PLENARY TALKS

Andersen, Gisle. “Phraseology in a cross-linguistic perspective: Introducing the diachronic-contrastive corpus method”

The inventory of idiomatic lexico-grammatical patterns in a language is not static, but new patterns of lexical co-occurrence evolve over time, for instance as the result of language-internal lexicalisation or conventionalisation of innovative metaphors, or the result of external influence due to language contact. It is well known that the English language exerts great influence on other languages at the lexical level, as seen from extensive borrowing of terminology and everyday words into many languages of the world (i.e. Anglicisms such as controller, blogg, awesome!, etc.). Although much less studied, it is also clear that the ‘phrasicon’ (Granger 2009) of a language, i.e. its ‘inventory of communicative formulae, catchphrases, slogans and other multi-word items’ (Fiedler 2017: 90) may be similarly affected by such external influence (cf. Andersen 2010, 2017; Fiedler 2014, 2017). This paper represents an early attempt to approach ‘the largely unexplored area of phraseological borrowing’ (Fiedler 2017: 90) in the Norwegian language. The aim is to explore new phraseological units via what I call the ‘diachronic-contrastive corpus method’, exemplified with expressions such as å gå den ekstra milen ‘to go the extra mile’; å gjøre en forskjell ‘to make a difference’, tingene er at ‘the thing is that’, etc.

I first present a typological survey of phraseological borrowing that draws on the work of Granger and Paquot (2008) and, more recently, Fiedler’s (2017) work on phraseological Anglicisms in German. Next, I show how the proposed corpus method can be used efficiently for investigating phraseology across time, for shedding light on the question of whether cross-linguistically parallel structures are the result of borrowing or parallel developments, and – importantly – as a vehicle for rejecting preconceived ideas about a form’s alleged origin (what might be called the ‘Anglicism fallacy’). For this purpose I consult diachronic and synchronic corpora of English (the Old Bailey Corpus, COHA and COHA) and Norwegian (The Norwegian Dictionary Corpus Nynorskcorpuset, the National Library’s Text Archive Bakkylla.no and the Norwegian Newspaper Corpus).

References


Fiedler, Sabine. 2017. Phraseological borrowing from English into German: Cultural and pragmatic implications. In Special Issue on Pragmatic Borrowing. Journal of Pragmatics, 113C.


D’Arcy, Alexandra. “Language history and linguistic corpora. Perspectives on ‘like’ and the like”

Although language change unfolds constantly, the innovation of apparent time (Labov 1963) as an analytic heuristic enabled detailed examination of a single ‘present moment’ that could be extrapolated to reflect diachrony over the short term. Although the emphasis of variationist sociolinguistics is often some current state of a language or variety, the roots of the field are grounded in questions emerging from the study of language in historical perspective. Apparent time is a powerful lens, but it cannot wholly replace the careful study of linguistic variation, structure and usage over the long term. Current states of language do not emerge context free. That is, they represent ongoing and continuous evolution and development, and though diffusion may introduce forms ‘abruptly’ (e.g. through inter-dialect contact), new forms rarely erupt spontaneously within a community. Regardless of their pathway into the language, footprints from earlier stages should be evident. In this talk I explore the ways in which synchronic perspectives can function as distractions from diachronic trends in variable systems. Operationalizing evidence from multiple corpora, both diachronic and synchronic, I argue that the examination of linguistic structure and usage beyond the living speech community is often necessary to tease apart bona fide recent developments from more longitudinal phenomena. On the basis of ‘like’ as a quotative, a discourse marker and a discourse particle, I conclude that when examining apparent synchronic cataclysms, it is necessary to probe the history of a variable system or feature, else risk falling for the Recency Illusion, the belief that uses noticed (or examined) in synchronic perspective are in fact recent (Zwicky 2005)

Davies, Mark. “New tools for research and teaching: The Corpus del Español and WordAndPhrase (Spanish), the NOW "monitor" corpora (English / Spanish), and the new iWeb corpus”

In this presentation, I’ll focus on four new corpora / resources from BYU, which have been released in the last two years. In each case, I’ll give a number of examples of how the corpora can be used for research and teaching, in ways that are probably not possible with other resources.

1. The Corpus del Español (Web / Dialects) contains two billion words from the Web (from 2013-2014). It allows users to compare the frequency of words, phrases, grammatical constructions, and collocates across 20 different Spanish-speaking countries, and it allows many types of searches that cannot be done with CREA or CORPES (Real Academia Española).

2. WordAndPhrase (Spanish). Users can browse through the top 40,000 words (lemmas) from the Corpus del Español (#1 above). For each word, they can see on one page a brief definition and translation to English, frequency information, synonyms, collocates, related topics (in many cases, more useful than collocates), and sample concordance lines. Users can also enter entire texts and see the keywords from the text, and then click on any word to see the information mentioned above.

3. NOW:English (News on the Web) contains 6 billion words of data from 2010 to the present time. Most importantly, every day 4-5 million words of data are added to the corpus, which allows researchers to look at changes in real time – the data is never more than 24 hours out of date. In August/September 2018, a similar corpus will be released for Spanish – more than 6 billion words of data from 2012, and updated (like NOW:English) every day.

4. iWeb (to be released in May 2018) contains 14 billion words of data from the Web, which makes it about 25 times as big as COCA (560 million words). Unlike other large corpora, it was created in a way that allows users to quickly and easily focus on particular topics and domains. Most importantly, it allows teachers, learners, and researchers to browse through the top 60,000 words (lemmas) to see a wealth of information from the corpus for each word (in ways not yet possible with the other BYU corpora).
Corpus linguistic methods open new avenues for our understanding of literary language. Carter (2004) has argued that the notion of a clear-cut distinction between literary and non-literary language is an unhelpful one. Instead, language use is best described in terms of a cline of literariness. In this paper, I will illustrate corpus linguistic methods that offer new ways of exploring the fuzzy boundaries between literary and non-literary patterns of language use. Focusing on the textual creation of fictional characters, I will look at patterns that account for the presentation of speech in fiction and the description of body language. The CLiC web application (http://clic.bham.ac.uk/) - that has been specifically designed for the study of literary texts - will be used to show how such patterns can be identified and analysed. The most recent version of the CLiC corpora now comprise more than 130 books across four subcorpora: the corpus of Dickens’s Novels, the 19th Century Reference Corpus (19C), the Corpus of 19th Century Children’s Literature (ChiLit) and the Corpus of Additional Requested Texts (ArTs). For all CLiC texts, direct speech and specific places around speech have been marked up (Mahlberg et al. 2016) so that CLiC can run searches across defined textual subsets. The ability to focus on subsets makes it possible to identify specific literary patterns that can then be usefully compared to non-literary language. To illustrate key functionalities of CLiC I will provide examples retrieved with the most recent version of the CLiC KWIC grouper as well as examples where manually added tags support the analysis of concordances lines.

References
Studies of varieties of English have shown that verb complementation, which is a core area of lexicogrammar, is often considered to differ regionally (Kjellmer 1985, Algeo 1988, Mair 1995, 2002, Mukherjee & Hoffmann 2006, Mukherjee & Schilk 2008) and may thus be a gainful beginning when depicting the emergence and development of New Englishes (cf. Schneider 2007:46). So far, studies of New Englishes based on corpus approaches have shown that verb complementation has been largely under-researched as a characteristic field of variation (Mukherjee & Hoffmann 2006). Most of the studies have concentrated primarily on the study of ditransitive verbs and transfer-causative motion verbs TCM (Olvarría de Ersson and Shaw 2003; Mukherjee and Hoffmann 2006; Mukherjee & Schilk 2008, Schilk et al. 2012 among others).

Other recent studies have focused on the variation between infinitival and gerundial complementation constructions (Bernaich 2015, Deshors 2015; Deshors & Gries 2016). This considered, my present work will contribute to the field of clausal verb complementation in World Englishes by studying the variation noticed in the sentential complementation including both finite and non-finite complement clauses (CCs) of what Quirk et al. (1985:1181) refer to as “factual verbs”, namely hope and decide in British English (as a reference corpus) and Singapore English (as a nativized L2 variety of English). The following objectives are pursued: 1) to analyze the distribution of finite (that-clause, zero complement clauses) and non-finite (to-infinitive) in the mentioned varieties of English; 2) to identify any existing divergence and/or convergence in the complementation clauses of hope and decide in both varieties; 3) to offer an analysis of some of the potential language internal factors influencing the choice of complement clauses after the stated verbs. These factors include subject denotation (co-referentiality between the main clause and the CC subjects), type and animacy of subject in CC, presence or absence of a modal verb in CC, presence or absence of a negative marker in CC, the voice in the CC (active/passive/copular), presence of intervening material in words between the complement taking predicate hope and decide and the CC (cf. Cuyckens & D’hoedt 2015). By analyzing the previously mentioned factors, the causes of the variation noticed are explained. The findings will be subsequently examined in light of an extensively used theory, namely Schneider’s (2007) dynamic model of postcolonial Englishes. The data are retrieved from the ICE corpora (See Greenbaum, 1991) which makes it possible to study and compare finite and non-finite complementation patterns of the verbs under review in both varieties. After manual pruning of irrelevant tokens, all corpus attestations are manually coded against seven grammatical factors and then entered into SPSS software package for statistical analyses.

The study shows a tendency for both varieties to favor non-finite constructions with both verbs, when the subject of the complement clause is identical to the subject of the main clause. However, the data also demonstrate that increasing preference for finite complementation clauses over non-finite constructions with the verb hope is due to the different denotation of the subject, where variation is only possible with finite complementation clauses. More results will be discussed in detail in the presentation.

References

elements of schematic text structure), *endophoric markers* (i.e., references to other parts of the text), *evidentials* (i.e., hearsay or attribution to reliable sources of information), and *code glosses* (i.e., expressions which supply additional information by rephrasing, explaining, or elaborating what has been said). In the methodology set for this study, the analysis was based on a corpus of authentic data consisting of the transcripts of those speeches and debates involving the leaders of at least the two parties topping opinion polls in each country and election (i.e. the PP and the PSOE [also Podemos in the 2016 election]) in Spain and the Democratic and Republican parties in the United States. Furthermore, the quantitative use of interactive resources was analyzed with the tool Metool developed specifically to detect metadiscourse strategies. Overall, the results demonstrate that the American political candidates make more use of interactive strategies than their Spanish counterparts. Significant differences were also found in the distribution and composition of such resources, specifically in the case of certain subcategories (i.e., transition markers, evidentials, and code glosses). Regarding the persuasive effect of these strategies, the particular linguistic/cultural contexts seem to influence the politicians’ rhetorical choices when delivering their speeches, as well as the personality of each political actor (Johnstone, 1999), and the message developed around his or her political persona (Lempert and Silverstein, 2012). The different results in both sub-corpora suggest that the use and distribution of interactive resources in political communication is not only conditioned by the candidate’s perceived public persona but also by the specific cultural context in which the campaign event takes place.

**References**


27. *Aldama García, Nuria. “Se” in spoken and written texts”*

*Se* is a high frequency word in Spanish that appears in a variety of structures acquiring different meanings. *Se* can be understood as a continuum having at one end pronominal functions such as *se* as direct or indirect objects, *se* as *le* or reciprocal and reflexive constructions. At the other end, *se* works as a syntactic mark indicating predicate valency reduction (anticausative, passive, middle and impersonal) (Moreno et al 1999). What anticausative, passive, middle and impersonal se constructions have in common is the omission of the agentive argument (Moreno Cabrera 1991). Somewhere in between those ends, *se* can be part of pronominal predicates or it can be interpreted as an anchoring or dative of interest, that is, as an emphatic and colloquial mark that does not necessarily establish a semantic relation with the verb (Bello 1847).

The aim of this paper is to estimate the usage distribution of *se* in written and spoken texts. To do so, 1000 sentences containing *se* were manually analyzed. There were two groups of sentences: the first one contained the first 500 sentences of C-ORAL-ROM (Moreno et al 2005) which is a corpus of Spanish spontaneous spoken texts. The second group contained the first 500 sentences of CORPES XXI which contains Spanish academic written texts (RAE 2016).

The 1000 sentences were annotated and classified under six tags: *ethic dative*, *pronominal predicate*, *direct object*, *indirect object*, *syntactic mark* (structures of predicate valency reduction) or disfluence (repetitions, incomplete sentences, reformulations that cannot be tagged within any of the categories mentioned above). The results show that the ethic dative and indirect object uses of *se* are remarkably more common in spoken than in written texts, whereas the direct object and valency reduction syntactic mark uses of *se* are remarkably more common in written texts. There are two conclusions that can be drawn from this search: first, academic written texts make use of *se* as a syntactic mark to avoid agency and possibly to present facts in a distant way. Second, the ethic dative use of *se* is more common in spoken spontaneous language because it helps speakers emphasize their discourse. Third, *se* as indirect object is more common in spontaneous spoken texts. Most of its appearances correspond to the sequence *se*-direct object-verb. Last but not least, *se* as direct object is more frequent in academic written texts due to predicate reflexive uses.

**References**


The two main problems that constituent grammars must face regarding parsing applications are capturing free word order and the computational cost of theoretical formalisms. These two factors have motivated the use of dependency grammars as an alternative to constituent grammars when it comes to parsing applications: dependency grammars capture word order through semantic and syntactic head-dependent relations without adding extra computational cost.

The UAM Spanish Treebank is a phrase-based corpus of 1.500 syntactically annotated newspaper sentences taken from El País Digital and Compra Maestra. The annotation format of the UAM Spanish Treebank combines a Part of Speech Level (POS) with the feature values specifying syntactic information for each element. There are four levels of annotation: syntactic categories, syntactic functions, syntactic features and some semantic features. The morphological analysis of the words was carried out with GRAMPAL (Moreno Sandoval 1993), which is a free access morphological tagger created by the LLI-UAM. To avoid redundancy, the
information is only specified in the most relevant layers although it might cause underspecification due to lack of syntactic detail. To reflect surface syntax, the corpus only includes two kinds of empty categories: null subjects and elided material. The annotation schema follows a simple, vertical intended format.

Transforming the UAM Spanish Treebank into a dependency treebank leads to a better understanding and study of Text Analytics and Natural Language Processing (NLP) in Spanish because dependency relations can capture and reflect Spanish free word order without increasing computational costs.

The aim of this paper is to explain the automatic transformation process of the UAM Spanish Treebank using Universal Dependencies annotation guidelines. Throughout the transformation process certain challenging issues have arisen: Spanish se or post verbal subjects constitute Spanish phenomena lacking an English counterpart.

The results of the automatic transformation procedure yield that, out of the 24836 words and punctuation marks contained in the treebank, 24842 were automatically tagged. The number of automatically assigned relations is 12149, that is 48.9% of the total number of relations. After a manual review, some errors concerning tagging and relation assignment were observed and corrected by hand.

References
Manzella, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard & David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. Association for Computational Linguistics (ACL) System Demonstrations.


119. Almela, Moisés. “Exploring the boundaries of complex collocations” The objective of this paper is to sketch out a methodology for determining the syntagmatic extension of lexical collocations. For over three decades the traditional focus of collocational research had been laid on word pairs and binary relations – for a critical review, see Evert (2009), Seretan (2011), and Gries (2013). However, in recent years the field has been shifting its interest to the study of more complex patterns of collocation, consisting of three or more lexical items (Alonso Ramos & Wanner 2007; Baker 2016; Brezina et al. 2015; Mollet et al. 2011; Nerima et al. 2010). This new line of inquiry has raised new questions. One of the problems that remains unresolved is how to develop criteria for determining the external boundaries of collocational patterns. If the maximum level of internal complexity is to be determined empirically for each collocational pattern, then specific techniques must be developed for the purpose. Daudaravičius and Marcinkevičiene (2004) have addressed a similar problem in relation to the extraction of collocational chains, i.e. “frequent uninterrupted strings of word-forms” (p. 322), such as just a few, for a long time, to a large extent, etc. In the present study I will analyze the boundaries of a different type of collocational patterns: those formed by lexical associations of three or more lemmas (e.g. PAY – CLOSE – ATTENTION). These patterns can be realized by more than one sequence of word-forms (e.g. close attention being paid to..., they paid close attention to, etc.).

The methodology proposed here is based on a specific corpus-linguistic application of Shannon entropy. Mason (2000) used this measure to obtain the profile of lexical gravity of a word, which he employed for calculating the optimal textual distance for extracting collocates. The strategy used in this study is an adaptation of Mason’s proposal, as it is based on comparisons of the level of entropy in neighbouring slots. The premise is that an extension of the number of elements in a syntagmatic chain accompanied by an increase in the level of entropy can be interpreted as a possible boundary point, and that a decrease in the level of entropy can indicate that the words in question are members of the same collocational pattern. The data will be extracted from the British National Corpus using Sketch Engine query tools.

References
Gries, Stefan Th. (2013). 50-something years of work on collocation: What is or should be next... International Journal of Corpus Linguistics 18 (1), 137–165.
120. Amador Moreno, Carolina P. “The intemperate habits of the people here, form my principal objection to this place: Empathetic deixis in the context of Irish emigration”

The expression of emotional attachment and psychological closeness is part of what is known as empathetic deixis, a term suggested by Lyons (1977: 677) to refer to what Lakoff (1974: 353) had called emotional deixis, where there is often at least an ‘intuitively evident connexion between physical and emotional proximity or remoteness’ (Lyons 1995: 311). The study of empathetic deixis has received little attention (see Rühlemann 2007: 222), and the context of emigrant letters, which represent a form of ‘intimate discourse’ (Clancy 2016) and where the sense of affective involvement in the narration of spaces appears so salient, seems like a particularly relevant place to investigate it.

This paper analyses the use of this, that, here and there in the Corpus of Irish English Correspondence (CORIECOR). CORIECOR is a corpus of personal letters which covers the timespan from 1750-1940. The corpus contains some 6000 texts (approx. 5 million words), of which 4100 (2.5m words) are correspondence maintained between Irish emigrants and their relatives, friends and contacts. 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 in IrE. The paper will discuss how, apart from pertaining to the spatial discourse of migration, deictic items like this, that, here and there have an empathetic or emotional meaning.

The study combines a number of different techniques using the tools of Corpus Linguistics to investigate the use of deictic forms in the letters. Quantitative and qualitative data analyses are employed in order to explore the notion of deixis in the corpus. Empirical research based on corpus data aims to potentially advance the already existing knowledge on this intriguing type of deixis, as suggested by Rühlemann (2007: 222). However, bearing in mind that from a theoretical standpoint, the frequency with which speakers produce certain linguistic features in a particular register ‘is likely to be a poor indicator’ (Evans Wagner et al. 2016: 224) of that feature’s stylistic and pragmatic functions, qualitative methods of analysis are also employed. The paper will discuss how, when dealing with pragmatic features like these, gross frequency counts in large-scale corpora would provide very little information on what speakers (letter authors in this case) are doing with a particular deictic form in different registers. For that purpose, qualitative data analyses based on smaller datasets would appear to be a more appropriate way to study pragmatic meaning than a large cross-corpora comparison.

References

78. Ávila Ledesma, Nancy E. “I thought you certainly had forgotten me by this time’: Exploring ‘I think’ and ‘I know’ in Irish emigrants’ letters using corpus pragmatics”

In her seminal book, Emotion Talk Across Corpora, Bednarek (2008:2) highlights the validity and potential of corpus linguistic methods ‘to examine the use of emotion talk in naturally occurring discourse’. Given their closeness to speech, personal letters have been extensively used to trace the emergence and development of linguistic features using corpus linguistics techniques (cf. e.g. Elspaß 2002; Dossena 2007; McLelland 2007; Moreton 2012; Amador-Moreno 2016). The study of emotion talk in the epistolary discourse from a corpus linguistics perspective, however, has long been a neglected area. Adopting a corpus-pragmatic methodology (Romero-Trillo 2008), this study sets out to investigate the conceptualisation of emotions and intersubjectivity in the personal correspondence exchanged between Irish emigrants to United States, Australia and New Zealand and their loved ones back in Ireland between 1840 and 1930. More specifically, the study proposes a corpus-pragmatic examination of projection structures such as I think and I know in an attempt to explain the function of such linguistic features in the construal of emotions in the epistolary discourse.

The letters used in this study – the USA and the Australia/New Zealand collections – were extracted from CORIECOR, the Corpus of Irish English Correspondence (McCafferty and Amador-Moreno in preparation) and searched using Sketch Engine (Kilgarriff et al. 2004). Corpus tools such as frequency lists, word sketch and concordance lines will be utilised to identify the most common collocations and linguistic uses of I think and I know. Secondly, a qualitative examination of the structures under study will be conducted in order to unveil the communicative functions embedded in the projection clauses and how they help to maintain emotional and physical links between emigrants and their families. Preliminary results highlight the value of personal letters to analyse emigrants’ emotional worlds and show that there is a significant difference regarding emotional responses between emigrants to USA and Australia and New Zealand – triggered by different migration experiences in the host countries.

References
Los programas de telerrealidad como corpus para el estudio de la interacción en español y alemán

93. Bertomeu Pi, Pau. “Los programas de telerrealidad como corpus para el estudio de la interacción en español y alemán”

El objetivo del presente trabajo es ahondar en el papel de los programas de telerrealidad como fuente legítima para la elaboración de corpus orales, y destacar su potencial para el análisis (inter)lingüístico de la interacción cara a cara. Concretamente, partimos de grabaciones de “Gran Hermano” en sus versiones española y alemana, en las cuales se observan situaciones comunicativas propias de la cotidianeidad.

En los últimos años se han realizado grandes avances en la elaboración de corpus escritos, orales y multimodales, por lo que en la actualidad es posible analizar con minuciosidad y rigor, por una parte, un espectro muy amplio de situaciones en el medio gráfico o fónico con concepción hablada o escrita (Söll 1985) y, por otra, aspectos muy concretos de la comunicación, como los gestos (Baldry y Thibault 2006). En el ámbito de la conversación cara a cara, destacan los trabajos realizados en el seno del grupo Val.ES.Co. (Grupo Val.ES.Co. 2014) en español, y del grupo FOLK (Schmidt 2014) en alemán. Mientras que en el primer caso el método de obtención de datos es principalmente la grabación secreta, en el segundo también es habitual el trabajo con vídeos.

En nuestro caso, el corpus que presentamos está conformado por conversaciones que reflejan la interacción diaria en contextos de cocina y reparto de tareas. Aunque son muchas las posibilidades de análisis que ofrece el corpus, el material resulta especialmente útil para el estudio de los actos directivos, como las peticiones o sugerencias, en la conversación coloquial.

En cuanto al proceso de creación del corpus, hemos recurrido al programa de transcripción ELAN y desarrollado nuestro propio sistema de etiquetas con conceptos clave de la Pragmática y el uso de gestos. Los momentos corresponden a diferentes funciones (e.g. para provocar argumentos). Future work will extend the annotation to a larger sample including TED-Talks transcripts to check whether the preliminary tendencies investigated in this study also apply to this genre.

References

Los préstamos y los extranjerismos han estado siempre presentes en la historia del español y constituyen uno de los principales mecanismos de ampliación del caudal léxico del español. Asimismo, los préstamos son la consecuencia más visible del contacto entre las lenguas y revelan la historia de las relaciones culturales entre los pueblos que las hablan. Con la globalización y los avances informáticos, se ha producido un mayor contacto entre las lenguas, lo que ha traído consigo una gran importancia de los préstamos y experiencias de otras lenguas al español, principalmente de aquellos que provienen de la lengua inglesa. Pero, ¿de qué forma se registran estas experiencias en el repertorio lexicográfico?

