX format, the study of semantically-related terms and co-ocurrents sets the basis for the development and improvement of translation resources.

References

Almela, Moisés. “Improving Corpus-Driven Methods of Semantic Analysis: A Case Study of the Collocational Profile of Incidence.” English Studies 92.1 (2011): 84– 99.

Bergenholtz, Henning, and Sven Tarp, eds. Manual of Specialised Lexicography. Vol. 12. Amsterdam: John Benjamins Publishing Company, 1995. Web. 21 Dec. 2018. Benjamins Translation Library.

Bowker, Lynne. “Towards a Corpus-Based Approach to Terminography.” Terminology 3.1 (1996): 27–52.

Cantos-Gómez, Pascual, and Moisés Almela-Sánchez, eds. Lexical Collocation Analysis. Advances and Applications. Springer, Cham, 2018.

Claveau, Vincent, and Marie-Claude L’Homme. “Discovering Specific Semantic Relationships between Nouns and Verbs in a Specialized French Corpus.” COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology. N.p., 2004. 39–46.

Heid, Ullrich. “On Ways Words Work Together -- Topics in Lexical Combinatorics.” Proceedings of the VIth Euralex International Congress. N.p., 1994. 226–257.

Kilgarriff, Adam, and David Tugwell. “Sketching Words.” Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins. Ed. Marie-

Hélène Conréard. N.p., 2002. 125–136.

L'Homme, Marie-Claude, and Claudine Bertrand. "Specialized Lexical Combinations: Should They Be Described as Collocations or in Terms of Selectional Restrictions?" Procceedings of Euralex 2000. Vol. 31. N.p., 1999. 221–225.

Lorente, M.; Martínez-Salom, A.; Santamaría, I.; Vargas-Sierra, C. "Specialized Collocations in Specialized Dictionaries." Collocations and Other Lexical Combinations in Spanish. Theoretical and Applied Approaches. Ed. E Torner, S.; Bernal. London & New York: Routledge.Routledge. Taylor and Francis Group: 2017. 200–222.

Patiño, Pedro. "A Specialized Parallel Corpus of English and Spanish Free Trade Agreements for the Study of Specialized Collocations." SYNAPS – A Journal of Professional Communication 26 (2011): 85–89. Web. 8 Oct. 2017.

Seretan, Violeta, and Eric Wehrli. "Multilingual Collocation Extraction: Issues and Solutions." Proceedings or COLING/ACL Workshop on Multilingual Language Resources and Interoperability. Sidney: N.p., 2006.

Semantic frames in *SciE-Lex*

Isabel Verdaguer[a]; Natalia Judith Laso[a]; Emilia Castaño[a]

[a]Departament de Llengües Modernes i Literatures i Estudis Anglesos, Universitat de Barcelona, España

Keywords: *Health Science Corpus; biomedical English; SciE-Lex*; *FrameNet; semantic frames*

Abstract

Following the current trend of applying frame semantics to specific registers and the growing body of research on FrameNet-based lexical resources for specialized language (Dolbey 2009 on molecular biology; L'Homme 2008, 2010, 2018, L'Homme and Robichaud´s 2014, L'Homme, M.C., Robichaud, B. and Subirats, C. 2014; L'Homme, Subirats and Robichaud 2016 on the fields of the environment and computing; Reimerink and Faber 2009, Faber 2015, Faber, León-Araúz and Reimerink 2016 on ecology; and Venturi 2013 on legal language) this paper presents a new stage in the development of *SciE-Lex* in terms of frame semantics. *SciE-Lex* (www.ub.edu/grelic/eng/scielex2/scielex.html) is a lexical database of biomedical English, based on the analysis of a corpus of biomedical English, the *Health Science Corpus*, aimed at helping Spanish scientists to write their papers in English, conforming to the conventions of scientific discourse.

SciE-Lex provides phonological, morphological, syntactic, semantic, collocational and phraseological information on the general words most frequently used in biomedical English. In order to highlight the connections between words that have similar syntactic and semantic patterns and to integrate the description of the meaning of individual words into a higher level of lexical organization, we are now introducing frame-based information. To this end we have resorted to *FrameNet*, an online lexical database

grounded in frame semantics, which links the meanings of words with their semantic frames and their syntactic contexts, systematizing the connections between related lexical units. The main objective of this new stage in the development of *SciELex* is to highlight the shared semantic background of the related words in the biomedical register. We also aim at identifying the collocational patterns of the lexical units which have a similar semantic and syntactic behavior, and establishing framebased semantic networks of related lexical units, which will be included in *SciE-Lex* in a later stage.

An initial search for the frames of the verbs in *SciE-Lex* has already been carried out, which allows the user to identify and connect the lexical units that evoke the same frame, and find similarities and subtle differences between them. As FrameNet is based on a corpus of general English, a further exploration (Verdaguer, forthcoming) has pointed to the main aspects that need to be treated in our future research on this specialized register: i) the customization of the ontology of semantic frames to make them conform to the characteristics of biomedical English, ii) careful analyses of frame elements, whose presence or salience may be highly relevant for the meaning of a word, iii) the establishment of intraframes and interframes networks of meaning.

References

Dolbey, A. (2009). BioFrameNet: A FrameNet Extension to the Domain of Molecular Biology. PhD. Dissertation. U.C. Berkeley.

Faber, P. (2015). Frames as a framework for terminology. In H.J. Kockaert and F. Steurs (Eds.), Handbook of Terminology (pp. 1:14-33). John Benjamins Publishing Company.

Faber, P., León-Araúz, P. and Reimerink, A. (2016). EcoLexicon: new features and challenges. In I. Kernerman, I. Kosem Trojina, S. Krek, and L. Trap-Jensen (Eds.), GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference, (pp.73-80). Portorož.

L'Homme, M.C. (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. Traduire, 217, 78–103.

L'Homme, M.C. (2010). Designing terminological dictionaries for learners based on lexical semantics: The representation of Actants. In P. Fuertes-Olivera (Ed.), Specialised Dictionaries for Learners (pp. 141–153). Berlin/New York: De Gruyter.

L'Homme, M.C. (2018). Maintaining the balance between knowledge and the lexicon in terminology: a methodology based on Frame Semantics. In Peters, P., J. G. Yongwei and J. Ding (Eds.), Medical Lexicography and Terminology. Special issue of Lexicography. Journal of Asialex 4(1).

L'Homme, M. C. and Robichaud, B. (2014). Frames and terminology: representing predicative units in the field of the environment. Cognitive Aspects of the Lexicon (CogAlex 2014), Dublin.

L'Homme, M.C., Robichaud, B. and Subirats, C. (2014). Discovering frames in specialized domains. LRec: 1364-1371.

L'Homme, M.C., Subirats, C. and Robichaud (2016). A proposal for combining "general" and specialized frames. CogAlex, 156-165.

Reimerink, A. and Faber, P. (2009). Ecolexicon: A frame-based knowledge base for the environment. In J. Hrebícek, J.H. Mírovsky, W. Pillmann, I Holoubek and T. Bandholtz (Eds.), European conference of the Czech Presidency of the Council of the EU TOWARDS eENVIRONMENT Opportunities of SEIS and SISE: Integrating Environmental Knowledge in Europe (pp 25-27). Brno: Masaryk University.

Venturi, G. (2013). A semantic annotation of Italian legal texts. A FrameNet-based approach. In M. Fried and K. Nikiforidou (Eds.), Advances in Frame Semantics. (pp. 51-84). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Verdaguer, I. (forthcoming). Semantic frames and semantic networks in the Health Science Corpus. Estudios de lingüística del español.