Esta investigación versa sobre la integración de los extranjerismos ingleses en los diccionarios españoles, especialmente en el Diccionario académico, con la intención de estudiar qué tipo de extranjerismos anglosajones incluye y de qué forma los integra la macroestructura. Para ello, partimos de la elaboración de un corpus en el que recogemos los extranjerismos de origen anglosajón en la última edición del Diccionario de la Lengua Española (2014). De este modo, nos centramos en estudiar la forma en la que la Academia integra estas voces en su repertorio lexicográfico, así como los cambios que se producen respecto a ediciones anteriores. Al mismo tiempo, analizamos la necesidad de introducir estos términos en la lengua española; es decir, si se trata de extranjerismos necesarios o si, por el contrario, son superfluos. Así pues, observamos si se cumplen los criterios que establece el Diccionario panhispánico de dudas (2005) sobre los extranjerismos, esto es, “que su incorporación responda en lo posible a nuevas necesidades expresivas y, sobre todo, que se haga de forma ordenada y unitaria acomodándolos al máximo a los rasgos gráficos y morfológicos propios del español”.

De este estudio, obtenemos varios resultados reseñables. En primer lugar, observamos que la presencia de extranjerismos anglosajones en el Diccionario académico es importante y que el número de voces inglesas en el DLE ha aumentado en los últimos tiempos, lo que muestra la influencia del inglés en nuestra sociedad y el mayor contacto que se ha producido en los últimos años con la lengua anglosajona. Por otro lado, comprobamos que, a pesar de las indicaciones del Diccionario panhispánico, el DLE sí incorpora extranjerismos superfluos, tanto adaptados como sin adaptar, incluyendo algunos con poco uso en el español actual y cuyo equivalente español posee gran vitalidad, como ocurre, por ejemplo, con rouge o tournée. Por último, vemos que es frecuente que en ocasiones, a pesar de que ciertos extranjerismos presenten ya una forma adaptada, el Diccionario académico opte por incluir tanto la variante adaptada como el extranjerismo o préstamo crudo debido a la frecuencia de uso que este último presenta, hecho que queda patente tras la consulta del Corpus XXI. No obstante, en estos casos, el DEL hace constar la preferencia de uso del préstamo adaptado, ofreciendo en él la definición por remisión.

58. Carrió Pastor, Mª Luisa. “La atenuación en el lenguaje académico: Un análisis comparativo entre la lengua española y la inglesa”

Los marcadores discursivos de atenuación son los que indican la decisión del escritor de poder reconocer diferentes proposiciones sobre un hecho (Hyland, 2005), mitigando las afirmaciones. Podemos también afirmar que es un mecanismo retórico destinado a regular la interacción entre los participantes (Briz, 2001). El uso de la atenuación permite a los escritores comunicarse con sus lectores de una forma que no indica compromiso sobre sus propuestas, lo que les facilita la aceptación en una comunidad discursiva y la posibilidad de dialogar con la audiencia de forma eficaz (Hyland, 1998). Algunos estudios del español coloquial conversacional (Albelda y Briz, 2010) también incorporan la atenuación como parte del registro coloquial. Este estudio se centra en el análisis de las estrategias de atenuación que se usan en el lenguaje académico. Los objetivos son, por un lado, investigar, analizar las estrategias de atenuación utilizadas en español e inglés para determinar pautas y patrones en ambas lenguas, comparándolas. Por otro lado, se persigue un objetivo docente, ya que los resultados de este análisis nos permitirán conocer en profundidad las estrategias de atenuación y utilizarlas comparaciones entre las dos lenguas con la lengua inglesa. La hipótesis de partida de este análisis es que las estrategias de atenuación son distintas en las dos lenguas, ya que la lengua inglesa utiliza frecuentemente el lenguaje académico mientras que no son utilizadas frecuentemente en español. Con el fin de obtener los objetivos que se proponen, treinta artículos científicos de ingeniería escritos en inglés y treinta artículos científicos de ingeniería escritos en español se seleccionaron para poder determinar el uso de las estrategias de atenuación. Una vez seleccionados, identificó las herramienta WordSmith Tool 5.0 los marcadores de atenuación en inglés y español basándose en las propiedades de los marcadores de atenuación en inglés y español en los estudios de Hyland (2005) y Albelda y Briz (2010). Una vez extraídas las frecuencias, se normalizaron por 10,000 palabras, comparándose el uso de las estrategias de atenuación en español y en inglés. A continuación, se analizaron algunos ejemplos de los casos más frecuentes para identificar sus patrones y se identificaron las estructuras más frecuentes en las que se encontraron los...
References

13. Crespo, Begoña. “Women's presence in scientific texts and prefatory material: Keeping track of persuasion”
In this work it is my aim to offer a description of the way in which women used linguistic devices denoting persuasion in samples that represent late Modern scientific writing, on the one hand, that is to say, samples from the Coruña Corpus of English Scientific Writing, and introductory material to these works, such as prefaces or dedications, on the other. Whilst scientific writing is considered to be presented in a neutral tone indicating objectivity and rigourousness (what in Lakoff’s words is known as “considerate styles”), prefaces can be said to evince a somewhat more intimate and persuasive tone (“high-involvement styles”) as, in principle, the stuctural rhetoric of these texts allows for a direct relationship between author and target readership. The features denoting persuasion, argumentation and interaction that will be analysed include predictive and necessity modals, conditional subordination and suasive verbs, infinitives, and split auxiliaries, considered as such by Biber 1988, Prelli, 19089 or Mischke, 2009, among others. Several variables will be taken into account in the present study: the diachronic variable (eighteenth, nineteenth centuries), the communicative format variable, also known as genre (treatise, textbook, catalogue...) and, finally, the scientific field or discipline variable (astronomy, history, life sciences, chemistry and philosophy).
This analysis will allow me to unveil how frequently and under what discursive circumstances female writers were present in their works as well as the kind of linguistic devices they preferred and how this could be interpreted in terms of writing style. Hopefully, the data found will shed some light on the type of social role prevailing for these writers: that of scientists or that of eighteenth- or nineteenth-century women living in an androcentric society.

References

96. Diaz Sierra, Sara. “Towards the compilation of a specialised corpus for the study of the literary representation of Northern Irish accent”
The variety of English spoken in the linguistic area of Northern Ireland –a term taken here to refer to the region that covers the historical province of Ulster– has not been the focus of much attention in the past and therefore needs some scrutiny. Even though there are some diverse and valuable studies on Northern Irish English (henceforth NIE) (Adams 1948; Milroy and Milroy 1978; Harris 1985; Corrigan 1990; Filippula 1995; Kirk 1997; Hickey 2002; McCafferty 2007), only a few of them deal with literary representations of this variety. In fact, the only literary works by Northern Irish authors that have been analysed more in depth are those by William Carleton and Patrick MacGill (McCafferty, 2010, p. 8). Although Kirk said that some years ago and several corpora representing the dialect of English spoken in Ireland have been developed since then such as the Limerick Corpus of Irish English (LCIE), ICE-Ireland and Hickey’s Corpus of Irish English, there is still no corpus of literary works produced in Northern Ireland containing literary speech representation. Thus, I intend to fill that gap by compiling such a corpus and the aim of this paper is to describe the way I plan to build it, focusing first on the main purpose to which my corpus will be put, then, on some key features that need to be considered before and during the creation of the corpus and, thirdly, on text collection and data entry, storage and retrieval.
Regarding the main purpose, the creation of my corpus is prompted by an interest in exploring how the fictional representation of NIE in different literary works can provide useful information about phonetic but not only phonetic features of this dialect in the context of present-day Northern Ireland. And it is this purpose that guides the building of the corpus in terms of overall and sample size, number of samples and annotation. All these aspects deserve some attention because they play a major role in making a corpus representative. The issue of representativeness is somewhat thorny since as Kennedy (1998) points out “it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even of a particular genre or subject field or topic” (p. 62). In spite of that, I will make my corpus as representative of a language variety, the English spoken in Northern Ireland, and of a literary genre, Northern Irish fiction, as possible.
Finally, it is worth noticing that the creation of my corpus is part of a greater project that aims at compiling a corpus of contemporary Irish literature produced both in the Republic of Ireland and in Northern Ireland that contains fictionalised representation of both Southern and Northern Irish accents.

References
73. Dömötör, Andrea. “All that predicates is not VP – Identifying predicate nominals in automatic parsing”  
This research aims to define an algorithm to detect predicate nominals in Hungarian sentences with an automatic parsing tool.

In Hungarian, and in many other languages, predicate nominals are morphologically unsigned. They only hold a default morphological case which is usually the nominative (Schütze 2001). For (automatic) parsing, however, it is crucial to distinguish “real nominatives” in subject position from default nominatives that form the predication of the sentence. In the case of Hungarian, the identification of the predicative role of a nominal is especially important due to the so-called zero copula. In present tense third person the Hungarian nominal sentences have no overt verb (see example [1]), therefore the head of these sentences in dependency parsing is the predicate nominal itself.

(1) a. János orvos.  
John doctor  
‘John is a doctor.’

b. Mi orvosok vagyunk.  
we doctor-PL be-PL1  
‘We are doctors.’

The parser in the focus of the study is psycholinguistically motivated which means that its parsing algorithm models the human language processing. Therefore, it parses the sentence in real time, word by word, based on the left context of the current word. The detection of nominal predicates follows the same principle: the algorithm tries to make a decision of the role of the nominal based on the preceding words.

For the elaboration of the algorithm I used a tokenized, morphologically annotated and disambiguated corpus (Oravecz – Váradi – Sass 2014). I extracted nominal sentences from the corpus and analyzed the left context of the predicate nominals with both automatic and manual methods. Based on the findings of the corpus analysis and also some linguistic considerations, I built a rule-based model and implemented it in Python. The program’s input is a nominal in nominative and the part of the sentence which precedes it. The output is a response of trivalent logic: True if the nominal is obviously a predicate, False if it is obviously not, and Undefined if its role is still unclear from the left context.

For the evaluation of the algorithm I used 300 test sentences which contained in total 534 nominals at issue. The test data was manually annotated using two kinds of annotation: one label showed the expected state of the nominal based on the left context (True, False or Undefined), the other one indicated its real syntactic role (True or False). This way the evaluation provides information which is interesting both for computational linguistics and psycholinguistics. The comparison of the program’s output and the first type of annotation shows in what measure can the left context help in the decision rather a nominal is predicative or not. The results revealed that in 78% of the cases the input preceding the nominal unambiguously specifies its syntactic status, no backtrack is needed.

The results are promising regarding the possibilities of rule-based automatic parsing. Besides, the findings of the model building corpus research provides interesting observations for theoretical linguistics as well.

References

The emergence and development of the Way Construction, exemplified in (1) and (2), was famously examined by Israel (1996) in a study which traced the modern form of the construction to three senses or subschemas, namely manner of motion (He stumbled his way to the back of the room), means of motion (cf. [1]) and accidental activity (cf. [2]).

(1) a. With swerdis dynt behuffis ws, Throw amyddis our enemys red our way.  
‘(It) behoves us by force of arms, (to) clear our way through among our enemies.’  
(1513 Douglas tr., Virgil Aeneid x. vii. 30; OED redd v.2 1.c.)

b. A fancyt animal, white as snow, brought by one of the visitors, purred its way gracefully among the wine-cups,

(2) a favourite animal, white as snow, brought by one of the visitors, purred its way gracefully among the wine-cups,

(1513 Douglas tr., Virgil Aeneid x. vii. 30; OED redd v.2 1.c.)
111. Fernández Pena, Yolanda. “Do syntax, semantics and agreement go hand in hand in complex subjects?”

Collective nouns in English have been the object of several investigations aimed at exploring their patterns of agreement (Levin 2001, 2006; Depreafere 2003; Hundt 2006, 2009). Nonetheless, to this date research in this field has mostly overlooked complex collective (i.e. NCOLL-of-N) subjects taking of-dependents and involving potential agreement conflicts such as:

(1) A group of sailors come to the front with the captain.
(2) The majority of the public are against [sic] hunting.

In prior research I have attested the role of overtly-marked plural of-dependents (e.g. (1)) in verb agreement: they reinforce the conceptual plurality of the collective and thus trigger significantly higher rates of plural agreement. The contribution of this study rests on: (i) the exploration of semantically plural (e.g. (2)) and/or morphologically singular dependents, and (ii) the interaction between the syntactic and semantic properties of the of-PPs and their influence on verb number. To this end, I have investigated the patterns of agreement of 23 collective nouns that take of-dependents (list retrieved from Biber et al. 1999: 249; Huddleston and Pullum 2002: 503) in samples of the written parts of the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA).

The results provide further statistical support for the strongly conditioning effect of plural of-dependents and show that the preference for plural agreement of complex NCOLL-of-N subjects is constrained by the morphology of the NCOLL-of-N itself.

References


(Alcaide, 2004; González Sanz, 2016) o emotivo (Nieto y Otero, 2003), pero en escasas ocasiones se ha observado qué influencia tiene, como medio persuasivo, la adaptación geolinguística de los gobernantes a sus propios votantes. Uno de los estudios más relevantes para conocer este tipo de variación es el realizado por Hernández Campoy y Cutillas (2010) en el que analizan el discurso de los gobernantes de la región de Murcia desde un punto de vista variacionista, y con el que advierten que estos suelen utilizar variantes vernáculas en la mayor parte de sus intervenciones, e incluso la propia presidenta de la Comunidad Autónoma suele adaptar su discurso a los hablantes.

Siguiendo este trabajo, el objetivo principal de esta investigación es comprobar cuáles son las semejanzas y diferencias lingüísticas entre el discurso político de los tres presidentes autonómicos de la democracia extremeña: Juan Carlos Rodríguez Ibarra, Guillermo Fernández Vara y José Antonio Monago. En un primer momento, y dado el carácter más o menos conservador de los partidos que han gobernado en Extremadura durante las últimas tres décadas (PSOE y PP), es posible que los dirigentes hayan adoptado una serie de marcas para poder tener una mayor influencia en el voto y el seguimiento de los habitantes de Extremadura. Por una parte, creemos que una de ellas puede ser la marca de solidaridad lingüística. Dado que la región extremeña es un área innovadora en la que se mantienen fenómenos como la aspiración de losson impresivos o de la velar /x/, la elisión de ciertas consonantes en posición final, la pérdida de /d/ intervocálica etc., los presidentes pueden hacer uso de estos rasgos para adaptarse lingüísticamente a sus votantes y lograr, así, una mayor cercanía.

Asimismo, creemos que pueden existir diferencias entre los discursos escritos y orales por influencia de algunos factores como el diseño de audiencia (Bell, 1984), y también entre los discursos autonómicos y los nacionales, dado que el público al que se dirigen, en este caso, es diferente. Será fundamental, también, indagar en cómo influye la procedencia (regional y familiar) de cada presidente, su formación académica, su puesto profesional, y la época en la que han gobernado.

Para poder comprobar estos objetivos se ha diseñado un corpus lingüístico compuesto por 15 intervenciones de los presidentes: 1 en la Asamblea de Extremadura, 2 nacionales (1 entrevista de radio y 1 en televisión) y 2 regionales (con las mismas características que en la anterior) que suponen más de 5 horas de grabación. Con este trabajo se abre una puerta al estudio del discurso político en Extremadura, un campo poco cultivado pero que puede abrir nuevas vías de investigación.

References

121. Flores Hernández, Ana Abigail. “Adquisición de los sufijos derivativos en aprendentes de inglés como L2”

Morphological competence is the ability to recognise the compositional units of a word (Koda, 2000; Tyler and Nagy, 1987; Liu & McBride-Chang, 2014; Zhang, 2014); the ability to identify the meaning and function of this units (McBride-Chang, Wagner, Muse, Bonnie, Chow & Shu, 2005); and the ability to use them in word formation through the mastering of their combinatorial rules and restrictions (Zhang y Widyastuti, 2010; Lardiere, 2006; Liu & McBride-Chang, 2010; Lardiere, 2005; Petrush, 2008). Acquisition of derivative morphemes in a second language is related to the growth of vocabulary size, the use of more lexical units (Schmitt and Zimmerman, 2002) as well as meaning decoding of complex words (Silva and Clahsen, 2008; Kim, Wang and Ko, 2011).

This research aim is to propose a possible order of acquisition of derivative suffixes in native Mexican-Spanish speakers, through the analysis of the morphemes appearing in a corpus of academic written language from undergraduate English learners; and the evaluation of the mastering of this suffixes through the creation of a morphological test. Additionally four aspects of morphemes have been evaluated to determine the impact of those in the order of acquisition proposed. The first aspect is related to the morphological equivalence, congruency or transparency (Jiang, Novoskanova, Masuda and Wang, 2011; Petrush, 2008), which refers to the existence of a L1 morpheme more or less similar in meaning and function to a L2 morpheme; the second one is related to the semantic complexity of morphemes (Goldschneider and DeKeyser, 2001), the number of meanings a form can take; the third aspect was the frequency, determined by the number of tokens and types of each morpheme in a reference corpus; and finally morphological regularity, the number of allomorphs a morpheme can have (Goldschneider and DeKeyser, 2001).

Corpus creation resulted in 156 texts 25,642 tokens, 3,224 types and a total of 2,807 complex words. The number of derivational suffixes found in the learner’s productions were 51 of the 282 considered at Morphoquantics (Laws and Ryder, 2014). 31 at level A2, 45 at level B1, 49 at B2 and 51 at C1 essays. Results indicated a consistent growth in the number of morpheme types from level A2-C1, showing a developmental process by emergency criteria. The classifying of derivative suffixes resulted in 43 levels of frequency, 5 levels of morphological regularity; 7 levels of semantic regularity and 7 levels of equivalence. The construction of the morphological awareness test is in progress, to build its items the appearing suffixes were selected according to their level of proficiency, equivalence, regularity and frequency considering four operations of the morphological competence, suffix meaning recognition, suffix syntactic category recognition and knowledge of suffix combinatorial rules.

References
Word2vec is mapping words to a D-dimensional vector representation, in such a way where semantically similar words tend to be close in the vector space. For our experiments, training and testing data are taken from the Corpus of Greek Texts (30 million words approx.). We have chosen twenty-four transitive verbs in Greek as used in 914 phrases including object complements. The distinction between literal and metaphor phrases has been based on the Metaphor Identification Procedure (MIP) suggested by the Pragglejaz Group (Stein). By applying MIP to phrases from the Corpus, we have manually created two lists, one with 459 literal sentences and another one with 455 metaphoric ones. In these two lists, the same verbs are used as sentence predicates with a variety of different objects. Collocations and delexical verbs were excluded from the testing corpus. An aggregate vector was thus created to represent each sentence and the label metaphor or literal was assigned as a consequence of the MIP. The generated representations were then used to train a SVM classifier to predict if a phrase is metaphor or literal.

In order to evaluate model performance a 10-fold cross validation was applied. The cross-validation accuracy of the model is 0.82 with f1-score 0.77, an especially good result for resource-poor languages.

References


Corazón tan blanco is in all probability one of the best works of the contemporary Spanish writer Javier Marias. This novel has enjoyed great success both in Spain and abroad, and has been translated into 37 languages, since it was first published in 1992. For Bösader (1998: 182), Corazón tan blanco succeeds in luring its readers with a repertoire of clichés of all kinds, and a combination of references to seemingly lofty discourse, such as the myth of New York, or Shakespeare’s Macbeth, which inspired the title. Other scholars (see, for example, De Maeseeneer, 2000, Christie 1998, Logie 1998, Gonzalez de Avila 1999, Simonsen 1999) have pointed out, however, that the most remarkable aspect of Corazón tan blanco is its language. Perhaps, one of the most prominent linguistic features present in Marias’ thriller is the literary use of repetition, particularly “phraseological repetitions”. Although repetition in the novel has been addressed by experts in the field of literary criticism (cf. Grohmann 1992, Basso, 2002, Flörchheim 2003), none of the earlier accounts has provided a systematic analysis of phraseological repetition in Corazón tan blanco. To our knowledge, earlier approaches have not addressed the quantitative, distributional and discursive peculiarities of phraseological repetitions together with their function in the narrative development, which is the focus of the present talk. Just to give an example from our research, the two-word sequence “futuro abstracto” yields no less than thirteen occurrences throughout the novel. Also, this same sequence is embedded in longer sequences such as “el futuro abstracto ... que es el que importa...”
porque el presente no puede teñirlo ni asimilarlo"; which appears in different chapters. We have been able to detect word repetitions by means of ngram techniques. However, the existence of phraseological variation required a subsequent careful manual approach. Other corpus techniques, such as file view and concordancing have also been carried out for a more qualitative analysis. The present study followed a corpus stylistics approach (cf. Mahlberg 2013; Mahlberg & McIntyre (2011) and was inspired, more specifically, by Stubbs' quantitative stylistic analysis of Conrad's Heart of Darkness (2005). The results of the quantitative and qualitative corpus approach allowed us to identify repetition and determine its literary function in the narrative of Corazón tan blanco.

References

Mahlberg & McIntyre (2011) A case for corpus stylistics: Ian Fleming's Casino Royale. Corpus Stylistics and Dickens’s Fiction


En 1978, John J. Nitti, uno de los editores del Dictionary of the Old Spanish Language (DOSL) y co-fundador del Hispanic Seminary of Medieval Studies (HSMS), describe en un artículo titulado “Computers and the Old Spanish Dictionary” un proyecto a largo plazo: the creation of the Old Spanish Archive (OSA), which is to be a repository ... of all the machine-readable manuscripts and concordances of those works represented in DOSL... OSA will be established as a research archive open to any interested scholars wishing to make use of its facilities ... eventually ... information retrieval will be carried out via the computer ... linking the magnetically-stored ... machine-readable text transcriptions and concordances (Nitti 1978: 52) Este proyecto superaba, en el momento de su concepción, las posibilidades informáticas disponibles en la época, siendo entonces el objetivo a medio plazo del HSMS la creación y divulgación de la vasta base de datos compuesta por las transcripciones electrónicas de manuscritos e incunable escritos en español entre los años 1000 y 1600, utilizando para ello las

microfichas y el CD-ROM. En el año 2011, con la aparición de la Biblioteca Digital de Textos del Español Antiguo (BIDTEA), el HSMS comienza a ofrecer acceso en línea a las transcripciones semi-paleográficas y a una serie de índices (alfabético, frecuencia, alfabeto inverso) y concordancias en formato KWIC. A partir del año 2015 el HSMS empieza a trabajar en el Old Spanish Textual Archive, un corpus lingüístico de cerca de 35.000.000 de palabras, basado en las más de 400 transcripciones semi-paleográficas de textos medievales escritos en castellano, asturiano, leonés, navarro-aragones y aragones realizadas por los colaboradores del HSMS. Tras la delimitación inicial del corpus textual comenzamos el proceso de etiquetado gramatical, para el que utilizando FreeLing, una herramienta de Procesamiento del Lenguaje Natural, y HSMS-app, una herramienta de análisis textual desarrollada para este proyecto.