Creating corpus and contrasting the discourse of foot races in València Ciudad de Running

Debra Westall

Departament de Lingüística Aplicada, Universitat Politècnica de València

***Keywords**: Corpus linguistics; Discourse analysis; Recreational races; Running; València Ciudad de Running*

Abstract

Using a corpus linguistics approach, the aim of this paper is to analyze the messages documented during one year of recreational foot races in the city of Valencia. Building on recent literature related to the rise in the popularity of foot races (Abadia et al., 2016; Abadia et al., 2014) and focusing specifically on the transformation of Valencia into a City of Running (Fundación Trinidad Alonso, 2014), I shall first explain the reasons for the popularity of foot races (Bennett et al., 2007; Llopis Goig and Llopis Goig, 2012) and then classify the races held during the 2016 calendar of "València Ciudad de Running". Second, I will describe the methodology used to create a corpus of linguistic expressions related to these events that have made their way into the general language, as well as those displayed during the actual races by the organizers, the participants, and the general public (Morant Marco, 2018). Third, I will describe the findings, focusing on the function of these expressions as well as the way they are exhibited during the five specific races selected for closer analysis: *XII Carrera de la Mujer, 34ª Volta a Peu a València, 4ª Carrera Solidaria Popular Cruz Roja, IV Carrera "En Manada" Bioparc Valencia*, and *I Marxa contra la Violència de Gènere.* Finally, conclusions will be drawn and future research studies proposed.

References

Abadia, S., Medina, F.X., Sánchez, R., Sánchez, J., Bantulà, J., & Morejón, S. (2016). Empresa, deporte y solidaridad: el caso de Oxfam Intermón Trailwalker. Revista

internacional de ciencias del deporte, 44(12), 164-181. http://dx.doi.org/10.5232/ricyde2016.04406 (Consulted 11/11/2018).

Abadia, S., Medina, F. X., Sánchez, R., Bantulà, J., Fornons, D., Bastida, N., Augé, A., Corderas, F., Vega, S. & Pujadas, X. (2014). Entre el boom atlético y la cooperación social. Las carreras solidarias y el ejemplo de la Trailwalker España 2013. Península, IX (I), 105-123. http://dx.doi.org/10.1016/S1870-5766(14)70122-9 (Consulted 11/11/2018)

Bennett, R., Mousley, W., Kitchin, P., & Ali-Choudhury, R. (2007). Motivations for participating in charity-affiliated sporting events. Journal of Customer Behavior, 6 (2), 155-178. Available at http://dx.doi.org/10.1362/147539207X223375

Fundación Trinidad Alfonso (2014). Nace Valencia Ciudad del Running (30 enero, 2014). Available at https://fundaciontrinidadalfonso.org/nace-valencia-ciudaddel-running/ (Consulted 11/11/2018)

Llopis Goig, R. & Llopis Goig, D. (2012). Una tipologia sociocultural de los corredores populares en España. Apunts. Educación Física y Deportes, n.º 108, 2.º trimestre (abril-junio), pp. 9-16. Available at http://www.revista-apunts.com/en/library?article=1536 (Consulted 11/11/2018)

Morant Marco, R. (2018). Personal communication.

How do Chinese students of Spanish use reporting verbs in their master's thesis? A concordance-based analysis

Gang Yao[a]; Biwei Li[b]

[a]Department of English Studies, University of Murcia, Spain [b]Department of English, French and German Philology, University of Oviedo, Spain

***Keywords**: academic writing*; *academic discourse*; *citation practice*; *reporting verbs*; *Spanish*

Abstract

Academic writing is one of the major learning activities for students of higher education.

In fact, many universities have included academic writing courses into their curriculum. As a useful and also must-have skill, reporting theories, concepts, experiments from other research, namely citation practice plays a crucial role in the foundation and development of writer's statements and in the connection and interaction with the academic community. Meanwhile, integrating source text into a claim is complicated and even causes problems for inexpert writers.

By using both quantitative and qualitative methods, previous research has examined citation work in a diverse range of writing settings, such as language backgrounds (native and non-native writers), cross-linguistic comparison (English-Chinese, English-Spanish, English-Czech), group comparison (novice vs. experienced writers), academic genres (research papers, theses), disciplinary variations (medicine, biology, linguistics), and chapter variations (introductions, literature reviews, discussions). Furthermore, some of those studies have built frameworks to categorize reporting verbs according to their

process functions, rhetorical functions, semantic differences, and stance types. However, the vast majority of relevant studies have been carried out in English setting; while reporting practices in Spanish context remain an under-researched area, and even less in Spanish as a foreign language (SFL) context.

The aim of the present study was to illustrate the way in which Chinese students of Spanish used reporting verbs in their master's thesis, by building on the pioneering work of Hyland (2002) and Charles (2006). To this end, we compiled a corpus of 23 theses (over 500,000 words) produced by L1 Chinese MA students and written in Spanish. Concordance was used to search for potential reporting verbs. We then provided a more inclusive coding scheme to categorize them systematically.

The results showed a diverse choice of reporting verbs used by Chinese students of Spanish, although there existed marked individual differences. The most common verbs were *proponer*, *señalar*, *indicar*, *creer*, and *considerar*, while many others only occurred once or twice. Integral sentences (the name of the cited author occurs inside the reporting context) were preponderant reporting structures. It was also found that discourse acts were the most favorable process functions (see Hyland 2002), followed by research acts. Additionally, the SFL student writers tended to choose more positive and conclusive verbs (e.g., *proponer*, *definir*, *afirmar*, *concluir*) when introducing cited sources. Finally, ARGUE verbs had the highest frequency among four semantic categories established by

Charles (2006). Apart from the statistical results, we also noticed that some verbs (such as *indicar*, *confirmar*, *apuntar*, *mostrar*, *demostrar*, etc.) can be classified into different categories depending on their contexts, which may lead to confusion in the theses.

The study, therefore, suggests that reporting verbs in SFL writing courses should be paid more attention. From a pedagogical perspective, SFL writers should be made aware of the functions and differences (may be subtle) among those reporting verbs in order to understand their conventional usage and use them in a proper way. Moreover, the coding scheme we employed in this research could be applied further to develop teaching materials and learning activities for academic writing courses. Lastly, although confined to SFL context, our research paradigm can provide insights into future work on other languages as well.

References

Charles, M. (2006). Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. English for Specific Purposes, 25(3), 310-331. doi: 10.1016/j.esp.2005.05.003.

Hyland, K. (2002). Activity and evaluation: Reporting practices in academic writing. In John Flowerdew (Eds.), Academic discourse (pp. 115-130). London, New York: Routledge.

Annotating Pragmatic Markers in Russian Spoken Speech: the Main Principles and Perspectives

Kristina Zaides

[a]*Russian Language Department, Saint Petersburg State University, Russia*

***Keywords**: pragmatic marker; corpus annotation; corpus data; speech corpus; spontaneous monologue, spontaneous dialogue; spoken speech*

Abstract

Pragmatic markers (PM), which are the subject of the study, are defined as special speech units that serve in spontaneous speech as meta-linguistic and functional tools and do not have any lexical and/or grammatical meaning (Bogdanova-Beglarian, 2014). Through the repeated speech usages, a lexeme or a phrase loses its meaning and grammatical characteristics, turning to be a functional word or a construction which helps a speaker not only to structure the utterance, but also to find a following items, to draw hearer's attention, to correct the previous fragment, to introduce someone else's speech to the utterance, etc. (Bogdanova-Beglarian et al., 2018). The list of pragmatic functions includes approximate, boundary-marking, deictic, hesitative, meta-communicative, reflexive, search, and many other functions (Bogdanova-Beglarian & Filyasova, 2018). The annotation of such markers in the corpus of Russian spoken speech is needed since these units certainly play an important role in speech production and perception.

The main principles of the pragmatic markers annotation concern the manually performed step-by-step process of markers revealing since existing ambiguity of several markers, which have the homonymic full-content analogue (Zaides et al., 2018), e.g., *smotrite* as a meta-communicative PM and a verb with the meaning of perception by the eye:

- *nu Andrej% / togda **vy smotrite** / znachit ya do devyati budu (...) nu (e) telefon vyklyuchu / i otvechat' ne budu / to yest' ya prosnus' gde-to v devyat' s kopeyechkami / budu uzhe (e) min... vy uzhe v eto vremya budete yekhat'* [S123] (during telephone conversation the hearer cannot actually see something at which the speaker points);

- *ryba ogromnaya / nado chtoby uvideli vse // yey! gospoda! **smotrite** / kakaya u menya ryba! a chto yesli () pokazat' yeyo gazetchikam / to to budet furor*! [S1, JUR, description of a picture].