Dedicamos el año 2017 a ampliar los recursos léxicos de FreeLing, trabajando en el reconocimiento de entidades nombradas (topónimos y antropónimos), de variantes ortográficas medievales y de palabras no identificadas por ninguna de las reglas desarrolladas. Para ello procesamos varios de los diccionarios del proyecto Dictionary of the Old Spanish Language del HSMS: el Diccionario español de textos médicos antiguos (Herrera 1996), el Diccionario español de documentos alfonsi (Sánchez 2000), el Vocabulario militar castellano (siglos XIII-XV) (Gago Jover 2002), el Diccionario de la prosa castellana del Rey Alfonso X (Kasten y Nitti 2002), el Diccionario herbario de textos antiguos y premodernos (Capuano 2017). De esta forma pudimos añadir unas 30.000 entradas más al diccionario, rebajando el número de formas sin analizar en todo el corpus a 700.000 (un 2% del corpus). En una fase posterior, la edición manual de las 3.000 formas sin identificar más frecuentes nos permitirá solucionar otros 240.000 casos, dejando el porcentaje de formas desconocidas muy cerca del 1%. En esta porcentajes pretendemos detalizar los procedimientos seguidos en este proceso de mejora y ampliación del diccionario base de FreeLing, y presentar una versión de trabajo de la interfaz de consulta de la base de datos, aún en desarrollo, que permitirá combinar búsquedas en todos los niveles de marcamiento anotados y filtrar los resultados según los campos recogidos en la meta-descripción de cada obra y de cada códice.

References
Sánchez-Marco, Cristina, Gemma Boleda, y Lluís Padró (2011): “Extending the tool, or how to annotate historical language varieties”, Proceedings of the 5th ACL-HLT Workshop on Language Technology for...
Garofalo, Giovanni. “Conectores contraargumentativos del español en las resoluciones procedentes del Tribunal Supremo de España y del Tribunal de Justicia de la UE: Análisis contrastivo basado en corpus”

Esta comunicación pretende analizar la diferente distribución de los conectores contraargumentativos (Montolío 2001: 45-98) en un corpus de sentencias del Tribunal Supremo de España (TS) (767.903 palabras) y en un corpus de resoluciones en castellano del Tribunal de Justicia de la Unión Europea (TJUE) (586.585 palabras).

Comparando el número de ocurrencias, las concordancias y la frecuencia normalizada de los conectores en cuestión, llama la atención la mayor cantidad y variedad de marcadores contraargumentativos empleados por el TS (15 tipos distintos) con respecto al abanico más restringido y a la menor incidencia de dichos conectores en las resoluciones del TJUE (11 tipos).

En el análisis se propone se baraja una doble explicación de esta divergencia distribucional, que puede deberse, en primer lugar, al distinto funcionamiento de los dos órganos jurisdiccionales implicados (Espaliú Berdud 2010, Prieto 2008, Caportori 1994). De hecho, la estructura de la sentencia del TJUE suele variar según la composición del órgano (salas de 3, 5, 7 miembros), mientras que el lenguaje de los tribunales españoles (TS) suele ser más marcador y a la menor incidencia de dichos conectores en las resoluciones del TJUE (11 tipos).

Para el análisis de los conectores se propone una doble explicación de esta divergencia distribucional, que puede deberse, en primer lugar, al distinto funcionamiento de los dos órganos jurisdiccionales implicados (Espaliú Berdud 2010, Prieto 2008, Caportori 1994). De hecho, la estructura de la sentencia del TJUE suele variar según la composición del órgano (salas de 3, 5, 7 miembros), mientras que el lenguaje de los tribunales españoles (TS) suele ser más marcador y a la menor incidencia de dichos conectores en las resoluciones del TJUE (11 tipos).

En el análisis se propone se baraja una doble explicación de esta divergencia distribucional, que puede deberse, en primer lugar, al distinto funcionamiento de los dos órganos jurisdiccionales implicados (Espaliú Berdud 2010, Prieto 2008, Caportori 1994). De hecho, la estructura de la sentencia del TJUE suele variar según la composición del órgano (salas de 3, 5, 7 miembros), mientras que el lenguaje de los tribunales españoles (TS) suele ser más marcador y a la menor incidencia de dichos conectores en las resoluciones del TJUE (11 tipos).

En el análisis se propone se baraja una doble explicación de esta divergencia distribucional, que puede deberse, en primer lugar, al distinto funcionamiento de los dos órganos jurisdiccionales implicados (Espaliú Berdud 2010, Prieto 2008, Caportori 1994). De hecho, la estructura de la sentencia del TJUE suele variar según la composición del órgano (salas de 3, 5, 7 miembros), mientras que el lenguaje de los tribunales españoles (TS) suele ser más marcador y a la menor incidencia de dichos conectores en las resoluciones del TJUE (11 tipos).

En el análisis se propone se baraja una doble explicación de esta divergencia distribucional, que puede deberse, en primer lugar, al distinto funcionamiento de los dos órganos jurisdiccionales implicados (Espaliú Berdud 2010, Prieto 2008, Caportori 1994). De hecho, la estructura de la sentencia del TJUE suele variar según la composición del órgano (salas de 3, 5, 7 miembros), mientras que el lenguaje de los tribunales españoles (TS) suele ser más marcador y a la menor incidencia de dichos conectores en las resoluciones del TJUE (11 tipos).

En el análisis se propone se baraja una doble explicación de esta divergencia distribucional, que puede deberse, en primer lugar, al distinto funcionamiento de los dos órganos jurisdiccionales implicados (Espaliú Berdud 2010, Prieto 2008, Caportori 1994). De hecho, la estructura de la sentencia del TJUE suele variar según la composición del órgano (salas de 3, 5, 7 miembros), mientras que el lenguaje de los tribunales españoles (TS) suele ser más marcador y a la menor incidencia de dichos conectores en las resoluciones del TJUE (11 tipos).

En el análisis se propone se baraja una doble explicación de esta divergencia distribucional, que puede deberse, en primer lugar, al distinto funcionamiento de los dos órganos jurisdiccionales implicados (Espaliú Berdud 2010, Prieto 2008, Caportori 1994). De hecho, la estructura de la sentencia del TJUE suele variar según la composición del órgano (salas de 3, 5, 7 miembros), mientras que el lenguaje de los tribunales españoles (TS) suele ser más marcador y a la menor incidencia de dichos conectores en las resoluciones del TJUE (11 tipos).

En el análisis se propone se baraja una doble explicación de esta divergencia distribucional, que puede deberse, en primer lugar, al distinto funcionamiento de los dos órganos jurisdiccionales implicados (Espaliú Berdud 2010, Prieto 2008, Caportori 1994). De hecho, la estructura de la sentencia del TJUE suele variar según la composición del órgano (salas de 3, 5, 7 miembros), mientras que el lenguaje de los tribunales españoles (TS) suele ser más marcador y a la menor incidencia de dichos conectores en las resoluciones del TJUE (11 tipos).

En el análisis se propone se baraja una doble explicación de esta divergencia distribucional, que puede deberse, en primer lugar, al distinto funcionamiento de los dos órganos jurisdiccionales implicados (Espaliú Berdud 2010, Prieto 2008, Caportori 1994). De hecho, la estructura de la sentencia del TJUE suele variar según la composición del órgano (salas de 3, 5, 7 miembros), mientras que el lenguaje de los tribunales españoles (TS) suele ser más marcador y a la menor incidencia de dichos conectores en las resoluciones del TJUE (11 tipos).

En el análisis se propone se baraja una doble explicación de esta divergencia distribucional, que puede deberse, en primer lugar, al distinto funcionamiento de los dos órganos jurisdiccionales implicados (Espaliú Berdud 2010, Prieto 2008, Caportori 1994). De hecho, la estructura de la sentencia del TJUE suele variar según la composición del órgano (salas de 3, 5, 7 miembros), mientras que el lenguaje de los tribunales españoles (TS) suele ser más marcador y a la menor incidencia de dichos conectores en las resoluciones del TJUE (11 tipos).

66. Giner González, Celia, Carlos Castelló Vercher and Danny Murillo Lanza. “¿Qué guay!: El uso de los corpus para la extracción de frases valorativas coloquiales y su enseñanza en la clase de ELE”

El presente trabajo surge ante la necesidad que se tiene en el aula de ELE de enseñar ciertos contenidos lingüísticos, funcionales y socioculturales a partir de muestras reales de la L2. En concreto, nos centramos en el estudio de ciertas frases valorativas coloquiales para su aplicación y enseñanza. Por frases valorativas coloquiales entendemos aquellas intervenciones que utilizamos para reaccionar y valorar determinados hechos lingüísticos y extralingüísticos. A partir del análisis de los datos extraídos de corpus conversacionales, tales como Val.Es.Co. 2.0 o Preseea, llevamos a cabo una clasificación de las expresiones en función de su polaridad positiva, negativa o ambigua, es decir, aquellas que se pueden emplear para las dos primeras. En este sentido, los corpus conversacionales, tanto en su formato oral como escrito, son excelentes recursos porque son fuentes fiables a la hora de elaborar nuestro estudio y, por otra parte, nos permiten que los contenidos de la enseñanza estén constituidos por muestras de habla reales y de uso actual. Asimismo, algunos autores como Albelda y Briz (2017) y Pitkowskij y Vásquez (2009) señalan en sus aportaciones la importancia y funcionalidad del uso de los corpus lingüísticos como herramienta pedagógica para la enseñanza y aprendizaje de ELE. En definitiva, nuestro objetivo es elaborar una propuesta metodológica para la enseñanza de este tipo de expresiones coloquiales a partir del uso de corpus. Los resultados son interesantes debido a que hemos encontrado muestras ideales de diversa índole en diferentes corpus que señalan la utilidad y necesidad de este recurso lingüístico en el aula de ELE. Tras este estudio de corpus, hemos diseñado la siguiente aplicación práctica, si bien serían posibles muchas otras. La actividad se desarrolla de la siguiente manera: (1) el profesor presenta expresiones valorativas tales como “¡qué guay!” o “¡qué horror!” en muestras reales de textos extraídos de los corpus conversacionales mencionados; (2) los estudiantes reconocen la polaridad positiva, negativa o ambigua de estas expresiones y sus contextos comunicativos de uso de manera autónoma con la revisión del profesor; (3) el profesor presenta unos memes, en los que se emplean estas expresiones asociadas a unas imágenes, y unas herramientas para que ellos mismos los creen en una actividad posterior; (4) los estudiantes elaboran dos memes que contengan expresiones de cada grupo (positivo, negativo y ambiguo) junto a una oración que contextualice cada imagen con su expresión y los publican en redes sociales; (5) los estudiantes reaccionan con un comentario en el que pueden hacer uso de las expresiones aprendidas en los memes de sus compañeros; (6) finalmente, los estudiantes exponen sus memes y el profesor, junto a los alumnos, los comenta; por otra parte, se premiará al mismo que ha tenido más reacciones de los...
compañeros. En conclusión, podemos afirmar que los corpus conversaciones resultan de gran utilidad en la didáctica de segundas lenguas.

Referencias


113. Gómez Jiménez, Eva M. “‘A divided nation’: The representation of social mobility in the right-of-centre British press from 1997 to present”

In one of his first speeches after being elected Prime Minister, Tony Bair committed his administration to re-build Britain ‘as one nation, in which each citizen is valued and has a stake; in which no-one is excluded from opportunity and the chance to develop their potential; in which we make it, once more, our national purpose to tackle social division and inequality’ (Blair, 1997). 20 years after such declaration of intent, a commitment that was also further shared by Brown’s, Cameron’s and May’s later cabinets, the social mobility policies have been assessed. The report, published in June 2017 by the Social Mobility Commission, shows that, although some of these policies have had a positive impact, these have not built Britain as one nation with opportunities for everybody to improve their social status, but, on the contrary, has led to a more divided, unequal country where improving social position is now more difficult than it was in 1997.

Parting from this background, and relying heavily in the social constructionist theory (Fowler, 1991; Fairclough, 2015), this paper aims at analysing how social mobility has been represented in the right-of-centre UK press during this 20-years period that has been assessed by the Social Mobility Commission. More concretely, three research questions will be addressed here: (1) Which are the different topics addressed in connection to social mobility in the past 20 years? (2) What are the keywords that characterize the representation of social mobility during this period? And finally, (3) is there any particular term social mobility has been associated to?

To answer these questions, this study combines the tools provided by corpus linguistics and critical discourse analysis (Baker, 2006, 2013; Jeffries, 2010; Jeffries and Walker, 2015; Partington et al., 2013). It uses data from two different corpora: the target corpus, which consists on newspaper material from 1998 to 2016 (the first and last full years in which social mobility policies where assessed), and the reference corpus, which consists on real language from the late 20th century, as collected in the BNC written sample (2007). The data for the target corpus has been extracted from Nexus UK (2017) by using que query phrase ‘social mobility’ specifically from the Times and Daily Mail, as these are two of the widest circulated right-of-centre newspapers in UK and they represent samples of both tabloid and broadsheet press, thus accessible for British people from many different backgrounds. The corpus contains 1,556,245 words and 1940 news stories. Different tools in Antconc 3.4.4m (Anthony, 2014) will be used to identify the most frequent words, the statistical keywords and the collocates of ‘social mobility’.

Preliminary findings suggest there has been a strong connection between social mobility and education during these 20 years. This is, currently, work in progress, so further results and conclusions are still to come.

References


The get-passive construction has received a great deal of attention over the last few decades, although it cannot still be compared with the extensive body of literature devoted to the regular be-passive. The study of this phenomenon has been approached from many different angles. Among them, the main areas of research have been register and style, regional distribution, historical development, possible relation with the middle voice, and the semantic and pragmatic features that distinguish it from the be-passive.

In most recent studies, it is widely assumed that the combination of get + past participle constitutes a fuzzy category where several subtypes of related constructions can be identified. In this sense, there is a general consensus that core get-passives exhibit the following defining features: a lexical verb with dynamic meaning, a be-passive counterpart, a corresponding active equivalent, and an explicit or implicit agent-phrase (see Collins 1996, Downing 1996). The first of these properties seems to exclude stative predicates of perception, cognition or emotion such as see, know, like or love from the get-passive. In fact, some authors clearly state that stative predicates are incompatible with the get-passive, hence the unacceptability of examples of the type: “*Freddie got seen” or “*The truth got known” (see Chappell 1980, Alexiadou 2005). However, get-passives with verbs from these three semantic classes have already been attested and are claimed to be possible, although infrequent, when properly contextualized (see Downing 1996, Thompson and Scheepers 2013). Out of these three classes, however, emotion verbs are still extremely rare in the get-passive. One possible reason for the widespread view that stative verbs do not combine with get in passive constructions may be the lack of real language corpora large enough to provide evidence of their actual occurrence in this pattern.

With this in mind and taking into account the above-mentioned strong resistance that verbs of emotion in particular apparently exhibit as regards the get-passive, the aim of this work is to show that this semantic class is indeed compatible with get in its passive use. For this purpose, five verbs of emotion --love, like, admire, dislike and hate--have been chosen and analysed. To
try to overcome the shortcomings of smaller corpora previously used, I have based my study on material retrieved from the Corpus of Global Web-Based English (GloWbE)—composed of 1.9 billion words from 1.8 million web pages in twenty English-speaking countries, and from the News of the Web Corpus (NOW)—consisting of 4.5 billion words from web-based newspapers and magazines. The data collected has been further examined in order to establish whether the instances of get followed by the past participle of these verbs found in the corpora display the semantic features and pragmatic implicatures commonly associated with the get-passive and absent from the be-passive, namely, resultative meaning, preference for a human subject, lack of an explicit agent-phrase, control and responsibility of the subject, adverse or beneficial consequences of the event, and emotional involvement of the speaker.

References

126. Gottlieb, Henrik. “Using existing text corpora, news archives and dictionaries to determine the inventory of Anglicisms: The making of the Danish GLAD contribution”
This presentation will focus on the trials and tribulations of the compiler of Anglicisms for the GLAD A-Z database.
The database, projected to be completed by 2021, will chart the English lexical and phraseological influence on languages worldwide. At present, fourteen languages are represented with contributions consisting of Anglicisms based on English etymons beginning with the letter O.
Danish is one of these languages, and I will show how corpora may contribute to heightening the quality of lemma selection and dating of (contemporary) Anglicisms. Having discussed the various criteria for inclusion in the GLAD database, I will demonstrate the usefulness of two Danish online sources, a 500-million word bona fide corpus accessed via the CoREST online tool, and a 25-billion word news archive (Infomedia), in ascertaining whether Anglicism candidates qualify statistically for inclusion.
It is worth noting that neither corpus-driven nor strictly corpus-based methodology is suited for the detection of Anglicisms. Only (some) unadapted English borrowings may be found that way, lexemes as ‘weekend’ and ‘quiz’. In a Germanic language like Danish, even many unadapted borrowings will go unnoticed if Anglicism-detection algorithms are used; neither ‘smart’ nor ‘internet’ would be identified automatically. And the wealth of adapted borrowings, loan translations and semastic loans would also be overlooked in the process of corpus-based retrieval.
Thus, instead of corpus-based retrieval, I use corpus-based testing of ‘manually found’ candidates. As this is a work-in-progress presentation, it remains to be seen how many of the 1,288 Danish Anglicism candidates now collected beginning with letters A or B will pass the frequency test. My guess is around 1,200. Knowing that A and B lexemes represent 12.7% of the entries in the corpus-based standard dictionary Den Danske Ordbog, the final version of the Danish contribution may present almost 10,000 entries – a significant share of modern Danish lexis.

References
Infomedia (1990 to date). www.infomedia.dk.

50. Hunt, Jaime. “What all the cool kids are talking about: Analyzing spoken corpora for anglicisms in German”
Speakers of German have often been criticised for using anglicisms such as cool and kids in their speech in order to sound sophisticated, intelligent, and/or trendy. Furthermore, many critics also suggest that Anglicisms are overwhelming the German lexicon and, therefore, threaten to turn the German language into Kauderwelsch or gibberish (see Dodd 2015). This paper will challenge this critique by providing evidence contrary to this opinion in the form of a lexical analysis of spoken German.
This analysis adopts Görblach’s (1994: 224) definition of an anglicism as “a word or idiom that is recognizably English in its form (spelling, pronunciation, morphology, or at least one of the three)”. This paper will expand upon the findings discussed in Hunt (forthcoming) by presenting a revised investigation into the spoken corpus of German spoken in 100 cities, containing 44,357 types and 1,185,151 tokens created by combining the corpora Regional Variants of German 1 (RVG1) and Hempel’s Sofa, from the Bayerisches Archiv für Sprachsignale ‘Bavarian Archive for Speech Signals’, as well as Deutsch Heute (DH) ‘German Today’ from the Institut für deutsche Sprache (IDS) ‘Institute for German Language’. This combined corpus is then merged with the Forschungs- und Lehrkorporus für gesprochenes Deutsch (FOLK) ‘Research and teaching corpus for spoken German’ (56,880 types and 1,672,863 tokens), from IDS’s Datenbank für gesprochenes Deutsch ‘Database for spoken German’, to create a corpus of over 2.8 million word tokens.
Spoken corpora have been underutilised in the analysis of the influence that English has on German (one notable exception being Glahn 2002), with the majority of publications (e.g., Gnatchuk 2015, Knospe 2007, Onyso 2007, Onyso 2012) employing the print media as corpora. However, these studies present only one genre of edited and often deliberately innovative language. Using a corpus of mostly spontaneous spoken language obtained in a variety of situations adds to this knowledge by providing insight into an arguably more accessible and ubiquitous form of German.
In addition to analyzing the frequency the anglicisms appearing in this combined corpus relative to autochthonous terms, this paper will investigate collocations of the top 10 most frequently occurring lexical Anglicisms in spoken German by employing the t-score collocate measure.
Preliminary results indicate that the lexical impact of English upon spoken German is relatively minor, with 6,175 tokens per million (or 0.6%) of the DH, RVG1, and Hempel’s Sofa corpora being of recognizable English origin. The types of Anglicisms in these corpora number 4.7% of all word types. Full findings will be presented at the time of presentation.

References
Tradicionalmente se ha venido considerando el siglo XVIII como una época en la que comienza el español moderno y en la que no se documentan procesos lingüísticos que susciten el interés de los investigadores, lo cual redunda en la escasez de trabajos documentados en textos del llamado primer español moderno (Octavio de Toledo, 2007 y 2008), a excepción de los estudios sobre el léxico. No obstante, la frontera entre el período clásico y el moderno se discute (Melis, Flores y Bogard, 2003; Girón Alconchel, 2008; Eberenz, 2009; Sánchez Lancis, 2012), debido a que recientemente han proliferado trabajos sobre esa época (García-Godoy, 2012; Carrasco Santos, 2015; Guzmán Riverón y Sáez de Rivera, 2016), tomando como base tanto obras literarias como corpus documentales no literarios, de España y de Hispanoamérica, trabajos que han demostrado que en el español dieciochesco se inician y desarrollan diversos cambios lingüísticos de índole gramatical, sintáctica o discursiva de indudable interés, muchos de los cuales quedan aún por analizar.

Además, como sucedía en siglos anteriores (Castillo Gómez, 2002; Fernández Alonso, 2009), no pasa despercebido el volumen de la correspondencia epistolar, no solo entre secretarios y cargos diplomáticos, sino también entre comerciantes y entre particulares, resultado de las relaciones sociales y familiares de esta época. Dada la importancia de esta vía de comunicación interpersonal, a lo largo del Setecientos vieron la luz diversos tratados sobre esa época (García-Godoy, 2012; Carrasco Santos, 2015; Guzmán Riverón y Sáez de Rivera, 2016), tomando como base tanto obras literarias como corpus documentales no literarios, de España y de Hispanoamérica, trabajos que han demostrado que en el español dieciochesco se inician y desarrollan diversos cambios lingüísticos de índole gramatical, sintáctica o discursiva de indudable interés, muchos de los cuales quedan aún por analizar.

Por todo lo expuesto, en este trabajo estudiaremos un corpus epistolar de indiscutible valor lingüístico procedente de la colección documental de los Marqueses de Valdeleroy del Archivo Histórico Provincial de Córdoba. Se trata de un total de 32 cartas manuscritas, fechadas en su mayoría a lo largo del primer lustro de los años veinte del siglo XVIII y emitidas por Fray Andrés Hidalgo Calderón a su primo, también sacerdote, Gaspar Gallego Peñaflé. A través del estudio de estos textos observaremos las distintas fórmulas de tratamiento en comparación con otras cartas remitidas al mismo destinatario, las fórmulas ritualizadas del género epistolar, como las de saludo y despedida, para concretar en qué medida se estaba produciendo esa modernización que señalan los tratados sobre misivas; la estructuración sintáctica, atendiendo especialmente a aquellos aspectos que puedan ser más cercanos a la proximidad comunicativa dentro de la retórica propia de esta tradición textual; así como otras cuestiones discursivas relacionadas con la temática concreta de cada uno de los textos.


Pseudo-Anglicisms are a particularly forceful reminder of the influence of English on other languages. The paper seeks to complement Furiassi and Gottlieb’s (2015) monograph focusing on false Anglicisms in Germanic and Romance languages with a pilot study of pseudo-Anglicisms in Czech as a representative of the Slavic languages. After a brief survey of the literature on the concept, the paper presents a sample of Czech pseudo-Anglicisms, its collection (with material excerpted from the Czech neology database Neomat and other sources, e.g. the corpora of contemporary Czech and analysis) and the sample displays an interesting range of (subtypes of candidates for the pseudo-Anglicism status manifesting both the creativity and motivation of Czech speakers (though some pseudo-Anglicisms appear to be shared by several...
languages). The findings show that the notion of pseudo-Anglicisms is indeed a scalar one with prototypic structure. As a result, all kinds of variants exist in the area between prototypical Anglicisms and prototypical pseudo-Anglicisms. In addition to the challenge of identifying suitable candidates, the decision where to draw the borderline between these two poles is difficult to make and depends on the criteria applied to the concept of pseudo-Anglicisms, such as their origin (created by non-native speakers), identifiability (not "recognizably English in its form", cf. Görlach, 1994: 224) or familiarity and comprehensibility (recognizably English but not familiar and/or understandable to native speakers). This leads to disagreements, for instance, between those who recognize lexical, morphological and semantic pseudo-loans (see Yang 1990). Plümmer 2000, Glahn 2002, and others) and those who reject morphological and semantic pseudo-loans as true pseudo-Anglicisms as not based on English models (cf. Onysko 2007). In fact, the use of and/or emphasis on different criteria my point to different solutions. The paper hopes to contribute to the discussion by exploring and commenting on the situation in Czech.

References


This contribution has been coauthored with Silvia Bonacchi, Ewa Jarmolowicz-Nowikow, Maciej Karpinski and Tomasz Wicherkiewicz, and has been funded by the National Programme for the Progress in Humanities in Poland (NPRH12H13 0524 82).

In this contribution, the design and structure of a multimodal database consisting of audiovisual data from task-oriented dialogues and interviews is discussed. The primary purpose of the corpus is to provide material for investigation of paralinguistic phenomena contributing to communicative accommodation in conversations among young inhabitants of the Polish-German borderland. The database consists of audio and video recordings of task-oriented dialogues and of individual tasks as well as extensive metadata. The speakers are Polish and German secondary school pupils from two neighbouring towns located on the Polish-German border: Slubice and Frankfurt (Oder). Teenagers dialogues based on a competitive and a collaborative dialogue scenario as well as individual sessions based on a gesture-evoking procedure were recorded using camcordes and an audio recorder. Afterwards, speakers were individually interviewed on language and identity-related topics. Questions were mainly focused on language competences and contacts with other languages and cultures, family and its structure, personal and collective identity and the feeling of belonging (group-in-mechanisms).

Speech utterances have been transcribed orthographically and phonetically, and annotated for prosodic features with Annotation Pro (Klessa et al., 2013). The ELAN template (Wittenburg et al., 2006) used for gesture annotation covers gestural units, phrases, and phases, and includes independent tiers for the features potentially relevant to accommodation analysis (Giles et al. 1991). Gesture annotation schemes and procedures have been designed specifically for the project (Karpinski & Klessa, in print). Head movements and gaze direction changes are annotated using separate tiers. Annotation quality and coherence have been controlled by experts by means of manual scrutiny and validation as well as relevant statistical measures (incl. kappa; cf. Holle and Rein, 2015).

For the purpose of safe storage, retrieval and further analyses, all the data are uploaded to a relational database system built with Microsoft SQL Server technology together with a client management tool, developed specifically for the present purpose in Visual Studio .NET WinForms C#. Recording sessions, including all the related media and metadata files, are available to annotators and project team members via safe Internet connections (VPN). Sessions can be assigned to authorized annotators and transcribers, and the workflow of the project can be supervised and managed in a consistent way which prevents data loss.

Borderland Database currently serves as a resource for multimodal discourse analysis (Kress & van Leeuwen 2001, O’Halloran 2004), focused on accommodation processes in Polish-Polish, German-German and German-Polish task-oriented dialogues. Preliminary analyses of the corpus have been focused on mutual influences in gesticulation and prosody, culture-related usage of gestures, and code switching that involves using English as a common language by German and Polish pupils. However, a wide range of further applications of the database are possible thanks to the availability of both multilayer time-aligned annotations and metadata (cf. e.g., Ide & Pustejovsky, 2010) collected during the recording sessions and stored within the central relational database system together with the audiovisual materials.

References

This study developed and assessed a learner corpus for analyzing pronunciation (henceforth, phonetic learner corpus; PLC). Our PLC involves learners’ subjective judgment for pronunciation difficulty and learners’ speech sound for the basic English pronunciation, differently from previous PLCs. These properties help analyze learners’ recognition of pronunciation difficulty according to their proficiency. Learners’ subjective judgment for pronunciation difficulty was determined sentence-by-sentence on a five-point Likert scale (1: easy, 2: somewhat easy, 3: average, 4: somewhat difficult, 5: difficult). Our PLC also included recording data of learners’ reading aloud two English texts (ten-sentence, and five-sentence texts). Each text included the basic consonants and vowels in English. Learners were 50 university students learning English as a foreign language (Japanese as a primary language). Their proficiency ranged from Beginner level (A1 in CEFR (the Common European Framework Reference) to Advanced level (B2 in CEFR) in the scale of TOEIC (Test of English for International Communication).

As a result, our PLC consisted of the subjective judgment data (n = 750, mean = 4.2, SD = 0.8). More than 80% of the instances appeared between the subjective judgment 3 (the average) and 5 (difficult). Our PLC was assessed using the test theory. The reliability was examined by internal consistency (Cronbach’s α ranging from 0.0 (absence of reliability) to 1.0 (absolute reliability)), and our PLC marked a higher reliability (α = 0.89) than empirical satisfaction (0.8). Construct validity was examined for the distinctiveness of the subjective judgment by three proficiency levels. Analysis of variance showed significant differences (p < 0.01, F (2, 747) = 38.3). Tukey’s post-hoc comparison showed significant differences (p < 0.01) between Beginner and Advanced levels and Intermediate and Advanced levels, but not between Beginner and Intermediate levels. Criteron validity was examined by correlation of subjective judgment with TOEIC scores, and the correlation was moderate (r = -0.51).

The moderate construct validity due to the similarity between Beginner and Intermediate levels might be caused by inadequate population of learners and/or problems regarding texts. Besides these possibilities, let us consider a learner’s factor that Beginner level learners judged the pronunciation difficulty lower than their pronunciation skill. This overestimation also explains the moderate correlation in the criterion validity, because outliers were most frequently found in the data of Beginner level learners. This overestimation can be solved by adding learner properties that explain learners’ pronunciation skill. Moreover, this should be taken as a pedagogical issue, and English learning/teaching needs to let Beginner level learners recognize their pronunciation skill properly.

This study assessed the reliability and validity of our PLC. Despite the moderate validity, our PLC achieved the high reliability. A pilot investigation of our PLC demonstrated issues for corpus research and for developing a learning/teaching assistant tool that measures the difficulty of materials for pronunciation practice.

References

47. Kotani, Katsunori, Takehiko Yoshimi and Mayumi Uchida. “Opinion Corpus for Assessment of Study Abroad Program”

This study developed a corpus for capturing students’ opinions on study abroad program (henceforth, opinion corpus, OC). Previous OCs dealt with course evaluation by local students. Our OC targeted inbound international students, and involved opinions for course contents (Academics), for university dormitories or off-campus houses (Accommodation), and for interaction with local students (Interaction). Our OC was also annotated with positive and negative opinion polarity.

This study examined applicability of our OC to the assessment of study abroad program, because the assessment is taken as a new challenge. Our corpus-based assessment helps improve study abroad program, which results in the increase of students’ learning outcome.

Our OC was compiled through course evaluation in Spring (n = 302) and Fall (n = 298) semesters. The survey questions and students’ comments were written in English, their primary language. Students’ answers were given on five-point Likert scale and as open-ended comments. Then, our OC consisted of 600 students’ comments consisting of 43,736 words (17,438 words in Academics; 13,767 words in Accommodation; 12,531 words in Interaction). The comments were evaluated by a host institution’s faculty member on three-point Likert scale (negative; neutral; positive).

The Likert scaled answers and comment answers demonstrated a contrastive opinion polarity in Interaction. The rate of the positive scaled answers (63%) was lower than those in Academics and Accommodation (approximately 80%), but the positive opinions were more frequently observed. This contrast motivated us to develop a tool that determines the positive and negative opinion polarity.

First, Chi-square analysis demonstrated significant difference on the comment length (the number of words in a comment) among Academics, Accommodation, and Interaction (χ2(2) = 893.69, p < 0.01). The Ryan-Einot-Gabriel-Welch post-hoc test showed significant differences (p < 0.01) between Academics and the others. The comment length reflected students’ interest and concerns, because students answered without limitations on the comment length and the time for completion.

Second, frequency analysis of word uni- and tri-gram demonstrated frequent use of “would” and “I would like” in Academics. A modal verb “would” expressed the presence of students’ counter expectation, which should be taken as problems of study abroad program. In addition, the use of a first-person subject indicated students’ negative comments, because a first-person subject was infrequently used for expressing positive appreciation.

Then, the multiple regression analysis was carried out for predicting the opinion polarity in Interaction. Dependent variables were the opinion polarity scores, and independent variables were scores of readability, sentiment, and mood. The result demonstrated a significant regression equation (F (5, 152 = 18.8, p < 0.01). The adjusted squared multiple correlation coefficient was moderate (R2 = 0.36).
This study analyzed our OC for assessing study abroad program. The results demonstrated that corpus-based analysis of students' opinions was promising for study abroad assessment, and our OC was effective for developing a tool that determines the opinion polarities.

References

110. Leiva, Jorge. “Interferencias ortográficas y ortotipográficas en un corpus paralelo de textos museísticos de la ciudad de Nueva York”
El presente trabajo tiene como objetivo el análisis de las interferencias ortográficas y ortotipográficas de la lengua inglesa que se dan en las traducciones al español de un corpus de textos museísticos de la ciudad de Nueva York. Dicho corpus, que es paralelo y alineado, se compone de 151 registros en inglés (381 074 palabras y 452 424 tokens) y el mismo número de registros en lengua española (429 019 palabras y 499 895 tokens). Tras una breve descripción del diseño y la compilación del corpus, se analizará el grado de influencia que tiene la lengua de origen de los textos del corpus (el inglés), en la lengua de destino (la lengua española) en lo relativo a los elementos ortográficos y ortotipográficos más significativos. Se estudiarán, entre otros, errores mecanográficos, cuestiones de acentuación, signos de puntuación —comas antes de conjunciones coordinativas, signos de exclamación e interrogación— y empleo de mayúsculas y minúsculas.

29. López Cuso, Mª José and Javier Pérez Guerra. “A corpus-based analysis of negative contraction in the recent history of English: This isn’t and is not an easy goal!”
The first clear orthographic signs of negative contraction in English (e.g. is not > isn’t) go back to the late sixteenth or early seventeenth century (cf., among others, Brainerd 1989/1993; Denison 1993: 309; Warner 1993: 207-208; Mazzon 2004: 62; 104-105), somewhat later than the first occurrences of auxiliary cliticization or verb contraction (e.g. he is > he’s; I will > I’ll). It is very likely, however, that such contracted forms were actually in use in speech long before they appeared in print with clitic spellings. In the following centuries, negative contractions were the object of the fierce campaign against contracted forms of all kinds initiated in the early eighteenth century by Jonathan Swift and Joseph Addison. Such reduced forms were openly criticized in grammars and spelling books of the time as being improper, vulgar, colloquial, inelegant, barbarous, harsh, and even ungrammatical, especially in writing, except for poetic use (cf. Sundby et al. 1991; Haugland 1995). In spite of this “war upon contractions” (Leonard 1929: 170), however, the Late Modern English period witnessed a considerable increase in the use of contracted forms in published material, an expansion which has continued to the present day. The aim of this presentation is to examine the role that different factors played in the selection of negative contractions and their uncontracted counterparts in eighteenth, nineteenth, and twentieth-century written English. For our purposes, data have been drawn from A Representative Corpus of Historical English Registers (ARCHER). Along the lines suggested by Kjellmer (1998), the selection of contracted and uncontracted negatives in Late Modern English and twentieth-century English is taken here to be the result of the interplay of promoting and inhibiting forces of various kinds, involving both structural and semantic factors together with aspects of style and dialect. The range of variables considered for the analysis includes, among others, the following: geographical variety (does American English show a stronger preference for contraction than British English?), register/text-type (are contractions more closely associated with speech and informal text-types than with formal writing?), verb form (do all contractible V + not sequences undergo contraction to the same extent?), verb function (e.g. does main verb be contract as frequently as auxiliary be?), subject type (does negative contraction behave alike with pronominal and non-pronominal subjects?), clause type (e.g. are contracted negatives more frequent in interrogatives or in imperatives than in declaratives?), and position (e.g. does clause-initial or clause-final position of a contractible V + not sequence favor contraction?). In an attempt to corroborate or refute the null hypothesis ‘negative contraction is not conditioned by the variables considered in this investigation’, we have applied a logistic regression analysis (functions ‘glm’ and ‘lmr’ in R) to the database so as to determine the list of variables with greater explanatory power within the model (Nagelkerke R2 and Concordance indexes). The interpretation of the results has also been supported by the ranked grouping of the main factors provided by the ‘chaidplot’ function.

References

87. Losey León, Mª Araceli. “Parallel Corpora-based Terminology Practice Tool for Interpreters in the Maritime Shipping Business Field”
In specialised communication an interpreter’s constraints in specific training on terminological, conceptual, prototypical text-types and contextual awareness can hinder an effective information delivery. This can get worse in any of the professional sectors where business communication is involved since linguistic misunderstandings may endanger the favourable prospects of commercial transactions. Maritime transport agents, stakeholders and related professionals often require the rendering of both simultaneous and consecutive interpretation
categories’ services. Maritime shipping business is a vast sector in the maritime industry that uses a myriad of written documentation containing formal agreements and standards. On the other hand, these documents are imbued with legal English terminology and phraseology so that further insights into the communicative situations and the adjoining typical genres and text types can be additionally required from the interpreter’s performance.

The main object of this study is to describe the process of parallel corpora exploitation towards the development of an English-Spanish bidirectional glossary to assist interpreters during their preparation and terminology training stage. It is also expected that the resulting product can help the interpreter along the entire workflow, from the practising stage to the interpreting performance in the booth or at the meetings. The main premises of this study are that Corpus linguistics can provide particular support to interpreters in the preparation or documentary stage and that a preliminary genre-based (Swales 1990, Bhatia 1993) text type identification is essential to feed the corpus substantially and provide a consistent specialised product. Using the methodological framework of Corpus Linguistics, the present work attempts to describe all the process from the documents’ selection and classification to the corpora compilation, corpora processing, term extraction and the glossary production. Advantages in interpreter’s training and competences in documentation skills using corpus methodology have been addressed by previous studies (Pérez Pérez 2013, Sánchez Ramos 2017, Seghiri 2017). Benefits from the implementation of corpus-based technology-assisted interpreting have also been pointed out by several scholars (Costa, Corpus & Durán 2014; Corpus and Durán 2017).

This study has been carried out within the framework of project VIP, Voice-text Integrated system for Interpreters, (ref. no. FFI2016-75831-P, MINECO) whose goal is to develop a web portal 2.0 with resources and computer-assisted interpreting tools intended to cover interpreter’s needs.

References

Mańczak Wohlfeld, Elżbieta and Alicja Witalisz. “Usefulness of the National Corpus of Polish for the study of Anglicisms”
The article offers an assessment of the National Corpus of Polish (NKJP) in reference to its usefulness for the study of English loans. We put forward a controversial claim that the National Corpus of Polish is not an adequate source for retrieving Anglicisms attested in Polish, since it lacks a tool to search for them. Also, it has been noted that the newest Anglicisms, e.g. *selfie* and *hashtag/hastag*, appearing with high frequency in both written and spoken Polish, and some of the well-assimilated English loanwords found in the mass media, e.g. *crowdfunding, followers, snapchatowanie, sekswerkerka*, are not included in the Corpus. This is accounted for by the fact that the Corpus was last updated in 2012. Secondly, while the National Corpus of Polish comes useful in verifying the use of the already identified (pre-2012) English loans of various types and in determining their frequency, as well as in finding KWIC concordances that display graphic (and, indirectly, phonetic) variants, morphological adaptation, inflectional forms, loan-based derivatives, contexts of use, collocability, and the co-existence of loans of various types, the tools provided in the corpus search engines fail to meet the requirements of more in-depth analyses of Anglicisms, which, generally speaking, are related to semantics and structure. For instance, no separate concordances can be obtained for English-sourced idiomatic loan translations which happen to be homonymous to Polish loose syntactic phrases. For lack of an efficient tool to find some structural patterns, the Corpus does not prove useful for finding instances of native compounds that have been composed with the use of an English-sourced word-formation rule. The instances of the Corpus deficiencies will be illustrated with authentic attempted-but-failed corpus searches. The NKJP-aided analysis of Anglicisms then is largely corpus-based, rather than corpus-driven, as the corpus is used to test hypotheses based on previously made assumptions regarding Anglicisms, and the researcher has to engage in a painstaking manual analysis.

References
NKJP - Narodowy Korpus Języka Polskiego [National Corpus of Polish]. Available at: http://nkjp.pl/.

En el marco del proyecto interuniversitario FFI2013-45769-P que lleva por título Combinaciones fraseológicas del alemán de estructura [PREP. + SUST.]: patrones sintagmáticos, descripción lexicográfica y correspondencias en español, la presente comunicación tiene por objeto presentar las diferentes aplicaciones de la lingüística de corpus en el ámbito de la fraseología. En concreto, se estudiarán concomitancias y divergencias de las combinaciones usuales del español y alemán de estructura [PREP + (DET) + SUST] con valor temporal, partiendo respectivamente de su comportamiento textual en los corpus de Sketch Engine eseuTenTen11 y deTenTen 13. Para ello se analizarán diferentes grados de fijación en los planos semántico y pragmático-textual con el fin de determinar la fijación externa (colocados verbales y sustantivos que aparecen antes y después del nodo) de los fraseologismos españoles de estructura [de + SUST]. Estas unidades fraseológicas presentan valor temporal y los sustantivos integrantes designan en su significado inicial movimientos reflejos o impulsos irreflexivos que tienen lugar en un espacio muy breve de tiempo: esp. de repente, de pronto y de golpe, y sus correspondencias al alemán auf einmal, mit einem Schlag; aus dem Stegreif, entre otros.
Stative possessive HAVE has a long history in the English language and has been subject to several changes over the centuries: the innovation of periphrastic DO and its eventual restriction to imperative, interrogative and negative declarative contexts; insertion of got; and later loss of HAVE in the got-construction. These innovations emerged over a long period between the fifteenth and nineteenth centuries (e.g., Crowell 1959; Jespersen 1961; Visser 1963-73; Tagliamonte, D’Arcy & Jankowski 2010; Lorenz 2016). They show very different distributions across international varieties of the English language (e.g., Nelson 2004; Kallen 2013:108-109), and their use varies according to linguistic factors, including: sentence type and polarity, type of subject and object, generic/non-generic reference, temporary/inalienable possession, contraction, etc. (e.g., Tagliamonte 2003; Lorenz 2016).

This is a diachronic study of stative possessive HAVE in Irish English from the eighteenth to the twentieth century, based on data from the Corpus of Irish English (Hickey 2003), which provides data from literary texts, largely dramas, containing representations of spoken Irish English, and the Corpus of Irish English Correspondence, which contains personal letters. These data sources permit comparison with studies of Late Middle and Early Modern English correspondence (e.g., Lorenz 2016), nineteenth-century correspondence from New Zealand, Australia, Great Britain and the USA (e.g., Hundt 2015), and twentieth-century British and American dramas (Jankowski 2008). The Irish corpora thus enable us to trace developments in the use of stative possessive HAVE from the earliest days of Irish English to the mid-twentieth century, and to compare usage in Ireland with some of the other major Englishes around the world. While in some respects, Irish English was ahead of the game, for instance in relation to the adoption of negative DO periphrasis with HAVE, in others it has apparently chosen to stand on the sidelines, as with the got innovations.

References


Esta propuesta de comunicación se basa en que los términos, además de caracterizarse como unidades gramaticales de triple vertiente (formal, conceptual y funcional), son también unidades pragmáticas de comunicación que aparecen en discursos específicos, realizados por individuos de diferentes características y en situaciones concretas de comunicación.

En este sentido, aquí estudiamos cómo la teoría lingüística se ha ocupado fundamentalmente de la gramática de las lenguas en general y ha construido su objeto de estudio a partir de una abstracción que no responde a ninguna realidad determinada. A tal efecto determinamos cómo la lingüística actual ha prescindido, así, de la diversidad en que se materializa cada lengua y ha obviado los aspectos comunicativos que estas evidentemente poseen.

Así las cosas, en esta comunicación studiamos cómo para cubrir este gran vacío, junto a la lingüística teórica hay ido surgiendo paralelamente algunas propuestas disciplinares que se ocupan del lenguaje dentro de un corpus cualquiera considerándolo como herramienta de expresión y comunicación.

References

89. Mora, Natalia. “Linguistic and computational aspects of Appraisal annotation: The case of mobile application reviews in English and Spanish”

Evaluation, opinion and subjectivity in discourse are related phenomena which are receiving much attention both in the Natural Language Processing (NLP) and in the Corpus Linguistics (CL) communities, from different approaches and perspectives. In the CL community, one of the most influential theoretical frameworks is ‘Appraisal’, a model for the description of linguistic resources used for negotiating emotions, judgements and evaluations (see Martin 2000; Martin and White 2005). In ‘Appraisal’ three major categories are proposed: ‘Attitude’, ‘Engagement’,
and ‘Graduation’. These have been fruitfully used both in theoretical and in NLP studies (see Taboada and Grieve 2004; Whitelaw et al. 2005; Read y Carroll 2012a, 2012b; Bloom y Argamon 2010, Neviarouskaya and Aono 2013, inter alia).

In spite of this work, the descriptive power of ‘Appraisal’ model for the characterisation of texts in terms of evaluation, opinion and similar notions, has not been exploited yet to its full potential in the NLP community. It is necessary to develop high-quality corpora annotated with Appraisal features in different languages in order to count on sufficient resources for automated processing. In this paper we present an example of a recent research effort at creating a bilingual English-Spanish comparable corpus of mobile application reviews annotated with Appraisal features. Our dataset consists of an initial training set of fifty reviews of mobile applications, randomly compiled from the automatic extraction of thousands of reviews from Google’s Play Store using a crawler designed ad hoc. The training set was filtered to ensure equal distribution of reviews in English and Spanish, similar length, and negative (1 and 2 star ratings) or positive (4 and 5 star ratings) polarity and app category (applications, games, films, books and music). This training set was used to perform agreement studies to validate the annotation scheme and guidelines of Appraisal features in English and Spanish. In addition, a larger dataset of two hundred reviews was also used as test data for the annotation of the validated Appraisal tags by one single annotator, also randomly extracted from Google’s Play store and containing the same distribution as the training set, both in terms of the languages used, the length, the polarity and the type of application. The annotation results reveal interesting genre-specific, language-specific and product-specific tendencies in the use of Appraisal features in these reviews. As to the genre-specific features, the preferred features are those expressing ‘Attitude’, more specifically ‘Appreciation’, and ‘Engagement’, while Graduation seems to be much less used in this genre. This is probably a reflection of the communicative purpose of this genre, i.e., presenting users’ opinions about a given product. As to the language-specific features, the Spanish reviews use ‘Attitude’ features much more frequently than the English ones, since the Spanish culture favours the expression of feelings and qualities more frequently than the English one. As to the tendencies observed in the review of different products, we observed that reviews of games and applications use more often categories such as ‘happiness’, ‘security’, and ‘satisfaction’, while book and film reviews use more frequently negative categories such as ‘unhappiness’ or ‘disatisfaction’. In spite of the preliminary results and the difficulties in achieving a reliable annotation with Appraisal features (see Mora 2017, Mora 2011), it is expected that the current work contributes to a wider use of the Appraisal model both in the CL and the NLP communities.