The annotation based on the typology of pragmatic markers, which the authors have already developed. For each PM-group is used special tag, but in spoken speech one marker can have many different functions, e.g., *kak by* serves as the approximator and hesitative marker, and *vot tak vot* is the deictic and meta-communicative marker that also marks the end of the speech production:

- *i ne prosto v vide turista a v vide cheloveka kotoryj / (e-e) v kachestve cheloveka kotoryj tam rabotayet / (e-e) i kak by zhivyot naverno // **vot tak vot*** [S20, MED, story].

Two corpora are annotated for pragmatic markers: the corpus of Russian everyday speech "One Day of Speech" (ORD), which contains all types of everyday spoken interaction

(Asinovsky et al., 2009), and the "Balanced Annotated Text Collection" (SAT), that includes only monological speech in different genres (Bogdanova-Beglarian et al., 2017). The annotation of dialogical and monological texts allows comparing them in pragmatic aspect. Moreover, the presence of different social and psychological groups of speakers recorded for the corpora can reveal the tendency of usage of markers by the particular social or psychological group. For now, the annotation of 15000 words sub-corpus of monologues and 75000 words sub-corpus of dialogues was implemented. As a result, the first observations on the frequency of groups and certain samples of pragmatic markers

in different types of speech are received. The correlation of types of PM used in speech with social and psychological characteristics of speakers is planned to be made.

References

Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., Sherstinova, T. (2009). The ORD speech corpus of Russian everyday communication «One Speaker's Day»: creation principles and annotation. In V. Matoušek, P. Mautner (Eds.), TSD 2009, LNAI, 57292009 (pp. 250-257). Berlin, Heidelberg.

Bogdanova-Beglarian, N.V. (2014). Pragmatemy v ustnoj povsednevnoj rechi:

opredeleniye ponyatia I obshchaya tipologiya [Pragmatems in spoken everyday speech: definition and general typology]. Vestnik Permskogo universiteta. Rossijskaya i zarubezhnaya filologiya [Perm University Herald. Russian and Foreign Philology], 3(27), 7-20. (in Russ.).

Bogdanova-Beglarian, N., Blinova, O., Sherstinova, T., Martynenko, G., Zaides, K. (2018). Pragmatic markers in Russian spoken speech: an experience of systematization and annotation for the improvement of NLP tasks. In S. Balandin et al. (Eds.), Proceedings of the FRUCT'23, Bologna, Italy, 13-16 November 2018 (pp. 69-77). FRUCT Oy, Finland.

Bogdanova-Beglarian, N.V., Filyasova, Yu.A. (2018). Discourse vs. pragmatic markers: a contrastive terminological study. SGEM2018 Vienna ART Conference Proceedings, 5, 123-130.

Bogdanova-Beglarian, N.V., Sherstinova T.Yu., Zaides, K.D. (2017). Korpus "Sbalansirovannaya Annotirovannaya Tekstoteka": metodika mnogourovnevogo analiza russkoj monologicheskoj rechi [Corpus "Balanced Annotated Text Collection": the methodology of multi-level analysis of Russian monological speech]. In D. A. Kocharov, P. A. Skrelin (Eds.), Analiz russkoj razgovornoj rechi (ARZ-2017): Trudy sed'mogo mezhdisciplinarnogo seminara [Analysis of Russian spoken speech (ARZ-2017): VII Interdisciplinary Seminar proceedings] (pp. 813). Saint Petersburg, Russia. (in Russ.).

Zaides, K., Popova, T., Bogdanova-Beglarian, N. (2018). Pragmatic markers in the corpus "One Day of Speech": approaches to the annotation. In Proceedings of Computational Models in Language and Speech Workshop (CMLS 2018) colocated with the 15th TEL International Conference on Computational and Cognitive Linguistics (TEL-2018) (pp. 128-143). Kazan, Russia.