References


In this paper we describe our methodology to integrate domain-specific sentiment analysis for Lingmotif (Moreno-Ortiz, 2016), a lexicon-based system initially designed for general language texts. Sentiment has been shown to be subject-dependent to a certain extent. As a consequence, the problem is adapting a classifier built for analyzing general-language texts to a new specialized domain. Our approach to dealing with specialized domains is based on the idea of “plug-in” lexical resources which can be applied on demand. In addition to single word terminology, multiword expressions (i.e. idioms and compounds and local grammar patterns) are an essential part of specialized lexical resources, where their correct sentiment classification is crucial to fully convey and understand domain concepts and their relations (Choi, Kim and Myaeng 2009; Jijkoun; 2010; Moreno-Ortiz et al., 2015, 2016).

From a methodological point of view, this contribution implements a study on the extraction of single words and multiword expressions in economic and financial language. This study has the twofold objective of approaching the methodological bases for the acquisition of a specialized corpus of news in the economy/finance domain from available web sources, and to obtain an approximation to the task of automatically extracting domain-specific expressions of sentiment to be used by our sentiment analysis software, Lingmotif. A simple, 3-step methodology is proposed to extract candidate terms from specialized corpora, which are then matched against our existing general-language polarity database to obtain a vast array of sentiment-bearing words whose polarity is domain-specific.

The acquired corpus, the Esmeraldas Corpus of Economic and Financial Contemporary English (ECECPE) is described, along with SentiEcon: Lingmotif’s financial and economic sentiment plugin lexicon. The former is a 12 million word corpus, which is the result of our compilation of economic news articles published between 2007 and 2015. On the other hand, SentiEcon is a complete specific-domain sentiment lexicon consisting on 5,400 specific-domain terms (single words and multi-word expressions) extracted from the corpus. Terms were extracted with Sketch Engine (Kilgarriff et al., 2014), which proved to be very useful for obtaining candidate terms and multiword expressions, which were then manually filtered by domain experts and, finally, assigned specific-domain sentiment scores. Test results show that our methodology is able to generate a comprehensive, fine-grained domain-specific sentiment lexicon. Finally, we discuss preliminary accuracy results of the application of our sentiment lexicon, compare them with other specific-domain sentiment analysis systems, and discuss our challenges and directions for future work.

References
46. Moreno Sandoval, Antonio. “La polaridad léxica en español oral: expresiones positivas y negativas en el corpus C-ORAL-ROM analizadas con LYNEAL”

El análisis de opinión y sentimiento es una tarea del Procesamiento del Lenguaje Natural que se ha hecho muy popular desde comienzos de esta década (Liu 2012, Moreno Ortiz et al. 2016). Se basa en la detección de expresiones positivas y negativas en los medios sociales (Twitter, Facebook) y en los blogs y comentarios en páginas de internet. Sin embargo, una metodología similar no se ha aplicado con frecuencia a otros tipos de corpus. El objetivo de esta investigación es mostrar una primera aproximación al uso de léxico positivo o negativo en un corpus de habla espontánea. Usaremos la parte española del corpus C-ORAL-ROM (Moreno-Ortiz & Fernández-Cruz 2015). Se trata de un corpus de conversaciones y monólogos espontáneos recogidos en 42 horas de grabación y conteniendo 320.000 palabras. Hay tres grandes subcorpus: el registro formal, el registro informal y el registro de los medios de comunicación, con variedad de temas y dialectos. Para cada hablante, están anotados los datos de sexo, nivel de estudios y edad, entre otros. Se realizaron sin plan previo, para garantizar la espontaneidad. No hay ningún sesgo explícito hacia ningún polo de la opinión. El corpus ha sido analizado con LYNEAL, una herramienta de gestión de corpus que incorpora búsquedas, análisis estadístico y representación gráfica de los resultados. El punto de partida ha sido la creación de una lista de 100 expresiones positivas y 100 negativas, elegidas de un diccionario de polaridad. Incluyen varias categorías (nombres, adjetivos, adverbios, verbos) y algunas expresiones multipalabra (“dar asco”). Aprovechando que el corpus está anotado con metadatos con rasgos sociolingüísticos, hemos comparado la aparición de estas palabras por nivel de estudios y edades. El análisis de opinión y sentimiento se ha realizado con LYNEAL para extraer características de la polaridad. Esta investigación nos permite ofrecer, con bastante seguridad, cuáles son las expresiones preferidas en español oral, objetivo último de esta investigación. Los resultados nos muestran, en primer lugar, que los hablantes prefieren claramente las expresiones positivas a las negativas, independientemente del sexo, edad o nivel de estudios. El análisis de Pareto nos ofrece la relación 71:1; 23:1, es decir, el 23% del léxico de polaridad aparece en el 71% de los casos. Esta observación nos permite ofrecer, con bastante seguridad, cuáles son las palabras positivas y negativas más características en el español oral.


This paper aims at offering a description of the way in which English-speaking women wrote when dealing with scientific matters during the outburst of ‘modern science’ (1700-1900). To this end, the work will revolve around two main focal points. On the one hand, the idea that corpus linguistics is an excellent methodology to carry out linguistic research and, on the other, that the selection of certain linguistic features for analysis can provide an interesting insight into intentional and non-intentional pragmatic functions in academic writing. With these two ideas in mind, the paper will analyse those texts written by women in several sub-corpora of the Coruña Corpus of English Scientific Writing, CC (Moskowich, 2011, 2015). This will be done, paying special attention to the Corpus of English Texts on Astronomy, Corpus of English Life Sciences Texts, Corpus of English Philosophy Texts and Corpus of History English Texts. These corpora, some of them in beta version, represent four disciplines, two pertaining to the field of the exact and natural sciences (astronomy, life sciences) and two to the field of the humanities (history and philosophy).

The analysis of the linguistic features in this material will probably allow some kind of textual characterisation (Quirk et al., 1985; Biber, 1988). The features selected for study are those expressing involvement (hedges, amplifiers, private verbs and pronouns). They will be used to measure to what an extent works by women may have some hidden pragmatic functions influencing the reader. The use of each of these linguistic elements will help us complete a portrait of women’s scientific discourse in late Modern English also taking into account the social context in which women of the Modern Age lived, their family background and access to education and science.

References


En este trabajo se ha analizado el estilo de Benito Pérez Galdós utilizando un enfoque de estilística de corpus. De forma más concreta, el análisis ha girado en torno a la proyección del discurso verbal de los personajes del universo galdosiano por parte de los narradores de las...
76. Núñez Nogueroles, Eugenia E. “What can CORPES XXI tell us about the thematic distribution of Anglicisms in the Spanish press?”

The study of Anglicisms in different recipient languages has a long tradition. However, only recently, a corpus-based perspective has been employed to deal with the analysis of English loanwords. One of the first pieces of research devoted to this phenomenon is the volume entitled *The Anglicization of European Lexis* (Furiassi et al., 2012), which contains a selection of papers delivered at a seminar of the 10th International Conference of the European Society for the Study of English (ESSE), held in Turin (Italy) in 2010. The chapter covering Anglicisms in the Anglicization of European Lexis (Furiassi, Cristiano, Virginia Pulcini, y Félix Rodríguez González, eds. 2012. *The Anglicization of European Lexis*. Amsterdam / Philadelphia: John Benjamins Publishing Company), Lorenzo Criado, Emilio. 1996. *Anglicismos hispánicos*. Madrid: Gredos. Oincs-Martínez, José Luis. 2012. «Newly-coined Anglicisms in contemporary Spanish. A corpus-based approach», in *The Anglicization of European Lexis*, 217-38. Amsterdam / Philadelphia: John Benjamins Publishing Company.


The study of Anglicisms in different recipient languages has a long tradition. However, only recently, a corpus-based perspective has been employed to deal with the analysis of English loanwords. One of the first pieces of research devoted to this phenomenon is the volume entitled *The Anglicization of European Lexis* (Furiassi et al., 2012), which contains a selection of papers delivered at a seminar of the 10th International Conference of the European Society for the Study of English (ESSE), held in Turin (Italy) in 2010. The chapter covering Anglicisms in Spanish within this volume, by José Luis Oincs-Martínez, makes use of the *Corpus Diacrónico del Español* –CORDE– (Diachronic Corpus of Spanish) and the *Corpus de Referencia del Español Actual* –CREA– (Present-Day Spanish Reference Corpus), both compiled by the Real Academia Española (RAE). They constitute extremely useful resources for the purpose of studying the introduction of Anglicisms into the Spanish language, since they fill a gap highlighted by several authors (such as Lorenzo Criado, 1996), i.e. the lack of real and reliable data. In fact, the use of corpora when studying the presence of English loanwords in a receiving language permits the researcher to carry out frequency analyses and to deal with this linguistic phenomenon in a wider scenario, which is evidently an improvement on previous research projects that could not adopt these methods. However, these tools have not been made the most of yet in relation to the area of borrowing. Therefore, in order to contribute to the field, the present short paper makes use of the last reference corpus that is being compiled by the RAE: the *Corpus del Español del Siglo XXI* –CORPES XXI– (21st Century Spanish Corpus). Given that this electronic resource offers the option of selecting certain search criteria, it is possible to refine the query in order to look for a word in subcorpora we are particularly interested in as opposed to in the whole corpus. In view of this, I carry out a corpus-based study in which I search ten Anglicisms in written texts originating in Spain, characterised as non-fiction and belonging to any textual typology appearing in the press. These ten Anglicisms have been selected because they are the most frequent ones (out of 2,198) found in the CREA (specifically, in the Spanish press published in Spain from 2001 to 2004 and covering all topics). In the present piece of research I aim to explore the thematic distribution of these ten loanwords in relation to their number of occurrences and the number of documents in which they are employed. In light of the findings, some possible reasons for explaining the results obtained are ventured, which helps to appraise the overall situation of these ten recurrent Anglicisms in 21st century Spanish press. For instance, the high percentage (86%) of occurrences of the word club in the field “actualidad, ocio y vida cotidiana” can be explained by the fact that “sports” –a sphere where this Anglicism is very common– is included within this thematic area.

References:


76. Núñez Nogueroles, Eugenia E. “What can CORPES XXI tell us about the thematic distribution of Anglicisms in the Spanish press?”

The study of Anglicisms in different recipient languages has a long tradition. However, only recently, a corpus-based perspective has been employed to deal with the analysis of English loanwords. One of the first pieces of research devoted to this phenomenon is the volume entitled *The Anglicization of European Lexis* (Furiassi et al., 2012), which contains a selection of papers delivered at a seminar of the 10th International Conference of the European Society for the Study of English (ESSE), held in Turin (Italy) in 2010. The chapter covering Anglicisms in Spanish within this volume, by José Luis Oincs-Martínez, makes use of the *Corpus Diacrónico del Español* –CORDE– (Diachronic Corpus of Spanish) and the *Corpus de Referencia del Español Actual* –CREA– (Present-Day Spanish Reference Corpus), both compiled by the Real Academia Española (RAE). They constitute extremely useful resources for the purpose of studying the introduction of Anglicisms into the Spanish language, since they fill a gap highlighted by several authors (such as Lorenzo Criado, 1996), i.e. the lack of real and reliable data. In fact, the use of corpora when studying the presence of English loanwords in a receiving language permits the researcher to carry out frequency analyses and to deal with this linguistic phenomenon in a wider scenario, which is evidently an improvement on previous research projects that could not adopt these methods. However, these tools have not been made the most of yet in relation to the area of borrowing. Therefore, in order to contribute to the field, the present short paper makes use of the last reference corpus that is being compiled by the RAE: the *Corpus del Español del Siglo XXI* –CORPES XXI– (21st Century Spanish Corpus). Given that this electronic resource offers the option of selecting certain search criteria, it is possible to refine the query in order to look for a word in subcorpora we are particularly interested in as opposed to in the whole corpus. In view of this, I carry out a corpus-based study in which I search ten Anglicisms in written texts originating in Spain, characterised as non-fiction and belonging to any textual typology appearing in the press. These ten Anglicisms have been selected because they are the most frequent ones (out of 2,198) found in the CREA (specifically, in the Spanish press published in Spain from 2001 to 2004 and covering all topics). In the present piece of research I aim to explore the thematic distribution of these ten loanwords in relation to their number of occurrences and the number of documents in which they are employed. In light of the findings, some possible reasons for explaining the results obtained are ventured, which helps to appraise the overall situation of these ten recurrent Anglicisms in 21st century Spanish press. For instance, the high percentage (86%) of occurrences of the word club in the field “actualidad, ocio y vida cotidiana” can be explained by the fact that “sports” –a sphere where this Anglicism is very common– is included within this thematic area.

References:


98. O’Donnell, Mick, Penny MacDonald and Oksana Polyakova. “Applying an error corpus to the classroom: from coarse-grained errors to teachable concepts”

Most error corpora tag errors only to a certain degree of delicacy (Dagneaux et al., 1998; Neff et al., 2007; Díaz-Negrillo, 2009; MacDonald, García Carbonell and Carot Sierra, 2013). Some only tag a handful of categories. Most tag with less than a 100 or so categories. In the Trecacle Project (MacDonald, Murcia, Boquera and Stuart, 2011; O’Donnell, 2012) we tagged a Spanish university EFL learner corpora using around 130 categories, which at the time, we thought was a rich degree of detail to allow us to identify the key problems of our learners.

In stage two of our research, we are using this error-tagged corpus to identify the grammatical concepts that need to be taught in the classroom. One thing we noticed quickly was that each of the error categories we identify does not by itself correspond to a single concept that the students need to be taught. For instance, the most common error in our corpus was article-inserted-not-required (e.g. “the drugs are a problem for the society”). However, one cannot simply say to learners: use fewer articles! As a teacher, we need to make clear the different
contexts in which the article is required and where it should not be used. This discovery led us to a more fine-grade coding of the article-inserted-not-required error in terms of the various contexts of use in which it is misused (e.g., generic-plural, generic-mass-noun, meal-names, percentages, places-of-work-etc.). By identifying the exact rules that govern the use of this form allowed us to discover which of them produces the most errors, and thus which should be taught to our learners.

It also provided us with concepts which are teachable: while one can’t teach on the basis of the general error, the more specific tags for this category relate to teachable concepts, such as “don’t use an article in front of a plural generic reference”, or “don’t use an article in front of a percentage” (“the 30% of my friends”) etc.

This is only a small finding by itself, but the same seems to be true for almost all error tags in our tag scheme. In nearly all cases, to derive teachable concepts from our error coding, we have needed to apply a more fine-tuned coding to each error, identifying the distinct sub-rules (distinct contexts of use) that produce something that, on the surface, looks the same. For instance, when exploring the set of subject-finite agreement errors, we see many cases where students do understand that the Subject and Finite need to agree in number, but still get it wrong occasionally. One big source for Spanish learners of English is that they take “people” to be singular (The Spanish equivalent, gente, is singular). Thus, to teach in the area of subject-finite agreement, one not only needs to teach the concept that the Subject should agree with the Finite, but also, that “people” is plural, that the Subject in an existential clause follows the verb, etc.

Overall, the paper argues not only for more fine specification of error categories, but also that these categories should correspond to teachable concepts.

References


55. Orrequa Barea, Aroa and Cristian Marín Honor. “Aspectos metodológicos de la extracción de contenidos generados por usuarios de Internet en la compilación de corpus” En los últimos años, el análisis de sentimiento o minería de opinión se ha convertido en un importante objeto de estudio en la lingüística computacional (Mankar e Ingle, 2005) y del procesamiento del lenguaje natural (NLP) (Schmunk et al., 2013). Debido al aumento del uso de redes sociales y a la gran presencia que tiene Internet en nuestras vidas hoy en día, se ha convertido casi en una costumbre leer opiniones de servicios o productos antes de consumirlos, así como evaluarlos una vez usados con el fin de tener una mejor experiencia. Esto es lo que se conoce como eWOM (electronic word of mouth) que proporciona información fiable de consumidores externos a las empresas y que se propaga fácilmente a través de Internet, estando así al alcance de todos (Goldsmith y Horowitz, 2010 y Schindler y Bickart, 2007).

En este estudio nos centramos en los aspectos técnicos y metodológicos de bajo nivel para la elaboración de un corpus de contenidos generados por usuarios (UGC) (Marine-Roig y Anton Clave, 2015) en páginas web relacionadas con el turismo, concretamente en las opiniones de clientes sobre una empresa de alquiler de vehículos sin conductor (Malagacar.com). Dado que el corpus propuesto pretende ser empleado para estudios de análisis de sentimiento, este primer paso resulta enormemente importante, pues determina el tipo de material constituyente del corpus el cual condiciona cualquier proceso posterior, como es el de anotación. Así, enfocamos nuestro análisis a dos de las tareas del modelo de implementación de Hippner y Rentzmann (2006) que son (i) descarga semi-automática de datos de páginas web y (ii) limpieza de texto con expresiones regulares.

En la primera tarea, usando el entorno de trabajo Scrapy, crearemos una “araña”, esto es, una serie de instrucciones que programaremos para extraer la información de la página web y almacenarla posteriormente en diferentes formatos (csv, txt, JSON). Al elaborar la “araña”, podremos decidir qué elementos del comentario queremos incluir en nuestro corpus, desde el texto como tal, hasta el autor, la fecha o incluso la puntuación que cada cliente proporciona. En nuestro caso, vamos a elegir el texto, la fecha, el origen y la puntuación. Consideramos importantes estos metadatos, pues, aparte de servir como referencia, la puntuación se utiliza en análisis de sentimiento como gold standard para medir el rendimiento del software y también para aprender la expresión positiva/negativa. Además, estos datos nos pueden servir para estudios posteriores de comparación entre el texto y la puntuación. La segunda tarea consistirá en la limpieza y formato de los comentarios obtenidos, empleando para ello un editor de textos y expresiones regulares. Los textos limpios y con formato adecuado serán clasificados para su posterior anotación mediante herramientas específicas, tales como BRAT o UAM Corpus Tool.

References


106. Oster, Ulrike. “The changing conceptualization of emotion words. Independent evolution or cross-cultural influence?” Do emotion concepts change over time? There have been completely different answers to this question in emotion research. On the one hand, universalists insist on the timelessness and pan-
culturalism of emotions, and on the other social constructivists believe that emotions are largely learned and subject to historical change (Plamper 2015). Cognitive linguistics, the framework this paper adheres to, places a strong emphasis on potentially universal aspects because one of its basic tenets is the recognition of the influence of the human body and mind’s functioning on conceptualization and categorization. However, it also recognizes the cultural specificity of emotion concepts (e.g. Kövecses 2005) as well as the intrinsic historicity of language (Geeraerts 2010: 333) and thus the gradual evolution of grammatical structures and concept configurations over time.

The relatively new field of history of emotions investigates the changing nature of emotion concepts from a historian’s point of view, tracing its causes back to the evolution of societies, cultures and values. So, if cultures and societies change, and if at times they inevitably do so under the influence of another culture, it seems plausible that emotion concepts in one language might be influenced by the conceptualization of an equivalent but not identical emotion concept of another language.

The aim of this paper is to establish whether, and to what extent, a basic emotion like fear might be conceptualized differently in different languages (English and German), if these conceptualizations change over time and whether or not the changes occurring in one language evidence a convergence with the tendencies of the other. Given the increasing communicative predominance of English over other languages in almost any field from science to popular culture, an influence of the conceptualization of English fear on German Angst is hypothesized.

In order to do achieve these aims, a corpus-based approach for the study of emotion words (Oster 2010) is combined with Ogarkova and Soriano’s (2014) concept of semantic foci. These foci (control, lack of control, visibility and internalization of the emotion) are used to organize and group the data obtained from a first phase of corpus analysis and to characterize the ways in which different cultures conceptualize given emotions and which aspects they foreground. Two large corpora of contemporary American English (COCA) and German (DWDS) are used to describe differences and similarities between the lexical items “fear” and “Angst” from a synchronic perspective. Then, with the help of the Corpus of Historical American English (COHA) and the Deutsches Textarchiv (DTA), we go on to trace how the relative importance of these aspects changes over the centuries.

The results of the analysis reveal several aspects in which the data for German Angst tends to converge with the English usage, which allows to conclude that the influence of another culture or language can be considered a contributory factor in diachronic variation. However, there is also evidence for independent, language-specific evolution in cases in which language contact is clearly not responsible for the changes.

References

15. Palacios Martínez, Ignacio M. “Language Contact and Language Change in Multicultural London English”

The majority of urban communities in the twenty-first century cannot be any longer regarded as monolingual since people from different origins, cultural and religious backgrounds tend to live together and interact in them due mainly to personal reasons and migrations resulting from natural disasters, war, political prosecutions, ethnic disputes or economic recession. This has had a clear effect on the linguiscape of these cities and, as a result, new linguistic practices have begun to emerge, notably in youth communication, which are generally known as contemporary urban vernacular (Rampton 2015), urban vernacular and urban youth speech style (Wiese 2009, Nortier & Svendsen 2015, Cheshire, Nortier & Adger 2015).

In the particular case of London a new variety of English has appeared, known as “Multicultural London English” (MLE) (Cheshire et al. 2011), which contains features from second language learners of English and from Afro-Caribbean Englishes, together with some other elements of local London and south-eastern varieties, to the extent that this new sociolinguist has been coined as the new Cockney (Fox 2015). Examples of these are seen/seen ‘OK’, whab ‘what’s going on’, yute ‘kid’ (from ‘youth’). The plural suffix -dem attached to personal nouns (boydem, mandem, peopledem) has emerged, its origin being in the speakers of Caribbean varieties. The same applies to the use of man as a pronoun of first and third person singular (e.g. before I got arrested man paid for my own ticket to go Jamaica). In addition, we find words from Arabic (akh as a short form for ‘brother’), Hindi (bandana, a kind of rag), Spanish (tapas bar, amigo), Italian (mafia, bravo), French (coup de mouche) and even Yiddish (schmuck).

Apart from this, we also find examples of innovation and change in grammar and discourse, such as the change of meaning of some words, for example, the adjectives wicked, sweet and sick with the meaning of great or good, the new quotative this is + pronoun (this is him; “blah”), the use of proper and bare as intensifiers (e.g. they were proper strict in school; the game is heavy, it’s bare addictive), vernacular negatives (ain’t, third person singular present don’t), negative concord structures (e.g. I don’t want nothing to do with you no more) and a large proportion of placeholders (e.g. I went to a thingy), general extenders (e.g. we had a street party and everything), invariant tags (e.g. we got two separate rooms innit) and address terms (bruv, dude, etc).

The findings will be drawn from the London English Corpus, which contains the Linguistic Innovators Corpus and the Multicultural London English Corpus, amounting to over 2,300,000 words together with material from additional sources, including dictionaries, glossaries, forums for teenagers, plus a questionnaire.

The analysis of all this data here will serve to understand better what is in fact happening in the communities of these speakers, and how change is emerging. Moreover, it may be possible to extrapolate some of the findings to similar groups of speakers in other European contexts.

References
Pennock Speck, Barry and Begoña Clavel Arroitia. “Using questionnaires and focus groups to gather quantitative and qualitative information on teachers’ perspectives in telecollaboration” This paper draws on the results obtained from a survey of 179 secondary school foreign language teachers and a series of focus group interviews (FGIs) in the context of a European telecollaborative project, TeCoLa (“Pedagogical differentiation through telecollaboration and gaming for intercultural and content integrated language teaching” -2016-1-NL01-KA201-022997). The main objective of the 36-question survey and the FGIs was to identify the amount of training that teachers would need regarding the tools to be employed in the project, video conferencing tools (VC), online games (OG) and virtual worlds (VW). These instruments were also employed to provide insights on how to design teacher development modules and teacher-training sessions and workshops.

Previous research (Dörnyei, 2003, 2007; Wilson & Dewaele, 2010; Wilson and McLean 1994; Cohen, Manion and Morrison, 2007) shows that questionnaires constitute a useful instrument for collecting data that can be combined with other sources of information like interviews and focus groups. FGIs, on the other hand, are particularly handy as a tool to elaborate the responses obtained in questionnaires and to help towards a better understanding of the opinions, beliefs and experiences of a group of informants (Al Ghazali, 2014; King, 2004; Gonzalez, 2011). As Gaižauskaitė (2012: 21-22) claims, “usually, focus groups are efficient in revealing covered, unexpected, or new information upon research questions” (see also Ansay, Perkins and Nelson, 2004; Engel and Schutt, 2010; Rubin and Babbie, 2010).

From the results of the questionnaire, which was made up of Likert-scale and opinion questions, we found that diversity and learning challenges and intercultural awareness and competence were very relevant to teachers while communicative competence was extremely relevant. This was not the case with Content and Language Integrated Learning (CLIL) probably because the majority of the respondents teach languages and are not engaged in CLIL itself. In answer to the questions on which of the above aspects, diversity, learning challenges, etc., are problematic, CLIL scored the lowest. With respect to their experience with telecollaboration tools, particularly related to VC, OGs and VWs, the data show that the respondents had very little or no experience. This was particularly so in the case of VWs. All in all, our tentative conclusions are that more training is needed in the case of OGs and VWs than in the case of VC. Regardless of their lack of experience, teachers seem to be clear about the potential benefits of the tools in all the topics and particularly in the case of communicative competence. In the open-ended FGI discussions, the participants identified the main challenges in their teaching context and explored the ways in which TC could help them cope with those challenges. The data obtained in the FGIs filled in the gaps of the quantitative survey data.

The advantage of the double-pronged approach to data collection, quantitative and qualitative is that it provides us with a more complete picture of teacher’s needs.

https://sites.google.com/site/tecolaproject/home-1

64. Pérez Béjar, Victor. “La elaboración de un corpus experimental de fonética acústica como herramienta de la macrosintaxis” En este trabajo queremos presentar la elaboración de un corpus experimental con grabaciones de audio obtenidas en el entorno controlado de una cabina de grabación para su posterior análisis acústico. Las grabaciones consisten en una serie de lecturas dramatizadas que emulan conversaciones coloquiales en las que se incluyen elementos críticos cuya entonación es lo que se somete a análisis.

Este tipo de corpus lo elaboramos como herramienta para caracterizar los límites de los enunciados en la organización textual, con especial atención a aquellos que terminan en una inflexión final en anticadencia, también llamados enunciados suspendidos. Estos nos sirven de ejemplos para demostrar la eficacia de un corpus experimental dramatizado en los estudios sobre estructura del discurso.

El marco en el que se desarrolla este estudio parte de la idea de que existe una sintaxis más allá de la estructura oracional y a la que se le suele dar el nombre de macrosintaxis (Berrendonner 1991; Blanche-Benveniste 2000; Gutiérrez Ordóñez 2016; Fuentes Rodríguez 2017a, 2017b). En este ámbito, surge cada vez más la necesidad de un análisis acústico de la prosodia del material oral, ya que la entonación constituye un rasgo indesligable de los límites de las unidades de la macrosintaxis. Muchos autores han propuesto modelos de segmentación discursiva basados en criterios acústicos: Cresti y Firenzueli 2001; Blanche-Benveniste 2011; Dehé 2014; Hidalgo Navarro 2016...

En nuestro caso, el corpus surgido de conversaciones fingidas y teatralizadas ofrece un material muy valioso que permite emular la melodía del habla espontánea, manteniendo las ventajas de un corpus controlado. Puede ser un complemento a las grabaciones de conversaciones reales, ya

References


que nos ha demostrado su eficacia en cuanto a análisis prosódico se refiere. Planteamos, así, demostrar una serie de ventajas en el análisis:

- Permite aislar variables con mayor eficacia que los extraídos de grabaciones reales, ya que es posible comparar dos variantes dadas en un mismo contexto.
- Permite buscar patrones acústicos habituales en una muestra representativa de hablantes.
- Las lecturas dramatizadas, además de fomentar la naturalidad de la conversación, permite la ocultación eficaz del elemento crítico.
- La calidad acústica de los ejemplos obtenidos en una cabina de grabación es notablemente superior a las grabaciones orales reales, que suelen tener mucho ruido de fondo.

Por el momento hemos recopilado un total de 36 grabaciones correspondientes de 18 textos distintos. Los textos de las lecturas, aunque adaptados, están basados en ejemplos reales de otros corpus, principalmente, el CREA, el CORPES XXI, el COFA o el corpus Val.EsCo.

Para el momento hemos recopilado un total de 36 grabaciones correspondientes de 18 textos distintos. Los textos de las lecturas, aunque adaptados, están basados en ejemplos reales de otros corpus, principalmente, el CREA, el CORPES XXI, el COFA o el corpus Val.EsCo. De hecho, Nicole (2014) y Fuentes Rodriguez, catalina (2016) apuntan que una metodología y un material indispensable en la gramática de los enunciados y del texto.

Según Cresi, Emanuela y Valentina Firenzuoli (2002) "L'articolazione informativa topic-comment e comment-

Para el momento hemos recopilado un total de 36 grabaciones correspondientes de 18 textos distintos. Los textos de las lecturas, aunque adaptados, están basados en ejemplos reales de otros corpus, principalmente, el CREA, el CORPES XXI, el COFA o el corpus Val.EsCo. De hecho, Nicole (2014) y Fuentes Rodriguez, catalina (2016) apuntan que una metodología y un material indispensable en la gramática de los enunciados y del texto.

Por el momento hemos recopilado un total de 36 grabaciones correspondientes de 18 textos distintos. Los textos de las lecturas, aunque adaptados, están basados en ejemplos reales de otros corpus, principalmente, el CREA, el CORPES XXI, el COFA o el corpus Val.EsCo. De hecho, Nicole (2014) y Fuentes Rodriguez, catalina (2016) apuntan que una metodología y un material indispensable en la gramática de los enunciados y del texto.

Research on pragmatic borrowing has noted that English-sourced pragmatic and discourse elements such as interjections, discourse markers, expletives, vocatives, etc., are frequently borrowed into a variety of recipient languages in the current era (see Andersen, 2014). In many cases, there is a clear pragmatic motivation for borrowing such elements (Peterson, 2017). While previous research has found, for example, that the English borrowing *plis* ‘please’ has been borrowed into everyday Finnish discourse for purposes of positive politeness (in addition to other motivations; see Peterson & Vaattovaara, 2014), an uninvestigated area of borrowing into Finnish is vulgar or impolite language, namely swearing.

This research looks at the incorporation of English-sourced swear words into Finnish discourse, using data from the Finnish forum Suomi24, available as a corpus from FinClarIn ([https://www.kielipanikki.fi/languagenbank/]). Building on previous work, where we found a surprisingly high level of integration of English-sourced swear words in Finnish CMC discourse, we have supplemented our original findings by exploring some of the social meanings of English-sourced swear words in relation to heritage Finnish equivalents.

This presentation first highlights the findings from our initial stage of analysis, which was corpus based and showed that English swear words 1) are not supplanting Finnish forms 2) are as likely to be used in other grammatical functions as they are in particle function (i.e., showing integration into Finnish morphology), and, 3) are used in a highly performative fashion, for example in bilingual puns and with spontaneous localized meanings.

The presentation then goes on to discuss the second stage of analysis, currently underway, which makes use of an online matched guise survey in audio format to gauge attitudes and perceptions of English-sourced forms compared to heritage Finnish forms. The quantitative (multivariate) analysis of the data looks at the relationship of English swear words with social factors such as gender, age and vernacular vs standard Finnish.

These combined methods offer a robust overview of how the English-sourced swear words are integrated grammatically, pragmatically and socially into everyday Finnish.

References


62. Petterson Traba, Daniela. “Semantic types and synonym choice: Examining the colloccational preferences of attributive adjectives in Late Modern and Present-day American English”

Near-synonyms, linguistic constructions with similar yet not completely identical meanings, denote the same concept but contrast in nuances of denotational meaning, or in other dimensions such as style, connotation, and/or collocation (Cruse 2000:159–160). Notwithstanding its omnipresence in language, near-synonymy has not received the attention it deserves, and several scholars have set out the need for more research in this field (e.g. Divjak 2006). This has led to the recent emergence of studies focusing on the semantic structure of groups of near-synonyms, including synonymous verbs (e.g. Divjak/Gries 2006), nouns (e.g. Liu 2013), adjectives (e.g. Gries/Otani 2010; Liu 2010), and adverbs (e.g. Liu/Espino 2012). However, these studies have all considered the semantic structure of near-synonyms from a synchronic viewpoint, disregarding their diachronic development. The analysis reported on in this paper aims at partially filling this gap by examining the attributive uses of four adjectival near-synonyms in the recent history of American English, fragment, perfumed, and scented, symbolizing the concept SWEET-SMELLING:

(1) [...] Mr. Minturn had to sip the generous and fragrant beverage slowly (COHA, 1892)  
(2) When Kilo buys scented soap she likes to have it really scented (COHA, 1907)  
(3) When he raised himself out of the water he placed perfumed oil on his body (COHA, 1999)

This is done by drawing on data from COHA (Davies 2010) and by conducting a logistic regression analysis (Levshina 2015:277) to determine the effect of two predictors, period and semantic category, on the dependent variable ‘synonym’ (i.e. fragrant/perfumed/scented). ‘Period’ has four levels, namely four 50-year periods: 1810-1859, 1860-1909, 1910-1959, and 1960-2009, whereas ‘semantic category’ refers to the semantic type of the nouns the synonyms modify, since several scholars (Geeraerts 1986; Gries 2001) have proven head nouns to be useful in revealing fine-grained shades of meaning of adjectives. The modified nouns are here categorized into semantic types (e.g. ‘food & drink’, ‘aesthetics’, ‘cleaning’) by employing the semantic classification of the HTOEED.

Findings ascertain that both predictors are statistically significant for the choice between the three synonyms. First, while fragrant typically accompany nouns that refer to entities with a natural pleasant smell (e.g. *nouns in the category ‘food & drink’,* perfumed and scented modify nouns denoting man-made, artificial objects (e.g. those in ‘aesthetics’, ‘cleaning’). Second, the results show that scented is gaining ground at the expense of fragrant in the 20th century. This replacement might be explained by the fact that the concept SWEET-SMELLING undergoes a notable change throughout the period analyzed, with nouns denoting a natural smell being...
modified significantly less frequently in the second half of the 20th-century, and those designating an artificial smell becoming much more frequent. This might suggest that speakers nowadays need to refer to artificially scented soaps and oils much more than a century ago which, in turn, could reflect a change to a much more sophisticated society characterized by mass-production and commercialism.

References

COHA = Davies. Mark. 2010
2009. Available at: <https://corpus.byu.edu/coha/> [Accessed online on January 23, 2018]


52. Pizarro Sánchez, Isabel. “Online Product Descriptions: A corpus-based contrastive analysis of their visual elements”

Words used to be the major carriers of information but nowadays they share this crucial function with visual aspects or modes of information presentation (Bateman 2008) since written texts are increasing their degree of multimodality. Images, tables, diagrams, videos, and icons are key elements in many genres as they illustrate the texts and, together with words, shape their meanings (Kaindl 2004).

‘Online product description’ is an example of a genre that makes use of the aforesaid visual elements. Following Swales (1990) and Bhatia (1993, 2004), these texts can be classified as a genre and can be described in terms of their rhetorical structure (moves and steps). Although visual elements have become an important issue for researchers, there are not many studies on the frequency of occurrence of these elements nor on their rhetorical distribution. The aim of the present study is to carry out an English-Spanish contrastive analysis of the visual elements of this genre, and more specifically it sets out to i) identify the visual elements within this specific genre and classify them according to the following categories: image, table, diagram, graphs, video, icon; ii) describe the rhetorical distributions and frequencies of the visual elements; iii) identify the verbal elements (introductory texts) that link the visual elements to the texts; and iv) establish differences and similarities in their English and Spanish usage in terms of frequency and rhetorical distribution.

This research relies on the specialized “Corpus Comparable de Descripciones de Galletas” (Comparable Corpus of Biscuit descriptions) that comprises 100 complete texts in each language, roughly featuring 13500 running words in English and 15000 in Spanish; all texts were written in European Spanish and British or American English and were published in the last three years (2014-2016). The corpus is tagged at the rhetorical level by means of the ACTRES text tagger and examples were retrieved from the corpus by means of the ACTRES browser http://contraste2.unileon.es/web/en/tools.html that supports queries within specific rhetorical moves and steps.

Empirical results reveal that, in general terms, visual elements are more frequent in English product descriptions than in Spanish. The rhetorical distribution of the visual elements tends to be similar in both languages when frequencies are close, which could be due to the fact that ‘Product descriptions’ do not only have an informative function but also a persuasive one that is even more significant than the informative. The study is significant as it includes refined quantitative data on visual element types and their rhetorical distributions in English and Spanish. Results can be applied to the training of tertiary level learners, technical writers and professional translators.

References


Conrad’s well-known characters tend to be strong men who are the head of their families, whereas his female characters are denoted to the neck which supports that head of the household. However, in Heart of Darkness there is a character who can be understood as an exception to Conrad’s writing, Kurtz’s African woman. The way she is presented to the reader is not only shows that she is different to other Conrad’s female characters, but because she does not follow the pattern of the time. It is during post-colonial literature when women are the Angel of the House; sweet, delicate, weak; educated only in the arts meant to entertain; thinking and wondering were male privileges. In this paper I will pay special attention to the five female characters and their brief roles, although, special attention will be paid to Kurt’s wife as well as his mistress, and to Marlow’s aunt. I will scrutinize the colloquations used by Joseph Conrad and the way they are presented and described by other characters. Using corpus tools, I will compare and contrast four Spanish translation of the novel; the one by García Rios and Sánchez Araujo ([1976] 2005), other by Diéguez Rodríguez (2002), a third one by Pitol Deméneghi (2009) and the last one by Alberto Alba (2011). Specifically, I will analyse the extent to which the four Spanish versions render the writer’s intention when he uses specific colloquations that contribute to shaping gender identity. First, Conrad’s use of the nouns ”woman”, ”women”, ”lady”, ”girl” and ”wife”, and the pronoun ”she” in the novel will be measured using a corpus methodology.
Then, using an aligned parallel corpus containing the English text and the four Spanish versions, the Spanish translators’ rendering of the above mentioned terms will be systematically scrutinised. The aim of the analysis is to show whether microtextual shifts take place in the Spanish versions and, if so, their effects from a stylistic point of view.

Most studies on English inverted constructions base their classifications on the preliminary distinction between two main types of inversion: full-verb inversion, as in “Out of the blue came a letter from Sinclair”, and subject-operator inversion, as in “Never did he manage to provide evidence for his claim”. Two main types of SOP inverted constructions can be distinguished in English: verb-first inverted and verb-second inverted constructions. Verb-first inverted SOP constructions are those inversions in which the operator is the first syntactic constituent in the clause, as in (1). Verb-second SOP inverted constructions, by contrast, are those in which the operator is not the first syntactic constituent in the clause, as in (2).

1) Should you make a mistake, let me know.
2) Only then did he prepare for the fight.

In my presentation I will offer a functional analysis of the SOP inverted construction in Present-day English. The focus of the study is V-2 placement; hence, no cases of verb-first SOP inversions are included. The study builds on Goldberg’s (2006) well-known example of subject-operator inversion schema, but is restricted to the use of verb-second SOP inversion in declarative clauses. On the basis of a statistical analysis using corpora of written and spoken English, I will argue that the different verb-second SOP inversions in declarative clauses distinguished in the literature are clear instances of constructions as understood within the Construction Grammar framework (cf. Goldberg 1995). It will be shown that despite their formal and functional dissimilarities, these inversions still relate to one another in systematic and predictable ways and I will show that in fact the different SOP inversion types are grouped in relation to a unit in the schematic network which is naturally most salient—the prototype (cf. Langacker 1987: 492)—and form with it a family of nodes—extensions from the prototype (cf. Langacker 1991: 548)—in the system. In sum, the analysis here will show that SOP inverted structures in declarative clauses are constructions, that is, form-meaning correspondences which in themselves carry meaning, independently of the words in the sentence (cf. Langacker 1987: Ch. 2). It will also be shown that these constructions form an interconnected, structured system or network and are best understood with reference to different forms of inheritance (cf. Goldberg 1995: 72 ff).

References


The study of television dialogue is without a doubt a burgeoning field, as attested by the current wealth of studies, books and journal articles available in the recent literature: Bubel 2006; Quaglio 2009; Bednarek 2010, 2015, 2017; Richardson 2010; Queen 2015; Beers Fägersten 2016; and Gregori-Signes 2017, to name but a few.

This being said, the purpose of the present paper is to analyze and critically evaluate the linguistic construal and ultimately the representation of sexual identities in a corpus of television dialogue. As a point of departure, a series of issues bound to arise in a corpus linguistic approach to the study of television dialogue will be addressed, namely corpus design, representativeness, data compilation, transcription and annotation, among several other factors to be taken into consideration.

Further, in order to explore the practical affordances of corpus linguistics in the study of television dialogue, the second part of the paper will concentrate on a case study currently in progress which takes a CADS (Corpus-Assisted Discourse Studies) approach (Baker et al. 2008; Partington, Duguid and Taylor 2013) to a corpus of television dialogue comprising the transcripts of two American television series: Will & Grace (NBC, 1998-2006 – henceforth WG) and Queer as Folk (Showtime, 2000-2005 – henceforth QF).

The motivation behind this choice is threefold. To begin with, each series has two distinct groups of people as main characters: one consists of two gay men and two heterosexual women (WG) whereas the other comprises five gay men and two lesbian women (QF). Moreover, each series addressed and/or attracted a different audience: WG was watched by “a mainstream (i.e. mainly heterosexual) audience” (Baker 2005: 95), whereas it is safe to assume that QF had more gay and lesbian viewers than heterosexual viewers based on its content. Lastly, each series belongs to a different fictional genre: sitcom (WG) and drama (QF).

To date, initial frequency-based surveys of the data have revealed an overall similar rate of lexical items used to name, label or classify sexual identity (Cameron and Kulick 2003; Baker 2005) in both series, yet some striking differences and other interesting insights have also emerged. For example, some sexual identities (such as bisexual, transgender or pansexual identities) are consistently underrepresented in both series. More surprisingly, some slurs or pejorative terms are more frequent in QF than in WG, a case in point being fag/s or faggot/s, with a normalized frequency per 10,000 words (henceforth NF) of 0.33 in WG and 3.3 in QF. However, the most interesting finding so far concerns gay/s (WG: NF=11.46; QF: NF=8.22) and straight/s (WG: NF=2.17; QF: NF=3.71), which could lead to the tentative conclusion that in each series the other group (i.e., the audience to which it is not addressed) is linguistically marked.

References
25. Puente Castelo, Luís. “‘If Heaven has bestowed any Sense on them’: Rhetorical Conditionals in Late Modern English Scientific Writing”

Conditionals are particularly valuable in scientific writing (Carter-Thomas & Rowley-Jolivet 2008), as they can perform a wide array of functions: they can be used to present hypotheses and make strong assertions which take the form of conditionals (Warchal 2010: 146), in which the blatancy of the truth value leads the audience to conclude that the author is trying to convey a strong assertion, rather than trying to restrict the validity of the statement. Even though this type of conditional could seem uncharacteristic of scientific writing, previous studies have found instances of this use in scientific texts (Warchal 2010).

The aim of this paper is to conduct a microscopic analysis of rhetorical conditionals in eighteenth and nineteenth century scientific writing. To do so, it will use three of the subcorpora of the Corpus of English Scientific Writing, dealing with astronomy (CETA), philosophy (CEPHNT) and life sciences (CELSST), and amounting to 1,200,000 words. These subcorpora will be searched for selected conditional particles and the occurrences will be analysed in order to identify rhetorical conditionals. Then, the examples of rhetorical conditionals will be analysed in detail and classified according to nine parameters, and their preferred contexts of use will be analysed both in socio-historical and linguistic terms.

References


This paper is a follow-up research on recent Anglicisms borrowed into Italian in the new millennium, an area which is still relatively unexplored. A previous study based on the electronic editions of some general Italian dictionaries (Grande dizionario italiano dell’uso, 1999; Lo Zingarelli 2017; Devoito-Oli 2017) has shown that the semantic fields most affected by English loans from 2001 to 2016 are IT (eg adware, dashboard, file sharing), the Internet (eg blogger, cloud, doodle), economy (eg crowdfunding, double dip, patent box) and sport (kiteboard, skicross, Nordic walking) (Author 2017). For the present study two wordlists have been created, containing the new Anglicisms included in these dictionaries in two time spans, the first in 2010-2013 (2014 edition, 146 new items) and the second in 2014-2016 (2017 edition, 141 items), with the aim of identifying whether any differences can be observed in the semantic fields of new English ‘entries’ in Italian. The underlying assumptions is that these general dictionaries of Italian, which are yearly updated, tend to be fairly open to the inclusion of English loanwords, especially those that are closely linked to topical, social events and boosted by journalism.

As lexicographic evidence fails to provide information about the currency (assimilation) and representativeness (frequency) of headwords, the further analytical step taken in this paper is to search these Anglicisms in the Italian Web Corpus (ItTenTen16), a 4.9 billion word corpus made up of texts collected from the Internet. The itTenTen 16 is a part of the TenTen corpus family, which is a set of web corpora built with a target size of 10+ billion words. The corpus query aims to establish frequency bands for the Anglicisms considered, verify their usage domains (specialized or general) and check their status with respect to Italian equivalents, when available. This research may also show whether a web corpus is an adequate reference tool for extracting this type of information, although this is the only large general corpus available for Italian.

References
4. Queralt, Sheila and Roser Giménez García. “La lingüística de corpus al servicio de la justicia: Estudio de un caso real”

This study proposes an application direct of the linguistic applied to a context social especifico: a judicial investigation. It will present a study of comparison forensic of corpus textuales con fines de atribución de autoría enmarcada in the growing bibliography of these cases comparables in that experts studied the characteristic linguistic of a corpus dubitado (de autor desconocido) in contrast with one or various corpus indubitable (de autor conocido) to establish the distance or similitud idiolectal between them (p. e.j. Turell, 2011; Queralt, 2015). The two corpus compared in this task are a corpus dubitado composed by two textos breves encontrado during the investigation policical of a homicide sin resolver in España and a corpus indubitable extenso composed by communications personales de una de las personas of interés principales in the investigation. Los textos in the first present menos de cien palabras cada uno, mientras que los del segundo se encuentran entre veinticuatro and ciento setenta y nueve palabras. Dadas la brevedad and la lengua of the textos (catalán), así como the falta of software automatico and of métodos cuantitativos que puedan ser aplicados con garantías of éxito a textos tan cortos, se utilizaron métodos cualitativos and of estadística descriptiva for the analysis and a comparación métrica and rigurosa of the corpus. Además, and a pesar of las críticas that los métodos cualitativos han recibido in the bibliografía (p. e.j. Nini and Grant, 2013; Solan, 2013), existe the necesidad of desarrollar métodos cualitativos que puedan ser aplicados or adapted for tasks of authoría similares in contextos forenses reales. Las conclusiones of esta tarea apoyan la visión recientemente defendida by varios autores of the disciplina sobre las ventajas of métodos combinados que puedan proporcionar índices de fiabilidad, así como explicaciones lingüísticas of the distintividad of the marcadores idiolectales identificados by (v. Wright, 2014). Los objetivos of esta tarea of authoría eran varios. En primer lugar, we debían identify marcadores estilísticos that caracterizan the estilo idiolectal of each corpus. En segundo lugar, debíamos identify the marcadores that pudieran ser comparados entre the corpus teniendo en cuenta the aparición or ausencia in ambas muestras, así como the número of realizaciones. In tercer lugar, se realizó a compare a comparación among between dichos marcadores with the goal of evaluar the distancia or similitud idiolectal of the corpus. Finalmente, se llegó a a conclusion sobre the probability of that los textos dubitados hubieran sido producidos by the autora of the textos indubitable. Con la presentación of this estudio of caso, se pretende dar a conocer a a aplicación especifica of the lingüística of corpus that pone directly a the servicio of the sociedad las herramientas of this rama of the lingüística aplicada and que tiene entre its objetivos complementar investigaciones judiciales or policiales mediante the estudio of the lengua.

References

43. Rabadán, Rosa, Hugo Sanjurjo González and Ana García Gallego. “Multilayered annotation, contrast and CNL construction”

Technology has changed substantially the traditional categories of text production. The borders between writing and translating are becoming increasingly blurred as sub-tasks in both processes are taken up by algorithms. Rather than relying on fully automatic processes, corpus-
based contrastive data obtained from comparable corpora offer an interesting alternative solution if formulated as controlled natural languages (CNL).

This paper starts from the principle that multilayered annotation plays a crucial role in this contrastive process. Our aim is a) to show how comprehensive tagging can improve the understanding of language data, b) discuss an already tested protocol, and c) to present tools that may considerably speed up the annotation process.

Designing and annotating genre-restricted comparable corpora in English and in Spanish shows that different layers of linguistic information are a productive way to streamline cross-linguistic expertise knowledge. While rhetorical (move and step) annotation help reveal where generic differences lie across languages and cultures (Moreno & Swales 2017), POS annotation facilitates grammatical analysis as well as functioning as a discriminating factor on which all the other layers rely. Semantic annotation defines the domain and pragmatic annotation encodes functions not presented in other tagging layers, but vital in cross-linguistic analysis (Culpeper et al. 2008).

In practice, the problem is whether unsupervised annotation is possible or manual annotation, a painfully time-consuming task, is not to be avoided. Unsupervised annotation tools, commonly known as taggers, have had different levels of development and success. Whereas unsupervised POS taggers have proven successful with a token accuracy up to 98% for both English (Manning, 2011, pág. 171) and Spanish (Padró, Collado, Reese, Lloberes, & Castellón, 2010) this is not so in rhetorical and pragmatic annotation, where human intervention is required and language dependency is stronger.

The annotation of semantic features, however, is somewhere in between. To date, most of the efforts are focused on tagging partial semantic information aiming to support activities such as Named Entity Recognition, Semantic Role Labelling or Sentiment Analysis (Piao, et al., 2017, p. 115). To build CNLs, an all-words tagging approach is necessary because it provides a deeper and more comprehensive analysis of language data (Piao, et al., 2017, p. 113). This can be achieved using taggers based on ontologies, for example EuroWordNet (Vossen, 1998), or lexical knowledge bases, as the USAS semantic lexicon (Piao, Bianchi, Dayrell, D’Egidio, & Rayson, 2015). For our purpose, USAS is a better option as it uses a well-known taxonomy (McArthur, 1981) applicable to both English and Spanish. Current coverage of Spanish in specific domains needs to be improved and re-focused to be useful in CNL construction. Our solution has been to expand the semantic, domain-restricted, databases using research based on machine learning algorithms thus avoiding human intervention.

We will illustrate this protocol in one genre of the domain ‘food and drink industries’. The ultimate goal is to parameterize the results obtained in both languages to be used in the implementation of a domain and genre-restricted CNL. The procedure is replicable for other domains/genres and for other language pairs.

References


Corpus linguistics gives us concrete proofs that certain words and terms are used more often than others. This privileged look at actual linguistic usage is of great use in the assessment of the real-world effects of terminological planning efforts, because it can be demonstrated that certain terms are indeed used where others are abandoned. Corpus linguistics thus gives us the means to examine whether the efforts to support certain linguistic behaviours actually have the desired effect or not. Within the ambit of languages for specific purposes, the virtues of the corpus-based approach have been extolled most notably by Quirion (2003: 60-61, 64) as embodying the ideal manner to analyse the effectiveness of terminological planning efforts.

Following in these footsteps, this study is based on the examination of the use of terminology in the domain of space sciences and technologies in two French language corpora. The first corpus is composed of 360 doctoral theses (containing approximately 20 million tokens), written in French between 2010 and 2015, while the second corpus is composed of French language popular science magazines published between 2010-2015, and contains approximately 9 million tokens. The terminology studied is composed of terms in the domain of space sciences and technologies, the terms being recommended by the French Government’s Language Enrichment Scheme between 2008 and 2017. After describing the corpora and the context of terminological planning in France, this paper will firstly consider the terms that are actually used in the corpora and compare them with the terms whose use is recommended. This analysis underlies one of the conclusions of this paper, that usage does not follow the recommendations in the bulk of cases; rather, the recommendations seem to generally reflect general usage. Secondly, this paper will compare and contrast the use of different terms present in the two different corpora to consider the relationships between the terminological usage observed, the different nature of the texts in question, and the terms themselves.

We shall also demonstrate that several general tendencies can be observed in the different terms employed and avoided in the denotation of nearly 100 concepts in the different corpora, notably the selective deletion of prepositions, the preferential implantation of terms maintaining a high formal similarity with English language terms and the implantation of English acronyms and initialisms. However, we show that the influence of English is not sufficient to eradicate genuine French terms that vary, both formally and semantically, from their English equivalents. This paper thus takes a corpus-based approach to describe and explain real-world usage in a highly technical domain where English bears an acute influence on the lexicon of a
major world language that has benefitted from a long-term effort to quell this external influence.

References

71. Rodríguez Abreuñearas, Paula and Sergio Maruenda Bataller. “A diachronic characterization of the Spanish word CRIMEN (‘murder’) in a corpus of gender violence news reports” The second half of the 1990s was a key moment in the development of social awareness of violence against women (VAW) in Spain (cf. Aran Ramspott and Medina Bravo 2006: 10). The role of the media to this end is undeniable, as they are one of the main forces that threw the issue of VAW into the public spotlight. As a result of the growing attention received by the media, the discourse around VAW has undergone remarkable changes. In order to discover how the media have contributed to the social construction of VAW (Comas d’Argemir 2014) in roughly a decade (2004-2015), this presentation explores an ad-hoc corpus of gender violence news reports based on three Spanish dailies (namely, El País, El Mundo and ABC). Our main aim is to find out how newsworthiness is constructed around VAW. Taking Bednarek and Caple’s (2012, 2014) framework, we will particularly focus on how negativity and impact are created to discuss the death of a woman at the hands of her (ex-)partner. The predominance of these two news values responds to the very same nature of the media, which are enterprises trapped in between two opposing interests, that is, to represent the gravity of the events while satisfying their need to grab the spectators’ attention, usually by means of sensationalism (Zurbano Berenguer 2015: 186). The corpus material shows that up to the early 21st century this type of murder was typically defined as a crime of passion (crimen pasional). This controversial label seems to imply that the aggression is justified as an act of love, or even that women may somehow be regarded as “culpable when they are being beaten, even murdered, by their partners” (cf. Chesney-Lind and Chagnon 2017: n.p.). However, according to our data, the second decade of the 21st century is witnessing a change towards other less “romantic” labels, such as crimen machista or, to a lesser extent, crimen sexual or crimen doméstico (literally, male chauvinist, sexual and domestic crimes). In addition, a closer reading of concordance lines unveils the main devices used to create negativity and impact, such as the characterization of the crime by means of adjectives like horrible, bloody or violent. Nonetheless, the data unveil a latent discrepancy among the three newspapers. Thus, whereas ABC shows a clear tendency to intensify the description of the murders, El País keeps a more moderate tone in such accounts, even favoring the use of mitigating forms such as alleged or supposed.

References


63. Rodríguez Arrizabalaga, Beatriz. “Corpus research on the morphological variability of Anglicisms in contemporary Spanish” Como ya pronosticó Lorenzo (1996: 30) a finales del siglo XX, y más recientemente han constatado Rodríguez González (2003: 574) y Oncins Martínez (2009: 116), entre otros, los corpus informatizados se han convertido en una herramienta esencial para los estudios lexicográficos y, como consecuencia, para las investigaciones que analizan la incorporación y presencia de voces extranjeras en las distintas lenguas del mundo. De entre todas ellas, los anglicismos ocupan hoy en día un lugar eminentemente privilegiado debido, principalmente, al estatus de lengua franca del que la lengua inglesa, asociada siempre con valores positivos tales como "modernidad", "innovación", "internacionalización", "prestigio" y "utilidad", entre otros, (cf. Kelly-Holmes, 2005: 104; 182-183; Batibia, 2008: 166), goza en nuestro mundo cada vez más globalizado. De hecho, son muchos los análisis basados en corpus (cf. Furias et al, 2012) que demuestran la incesante entrada de voces inglesas, tanto necesarias como innecesarias, en distintas parcelas léxicas —gastronomía, ciencia, tecnología, economía, deportes, moda, entre otras— de numerosas lenguas, entre las que Hjarvar d (2004: 76), por ejemplo, sitúa lenguas europeas y afro-asiáticas de origen muy diverso: germánicas (alemán), romances (francés y español), eslavas (ruso) y semíticas (árabe). La utilización de los estudios sobre anglicismos basados en corpus va, sin embargo, mucho más allá de demostrar su productividad e índice de frecuencia en otras lenguas ya que estos pueden aportar datos muy reveladores, entre otras cuestiones, sobre su ortografía, sus propiedades morfológicas, los registros y variedades dialectales en los que se utilizan, etc. En este estudio, en concreto, ofrecemos los resultados más significativos de un análisis de corpus en el que investigamos el comportamiento morfológico de los denominados “anglicismos hispánicos” (cf. Lorenzo, 1996) nominales y adjetivas en español peninsular contemporáneo con objeto de demostrar que se trata de dos clases de palabras morfológicamente muy variables, en cuanto a las categorías de número y género se refiere, debido, a nuestro juicio, al comportamiento morfológico tan dispar que los sustantivos y adjetivos ingleses y españoles presentan en sus respectivas lenguas, y que precisamos, como tales, de una inminente normalización para la que todavía no parece haber consenso entre los académicos. Para ello, analizamos en dos corpus diferentes de español contemporáneo —el Corpus del español del siglo XXI (CORPES XXI) de la Real Academia Española y el Corpus del español de Davèrs (2016)— el comportamiento morfológico de nueve sustantivos —it-bag, blazer, celebrity, clutch, front row, sneakers, T-shirt, trench y sneakers— y seis adjetivos —cool, fashionable, nude, preppy, sporty y trendy— pertenecientes al léxico de la moda porque, como consecuencia del innegable poder de seducción comercial que caracteriza a la lengua inglesa, éste se ha convertido en una de las parcelas léxicas del español peninsular actual más receptivas a la entrada de anglicismos; dicho de otro modo, siguiendo a Lorenzo (1996: 11), Baumgardner (2006: 252) y García Morales et al. (2016: 19), entre otros, porque resulta inescusable que hoy en día “el inglés vende”.

References
Referencias
Seoane, Elena. 2006b. Changing styles: On the recent evolution of scientific British and American English. In Badger, Richard (ed.), Questions of style. Legal drafting manuals and scientific style aims at determining whether law reports present an overall frequency of passives similar to other written legal documents and most scientific discourse (see e.g. Seoane & Loureiro-Porto 2005; Seoane 2006a). However, the paper assesses whether significant changes can be observed in the frequency of passives over time or whether legal written texts are indeed resistant to change (Tiersma 1999: 135; see also Görlich 1999: 145; 2001: 205; Williams 2013: 353). Results suggest that the passive voice is extensively used in law reports, and that their frequency has increased from the sixteenth to the mid-twentieth century. However, during the second half of the twentieth century the frequency of passives is reduced, probably as a response to the Plain Language Movement which advocates for reducing passive verbal forms wherever possible (Williams 2007: 177; 2013; see also Seoane & Williams 2006: 124).

69. Rodríguez Puente, Paula. “Passives in legal English: A diachronic approach”

Passives are typically associated with formal academic prose, as markers of a detached, impersonal style which characterises writing (Biber 1988: 228; Biber et al. 1999: 476; Seoane 2006a, 2006b, 2013). As a formal written type of discourse, legal English makes extensive use of passive structures (Tiersma 1999: 75-77; Tiersma 1999: 74-77; Williams 2007: 35-36). Williams (2004: 231) suggests that approximately one quarter of all verbal constructions in prescriptive legal English take the passive form. According to Tiersma (1999: 75-76), passives (and also nominalizations) function as a device to obscure the actor, especially in prescriptive text types, such as statutes and court orders (cf. We shall punish those who skateboard on the sidewalks. Those who skateboard on sidewalks will be punished). For Bhatia (1994: 142), in turn, passives can help legal drafters to make provisions “more compact and yet precise and all-inclusive” (Bhatia 1994: 142).

This paper examines the development, usage and frequency of passives in A Corpus of Historical English Law Reports, 1535-1999 (Rodríguez-Puente et al. 2016), a half-million-word diachronic corpus of law reports: records of judicial decisions which interpret and fill the gaps where there is no statute law (Kearns 2007: 9; Rodríguez-Puente 2011; Fanego et al. 2017). Law reports constitute a specific type of legal document with a complex nature, which has led researchers to designate them as ‘hybrid’ texts, fulfilling both prescriptive (normative) and descriptive (non-normative) functions and containing both expository (more objective) and operative (more subjective) linguistic features (see Šarčević 2000: 11; Tiersma 1999: 139-141; Williams 2007: 28-29). In spite of these peculiarities, research on the linguistic features of law reports is scarce and limited to a small number of comments about their function, their usefulness as materials for teachers of English for Academic Legal Purposes (Badger 2003), and their hybrid nature. Although the release of the British Law Report Corpus 2008-2010 (Marín Pérez & Rea Rizzo 2012) has opened the door for research into the contemporary features of law reports, diachronic developments in this text type are largely unexplored. The current paper aims to fill part of this gap by investigating the use of passives here, focusing on two main research questions. First, it aims at determining whether law reports present an overall frequency of passives similar to other written legal documents and most scientific discourse (see e.g. Seoane & Loureiro-Porto 2005; Seoane 2006a). Second, the paper assesses whether significant changes can be observed in the frequency of passives over time or whether legal written texts are indeed resistant to change (Tiersma 1999: 135; see also Görlich 1999: 145; 2001: 205; Williams 2013: 353). Results suggest that the passive voice is extensively used in law reports, and that their frequency has increased from the sixteenth to the mid-twentieth century. However, during the second half of the twentieth century the frequency of passives is reduced, probably as a response to the Plain Language Movement which advocates for reducing passive verbal forms wherever possible (Williams 2007: 177; 2013; see also Seoane & Williams 2006: 124).


Desde que, con la lectura distante (Moretti), se propusiera un enfoque cuantitativo para complementar otras metodologías más clásicas del análisis literario, han sido pocos los análisis que con esta metodología se han centrado en escritoras latinoamericanas. El presente trabajo quiere contribuir a desentrañar, de las miradas de escritoras mexicanas contemporáneas, la perspectiva literaria de Cristina Rivera Garza, a quien algunos consideran una de sus máximas exponentes (Estrada; Palma Castro y Quintana).

Para indagar en las temáticas y sus tratamientos se recopilaron dos corpus propios: uno de la obra en prosa de CRG (novela, cuento y ensayo); y otro, como corpus de referencia, de Narrativa Contemporánea escrita por mujeres mexicanas. Este último es un corpus basado en un canon (Pacheco) que integra la obra de una veintena de escritoras del siglo XX y XXI. Tras procesar los textos mediante etiquetado automático con Freeling (Padró y Stanilovsky) se procedió en primer lugar a realizar una extracción y un análisis de palabras clave de la obra de CRG con AntConc (Anthony). Posteriormente, se procedió a agrupar manualmente algunas de estas palabras en áreas temáticas, lo cual finalmente permitió pasar a estudiar en profundidad colocaciones y líneas de concordancia.

Del análisis de las primeras 200 palabras clave, se identificaron cuatro áreas temáticas interconectadas:

1. En primer lugar, un área vinculada con el proceso de escritura. Una escritura que transgrede géneros y fronteras mediante el lenguaje. Aquí se identifican elementos que aparecen constantemente en los textos de Rivera Garza. Leer, investigar, analizar, manipular son acciones realizadas por personajes que captan, conjeturan, revelan e interpretan representaciones de diversas imágenes de la realidad: el fotógrafo, la periodista, el traductor y la detective. La mirada es clave en el arteficio literario.

2. En segundo lugar, se identifican proyecciones del cuerpo que tienen que ver con ese proceso de creación –el cuerpo textual se desdoba en el cuerpo de los personajes–. Por ello, estos elementos son relevantes en la obra de CRG. De ahí que resulte relevante la voz narrativa como la posibilidad de dar voz a los que carecen de ella: víctimas, desaparecidos, migrantes y mujeres.

3. En tercer lugar, se identifican partes de un cuerpo doliente/dolido cuya presencia jalona toda la obra analizada y apunta a la denuncia y la presencia de la violencia en la historia (reciente) de México. Finalmente se destacan, entre los verbos clave, el uso de varios verbos de cognición (recordar, pensar, suponer), que vienen a cerrar el círculo hermenéutico que se inicia con la escritura procesual.

Para terminar se realiza un examen textual con ProtAnt (Anthony y Baker) para determinar cuál de las obras en prosa de la autora podría resultar la más representativa de su producción literaria.

Esta investigación muestra que, mediante técnicas propias de la Lingüística de Corpus, se pueden llegar a conclusiones y lecturas casi idénticas a las que la crítica literaria llega por la vía clásica del análisis.

**References**

Anthony, Laurence. AntConc (v.3.4.3). Waseda University, 2014, http://www.laurenceanthony.net/


81. Rojas García, Juan, Pamela Faber and Riza Batista Navarro. “Application of Topic Modeling for the compilation of comparable specialized corpora thematically balanced”

In order to extract useful information from a bilingual comparable corpus (BCC) on a specialized domain of the science, let’s say Coastal Engineering, Bowker & Pearson (2002: 93 and 220) point out that it is essential that the two monolingual subcorpora contain original language material and have some features in common, which, at least, include subject matter (Coastal Engineering), text type (research papers and textbooks), degree of technicality (specialized and semi-specialized texts), period in which the texts were written (in the last twenty years), and similar size. However, it is difficult to manually ascertain, on the one hand, what individual topics are covered in the entire corpus (e. g. within the subject matter of Coastal Engineering, numerous related topics could be dealt with, among others, characteristics of coastal engineering structures such as groins, jetties, or seawalls; hydrodynamics of the coastal zone; field measurement techniques and analysis, etc.), and on the other hand, to what extent each topic is equally represented in both subcorpora.

We assumed as hypothesis that it would be most desirable that the BCC should be thematically balanced to obtain better results when searching for:

- a) an equivalent for a specialized term in the other language;
- b) how to use this term in an idiomatic way;
- c) different senses of a polysemous term, since the most typical terms on a specialized domain are not the most monosemous ones (Bertels, 2006 and 2011), and the topics in which a polysemous term appear can be useful for discriminating its senses (Reisinger & Mooney, 2010; Huang et al., 2012; Liu et al., 2015);
- d) what semantic relations a term held with other terms belonging all of them to a certain topic; and
- e) what lexico-syntactic patterns between two terms reveal a specific semantic relation.

Using an English specialized subcorpus on Coastal Engineering comprising 6 million tokens as starting point, the overall aim of this paper was to establish a procedure to compile a Spanish...
subcorpus, comparable to the English one, in such a way that both subcorpora were thematically balanced. More specifically, the English specialized subcorpus was built to represent a genre (specialized and semi-specialized texts), a register (science), a discipline (Environmental Engineering), a specific domain (Coastal Engineering), and a language (English). To maximize representativeness, the corpus was designed based on criteria proposed by Sinclair (1991, 2005), Meyer (2004), and Biber (2008): balance, diversity of sources, availability of texts in electronic form, period, size, use of complete texts, and variety of writers.

To achieve our goal, a Topic Modeling technique, based on Latent Dirichlet Allocation (Blei et al., 2003), was applied to the English subcorpus in order to automatically find out the different topics covered, along with the twenty more relevant terms associated with each topic. The topic-corresponding lists of relevant terms were used as keywords to query multidisciplinary research citation index databases, such as Web of Science or Scopus, to collect research papers on Coastal Engineering originally written in Spanish language.

References


22. Ruano García, Javier. “On the enregisterment of the Lancashire dialect in Late Modern English”

Over the past few years, there has been a growing scholarly interest in the processes whereby varieties of English are constructed and publicly circulated as (relatively) stable sets of linguistic features that index sociocultural values. This has been largely prompted by Asif Agha’s groundbreaking framework of *enregisterment*, which, as is known, he defines as “the processes through which a linguistic repertoire becomes differentiable within a language as a socially recognized register of forms” (2003: 231-2). Whilst modern varieties of English have received a considerable amount of scholarly attention in light of Agha’s framework—especially visible in Johnstone’s extensive study of Pittsburgese (see Johnstone 2006, 2011, 2013; Johnstone et al. 2006)—, little research has been conducted in order to explore the enregisterment of older varieties of English: Picone (2013) and Paulsen (2016), for example, have examined these processes in the US, whereas Beal (2009, 2012a, 2012b, 2016), Cooper (2013), and Beal and Cooper (2015) have addressed the enregisterment of northern varieties of British English during the Early / Late Modern periods. They have considered different types of discourse, namely dictionaries, newspapers, and especially dialect writing, which, because it relies consciously on regional speech forms, Clark (2013: 261) highlights, “may not only be an intentional act, but an act of enregisterment”.

This paper takes a preliminary approach to the enregisterment of the Lancashire dialect in the Late Modern English period. Drawing on The Salamanca Corpus, I look at Lancashire dialect writings published between 1700-1900 and undertake a quantitative analysis of the data to determine, firstly, the repertoire of forms that were circulated in representations of the dialect and the values they indexed; that is, to explore how the Lancashire dialect was enregistered in Late Modern England, for which purpose Silverstein’s framework of indexicality is likewise taken into consideration (see Silverstein 2003). Secondly, the paper attempts to ascertain whether some of the features associated with the dialect lost saliency over this period (i.e., were deregistered—see Cooper 2017), pointing at changing attitudes towards and perceptions about the dialect. In doing so, this paper may contribute to recent research on enregisterment in historical contexts, as well as on the Lancashire dialect in particular, which has not been thus far explored through the lens of Agha’s pioneering framework.

References


Clark, Urszula 2013: “‘er’s from off: the indexicalization and enregisterment of Black Country dialect”. *American Speech* 88.4: 441-466.


59. Rufat, Anna. “Dar como verbo de apoyo y como verbo pleno en el discurso escrito nativo y no nativo”

Nuestra investigación muestra la manera en la que un enfoque lingüístico basado en corpus informatizado aplicado a los estudios de interlengua (Contrastive Interlanguage Analysis, Granger [1996, 2015]), en este caso del español, contribuye a revelar las semejanzas y diferencias entre el español de un grupo de hablantes no nativos y un grupo equivalente de hablantes nativos. En concreto, se analiza de la lengua escrita de un grupo de aprendientes anglofonos universitarios el uso de dar como verbo de apoyo (dar un salto) y como verbo pleno (dar un libro/dar alojamiento) en la construcción sintáctica más distintiva de este verbo, a saber verbo + nombre.


Los verbos frecuentes han desarrollado en cada lengua muchas especificidades (Altenberg y Granger, 2001, p. 174); así, pues, por las características que frecuentemente se atribuyen a los verbos frecuentes (Vivero, 1996) y debido a la falta de acuerdo entre los investigadores de la interlengua del inglés acerca de si estos verbos se sobreutilizan (Källvist, 1999; Guo, 2006) o infrautilizan (Altenberg y Granger, 2001, o si son seguros para el no nativo (Hasselgren, 1994: 250, en relación con su propuesta teddy bear principle,) o una fuerte de errores (Källvist, 1999; Altenberg y Granger, 2001; Guo, 2006), unido, a su vez, a la escasez de trabajos que se centran en este aspecto de la interlengua del español, la necesidad de este análisis se ve reforzada.

El corpus utilizado para este análisis es el corpus escrito del español como lengua segunda o CEDEL2 (Lozano, 2009; Lozano y Mendikoetxea, 2013), compilado por la Universidad Autónoma de Madrid y la Universidad de Granada. Dado que el español en el corpus de los hablantes nativos presenta ciertas marcas de distinto orden (diástricas, diatópicas y diafásicas) y la interlengua debe analizarse sobre la base de algún tipo de norma prescriptiva que permita determinar la adecuación de las muestras lingüísticas a la variedad del español seleccionada como modelo, se ha aplicado un estándar panhispanico, resultado de la unión de los estándares peninsular y americano, que constituye un referente lingüístico operativo para categorizar los datos del CEDEL2, en el que los hablantes nativos y los hablantes no nativos manejan variedades diatópicas distintas. Las diferencias y similitudes entre el español de los nativos y el español de los no nativos permite identificar las necesidades de estos referidas al verbo dar y determinar si estamos ante un verbo seguro (su uso no muestra desviaciones perceptibles) para el no nativo, o, por el contrario, problemático. Los resultados muestran que los aprendientes, incluso en un nivel avanzado, tienen dificultades con un verbo tan frecuente como dar.

**Referencias**


De Miguel, E. (2011). “En qué consiste ser verbo de apoyo”. En M.V. Escandell Vidal et al. (Eds.), *60 problemas de gramática* (pp. 139-146). Madrid: Akal.


70. Sánchez Barrera, Ana Ruth y Ignacio Rodríguez Sánchez. “Miradas furiosas desde el baño. Un análisis de la novela rosa mexicana contemporánea”

Contrariamente a la telenovela, la novela rosa mexicana (NRM) es un género que ha despertado escaso interés en el mundo académico hispanohablante: existen pocos estudios académicos, limitados a muy pocas obras, y casi siempre tomándola como subproducto cultural. Desde los estudios culturales y literarios se ha identificado, como rasgo distintivo del género, una línea discursiva donde la heroína desconfía y se defiende de un héroe que se burla de ella (Cane, 1994; Modleski, 2008; Valero, 2012). Desde la Lingüística de Corpus se ha estudiado a la codificación lingüística del sexo, el romance y las relaciones de género en novelas de géneros populares angloamericanos (Baker, 2010; Harvey & Shalom, 1997).

En la presente investigación se analizan las representaciones del rol de las mujeres en la NRM desde la conjugación de tres disciplinas: Lingüística de Corpus, Estudios Literarios y Estudios Culturales. Siguiendo la propuesta canónica de Stubbs (1995), se propone un análisis desde dos niveles: el de los elementos de la lengua utilizados en la codificación y lo que estos revelan de la cultura de los hablantes.

Metodológicamente, proponemos un trabajo en que se cruzan datos de cuatro corpus distintos: un corpus experimental de Novela Rosa Mexicana Contemporánea (CNRMC, de elaboración propia); y tres corpus de referencia: Corpus de Referencia del Español Actual (RAE, 2008), un Corpus de Novela Rosa Clásica (de creación propia); y un Corpus de Narrativa Contemporánea escrita por Mujeres Mexicanas (de elaboración propia). Cada uno de los corpus propios contiene aproximadamente un millón y medio de palabras y su composición siguió criterios de muestreo escrupulosos y la asesoría de expertos en cada área.

El análisis de palabras clave y palabras clave-clave arroja atención sobre datos centrados en varios ejes, de los que destacamos los siguientes:

a) El cuerpo. En las relaciones amorosas descritas en la NRM, el cuerpo se describe en situaciones graduadas de poder y control. Se presenta un análisis de las palabras clave como cuerpo, mirada, ojos, cabello y labios. Por ejemplo, las siguientes colocaciones de mirada en el CNRMC (considerando frecuencia, dispersión y probabilidad logarítmica) parecen subvertir algunos roles...
clásicos del héroe y heroína: los adjetivos reprobatoria, inquisitiva, matadora, asesina, perdida, intensa, furiosa y los verbos clavar, sostener, esquivar, fulminar, dedicar.

b) Percepción del otro y autopercepción. Por ejemplo, la colocación sentir ganas se utiliza frecuentemente en contextos donde un personaje masculino expresa a su interlocutora un deseo especialmente violento.

c) Espacio. La intimidad es un ámbito que las heroínas buscan en dos sentidos: para tener privacidad y para el erotismo. En el CNRMC, el baño (WC) aparece como un nuevo espacio donde aparecen algunos personajes femeninos.

Como conclusión, se propone una visión contemporánea de este género en el que se proyecta una representación alternativa de algunos personajes femeninos (en mayor control de su sensualidad y su vida), pero que no excluye sino que se conjuga y convive con rasgos prototípicos correspondientes a una visión tradicional conservadora.

Referencias
Real Academia Española. (2008). Banco de datos (CREA) Corpus de referencia del Español Actual v.3.2.

67. Sánchez Berriel, Isabel and José Luis González Ávila. “Técnicas de aprendizaje profundo aplicadas al análisis visual de colocaciones léxicas en español”

En los últimos años han irrumpido las técnicas de aprendizaje profundo en el campo del Procesamiento del Lenguaje Natural (PLN) de forma exitosa. Sin embargo, es difícil encontrar investigaciones lingüísticas que se apoyen en el uso de las ventajas que aporta la representación de palabras en un espacio vectorial continuo. Este problema es aún más inexplorado cuando se habla de su aplicación al español.

La investigación que se presenta se centra en aplicar estas técnicas punteras en el análisis semántico de las colocaciones en español. Se estudian las técnicas de aprendizaje profundo y su aplicación al español en combinación con la librería de PLN, FreeLing. Se usa la representación obtenida para los coloquios en colocaciones léxicas del español para mostrarla de forma amigable mediante técnicas de visualización que faciliten el análisis semántico de las mismas.

Las bases con las que combina un determinado verbo, adverbio o sustantivo en las colocaciones verbo + sustantivo, adverbio + verbo o sustantivo + adyacente pertenecen a grupos semánticos. Algunos ejemplos son: sentir +(náusea, escalofrío, hambre), sentir+(compasión, lástima), abundantemente+ (proveer, producir, surtir), abundantemente+(llorar, sudar). En (Bosque, 2001) se resalta la interpretación de las colocaciones como un caso particular del concepto más amplio de selección léxica. Una determinada base de forma aislada no constituye el único caso compatible para un determinado coloquio. Con contraposición, éste se selecciona por sus rasgos semánticos, que son compartidos por otros sustantivos en el caso verbo + sustantivo, otros verbos para el tipo adverbio + verbo, o bien sustantivos cuando se trata del caso sustantivo + adjetivo.

El método word2vec (Mikolov et al., 2013) mapea las palabras del corpus en un espacio vectorial continuo utilizando una red neuronal que aprende representaciones que predicen los contextos de una palabra dada. Estos modelos del lenguaje captan propiedades sintácticas y semánticas que además pueden ser expresadas mediante analogías como: mujer : hombre :: reina : rey. Esto se refleja en las transformaciones lineales dentro del espacio vectorial: v(mujer)-v(hombre) = v(reina) - v(rey). Una de las grandes virtudes de este modelo es que palabras con comportamiento lingüístico similar quedan representadas en vectores cercanos en un espacio dimensional continuo. Esta última propiedad permite realizar análisis de las colocaciones apoyados en la visualización de los datos.

Por una parte, determinamos grupos de palabras entre los argumentos aplicando análisis cluster mediante el algoritmo de las k-medias, por otra, se generan visualizaciones usando t-Distributed Stochastic Neighbor Embedding (t-SNE). Se dispone de una plataforma que recupera textos de la DBpedia, Twitter, RSS de la prensa y el corpus de actas de plenos municipales publicado en la sección de Transparencia del sitio web del Ayuntamiento de Santa Cruz de Tenerife. Se ha seleccionado la librería java DL4J (DeepLearning4j) que incorpora word2vec. Freeling 4.0 se utiliza previamente con objeto de normalizar los textos manejados, el nivel de análisis aplicado es la lematización. Se generan los gráficos sobre las colocaciones en español que se presentan como ejemplos en los trabajos lingüísticos consultados, entre las que aparecen 1673 palabras diferentes.

References

Intimate Partner Violence (IPV) is undoubtedly one of the most pervasive concerns for many contemporary societies. Although efforts have been gradually introduced to eradicate this power imbalance, recent figures suggest that this situation is far from improving (WHO, 2016). As a matter of fact, 25% of women today are believed to go through abusive experiences of this sort during their lifespans. Even more alarmingly, there seems to be an increasing tendency among youngsters to assume that this sort of violence is a normal attitude within an intimate partner context (FAO, 2017).

Studies aiming at getting a better understanding of IPV and its dynamics are mostly carried out from disciplines that are thought to have a more direct connection with this social issue, namely psychology or the health sciences. Bearing in mind that one of its main motivations is the analysis of linguistic aspects of social problems and processes (Wodak and Meyer, 2009), it is striking that research from a Critical Discourse Studies (CDS) perspective on this issue is still relatively scarce. Although most attempts have focused on recontextualised discourses around/on IPV (Santaemilia & Maruenda, 2014), discourses by key actors, namely abused women and perpetrators, are also gradually blooming (Holmes et al., 2007).

In the attempt to furnish this niche, and endorsing many of the tenets that extol the useful virtues that spring from combining critical discourse analysis and corpus linguistics (Baker et al., 2010), this research sets out to shed light on the discursive choices of women survivors of IPV focusing on a corpus-assisted examination at the textual level. To this end, I am assisted by a 120,000-word corpus collected from an anonymised, public, online forum available to IPV survivors in which they share their experiences with this type of violence. More specifically, three online communities embedded within this forum (‘Is it Abuse?’, ‘Getting out’ and ‘Life after abuse’) are contrasted in order to sketch out how the discursive output can vary across these three communities, which at the same time mirror the usual stages women experiencing this type of abuse tend to undergo.

This research builds upon recent studies in this vein (Sánchez-Moya, 2017), in which the text-analysis software tool LIWC (Pennebaker et al., 2011) is applied to, inter alia, identifying the differences in the emotional tone across the three above-mentioned communities. For the purposes of this paper, also following a sentiment analysis perspective, I make use of Lingmotif (Moreno-Ortíz, 2016) in the attempt to determine the semantic orientation of the online posts pertaining to the three communities under scrutiny by relying on Lingmotif’s ability to detect linguistic expressions of polarity. Interestingly, this study seeks to provide insights into the ways these three communities are characterized in their (collective) semantic orientation. By doing so, this study arguably contributes to the discussion on to what extent the expression of sentiment can alter depending on different domains and their specificities (Choi et al., 2009).

References

103. Santaemilia, José. “‘El hombre presuntamente asesó más de veinte puñaladas a su pareja’: What is the role of men in media representation of violence against women?”

The way we think, describe and evaluate sensitive issues –such as violence against women (VAW), homosexuality or abortion, to name just a few- is one of the most powerful discursive constructions of contemporary society. Media language crucially contributes to the changing social definitions of social issues, providing a language that is highly evaluative and ideological (Hunston 2010). Contemporary mass media are essential in defining key terms for day-to-day communication purposes. According to Bergvall (1999: 284), the mass media “impose images and constructions of behaviour”, thus determining one’s ethical and ideological attitude towards sensitive social issues.

In VAW news stories, women are omnipresent, as they are both the victims and the targets of social empathy. They constitute the focus around which media discourse on VAW is articulated as an unstable mixture of police report, social denunciation and institutional support. And what happens with men? In this paper I explore the term ‘hombre’ (‘man’ in Spanish) in a large corpus of VAW news from the digital versions of quality Spanish dailies such as El País, El Mundo and ABC, published over a large period of time (2004-2015), and constituting a sub-corpus of GENTEXT-N, a large ad-hoc corpus compiled at the Universitat de València and revolving around contemporary gender and sexual (in)equality issues. Our methodological approach is a mixed one, bringing together corpus techniques (word frequencies, collocations, and concordances – see Baker 2010) and critical analytical tools (CDA, news values –see Hoey 2005, Baker et al 2008, Baker & Levon 2015, Bednarek 2017). Our aim is to unveil meanings and associations that sometimes differ significantly from the values promoted either by Spanish traditional lexicography or by common language.

Whereas ‘woman’ is clearly foregrounded in contemporary Spanish media discourse on VAW (see Santaemilia & Maruenda 2016), ‘man’ is somewhat underrepresented and is often omitted. Traditional lexicography (e.g. DRAE 2014) still presents men and women in a more or less stereotypical light (Calero 1999, Bengoechea 2000, Forgas 2000, Lledó 2002), thus perpetuating asymmetries (Cameron 1998), prejudices (Hoey 1996) or sexist traces (Stoll 2003, Cabeza & Rodríguez 2013) that have repeatedly been denounced or condemned.

Attention, then, will be paid to a number of traits (word frequencies, strongest collocates, concordance data, discourse prosodies, lexical priming, agency, critical or ideological values, positivity/negativity and other news values, etc.) that are likely to reveal certain ideological values in the presence/absence of men in VAW news stories or in the roles they play (or fail to play). It becomes apparent that blame is increasingly assigned to men in gender-based violence episodes, but women routinely still continue to be –linguistically, rhetorically and ideologically-
at its very centre. By revising contemporary social or ideological definitions of key terms (woman, man, violence, victim, aggression, and many others) in sensitive issues we are likely to gain a greater understanding of the changing public discourses on socioideological issues (such as VAW) in the Western world.

References


33. Sentí, Andreu and Sandra Montserrat. “Elaboración de un corpus oral y coloquial del catalán valenciano”


Carrera-Sabaté, Josefina y Joaquim Viaplana (ed.): *Corpus Oral Dialectal (COD)*. Textos orales del nordesticentral. Dipòsit Digital de la UB.

CIVAL = Academia Valenciana de la Lengua: *Corpus Informatitzat del Valencià [http://cival.avl.gva.es/]*


COD = *Corpus oral dialectal*. [http://www.ub.edu/ccub/corpusoraldialectal-cod.html]

CTILC = Institut d’Estudis Catalans: *Corpus textual informatitzat de la llengua catalana* [https://ctilc.iec.cat/]


Perea, María-Pilar y Joaquim Viaplana: *Corpus Oral Dialectal (COD)*. Selecció de textos. Dipòsit Digital de la UB.


54. Serrano Losada, Mario. “The rise and development of mirative ‘end up’ in American English”

Mirativity is a linguistic category that conveys new or unexpected information with overtones of surprise (DeLancey 1997; 2012). Recently it has received increasing scholarly attention, especially from the perspective of those languages that do not encode it morphologically, such as English (e.g. Gentens et al. 2016). The Late Modern English period bears witness to the appearance of several mirative verb constructions. This group of constructions already included verbs like prove, attested from Middle English, and appear, from Early Modern English (Visser 1963: §235). During the Late Modern English period it incorporated other raising verbs such as turn out, in the eighteenth century, and, more recently, end up, first attested in late nineteenth century American English.

This paper zooms in on *end up* mirative constructions to explore the processes by which this phrasal verb emerges and shifts from its original lexical resulative meaning (1) to its conventionalized mirative meaning (2):

1. (1) Moor, Mrs. Myers, Marvin, and William Leland all signed the articles of capitulation, and the affair *ended up* with a grand ball at the Clarendon. (COHA, 1872)
2. (2) I couldn’t bring myself to vote for Trump, but I *ended up* being happy that he was elected and I’ve spent a lot of time since the election trying to understand his appeal. (NOW, 2017)

The process whereby the present-day verb has come to signal information which is new or unexpected, encoding overtones of surprise, is not an unfamiliar pathway, as the extant literature has observed a tendency for resulative and inchoative senses to develop mirative nuances (González-Fernández & Maldonado 1998).

In present-day American English, *end up* seems to be undergoing further development, as it can now be used parenthetically (3). A similar change has been observed for other evidential and mirative verbs (e.g. *it seems, turns out*; Heller & Howe 2008).

3. *Ends up* the local police chief set the fire so he could rescue his girlfriend, who lived in an upstairs apartment. (COCA, 2008)

This paper argues that the rise of the various mirative *end up* constructions is the result of both pragmatic enrichment and paradigmatic analogy, motivated by a series of semantically and formally related expressions. Thus, the incursion of the verb into the paradigm of evidential and mirative raising verbs was brought about by both concrete and structural analogical modeling (cf. Fischer 2015), following preexisting members of the category such as *happen, prove* and, most prominently, *turn out*. These verbs do not only share their status as raising predicates that express evidential and/or mirative meanings, but also show similar constructional behavior (e.g. most can be used in as-parenthetical constructions). Data for the present study were drawn from COHA, COCA, NOW and SOAP.

Sources


**NOW** *Corpus of News on the Web: 3+ billion words from 20 countries, updated every day*, 2010-2016. Compiled by Mark Davies. http://corpus.byu.edu/now/


References


83. Sharygina, Sofia. “A Comparative Corpus-based Analysis of the Means of Expressing the Future and the Prospective Aspect in English and Spanish”

The present paper investigates means of expressing future time reference and prospective aspect in Spanish and English.

From the point of view of both typology and grammar, future tense and prospective aspect are differentiated. While forms and constructions with prospective meaning usually express the meaning that can be defined as ‘the state of X is such that later P will occur’, i.e. point to a precondition for some subsequent situation [Comrie 1976: 64], the meaning of expressions of future time reference is identified as ‘from my point of view the situation P will proceed after the moment of speaking’, i.e. the use of future forms means that an assertion about the future is being made [Bybee et al. 1991: 19].

Nevertheless, the expressions of prospective aspect and future tense appear to be quite close semantically, which leads to their interoperability in spoken and written language as well as to their being sometimes mixed up in academic literature.

The research described in the paper has been done on the basis of data provided by the parallel English-Spanish corpus which is a part of The Czech National Corpus. To improve the quality of the analysis the texts provided by the corpus were divided into two groups: the texts of written formal language (group A) and the texts of spoken informal language (group B).

In the research 200 examples of future forms being used in English and Spanish texts and translated to the second language were analyzed, as well as 200 examples of each of these English prospective constructions being used and translated to Spanish:

To be going to + Inf;
To be about + Inf.

And these Spanish prospective constructions being used and translated to English:

Ir a + Inf;
Estar por + Inf;
Estar a punt de + Inf.

The results indicate that the following general conclusions can be drawn:

- the choice of the translation of each form or construction is greatly affected by the type of the text: whether it is a text of written formal language or a text of spoken informal language;
- there is a tendency to use a form or construction that is ‘parallel’ to the original one;
- nevertheless, one cannot claim that they have the same meaning.

References


104. Suleymanov, Dzhavdet and Olga Nevzorova. “Vowel Harmony Violations in Tatar: Studying Corpus Data”

The paper presents the results of exploring vowel harmony violations in Tatar using the data of Tatar National Corpus (http://tugantel.tatar/?lang=en). This Corpus contains texts of different styles and genres; its current volume is 150,000,000 tokens. It has a system of grammatical annotation that is oriented at presenting all the existing grammatical word forms.

The Tatar language is related to the Turkic family and is characterized by a rich agglutinative morphology; presently it uses the Cyrillic alphabet.

The vowel harmony is the main feature of phonological organization of Turkic languages. The vowel harmony in Tatar is a morphonological assimilatory process involving agreement between vowels within a word form. Tatar words usually contain merely back vowels or merely front vowels; stems do not alternate and alternating affixes (derivational and inflectional ones) follow vowel harmony rules. Nevertheless a large number of exceptions of different origin and structure exist, and their study is of theoretical and practical significance. Understanding vowel harmony is a crucial step towards implementation and improvement of text processing systems for Tatar.

Written texts represented in the Corpus collection cannot reflect all phonetic peculiarities and dimensions of vowel harmony and its violations within linguistic items (for instance Tatar has both vowel backness harmony and vowel roundness harmony, but the latter is not reflected in writing). Nevertheless corpus data research may shed light upon such cases.

Thus we identified word forms with mixed sets of vowels and carried out their quantitative analysis. We distinguished their basic grammatical classes. We also compiled a frequency list of words of different origin and structure with mixed sets of vowels.

Corpus evidence allows us to conclude that commonly used loanwords of Arabic and Persian origin make the main contribution to violating the law of vowel harmony in Tatar.

The second most important class is formed by Russian loanwords and borrowings from European languages through Russian. The vowels e and i found in Russian borrowings look like corresponding Tatar vowels, however they require back vowel affixes, so Russian front vowels behave in a non-trivial way blending into the Tatar phonological system. Another important lexical class is made up of compound words consisting of two stems containing back and front vowels; the most frequent of them are compound pronominal and numeral.

We studied also the influence of grammatical factors on vowel disharmony and described the most important ones.

The paper presents statistical distribution of classes of words with vowel harmony violations and typical models of vowel harmony violations in Tatar. The results may be used to clarify a number of provisions of Tatar grammar concerning interaction of phonetics and morphology and to improve the Tatar morphological parser for which the words with mixed sets of vowels are the sources of errors. The obtained corpus data can also be applied for development of systems of synthesis and recognition of Tatar speech.
38. Terrazas Calero, Ana Mª. “‘I mean, he’s a bogger – there’s the whole language...’: The use of Irish English in the Ross O’Carroll-Kelly plays”

Literary dialects have traditionally been criticized by some academics who challenged their role as representatives of real spokenness due to their having been created by writers and for being ‘watered down’ representations of orality. However, they are created according to models of naturally-occurring language and, as such, they can reveal important sociolinguistic information about their speakers while enriching the characterization of the book. In addition, they have proven to be reliable sources for linguistic evidence as they pinpoint the presence or absence of linguistic features. Thus, they function as great tools for documenting linguistic change and for the enregisterment of speaker identities.

The concept of enregisterment was developed by anthropologist Asif Agha and refers to the way a set of linguistic features, which were previously unnoticed by a community of speakers, undergoes a series of processes whereby it becomes “indexical of speaker attributes” (Agha 2005: 38), thus conveying speaker identity. Johnstone et al. (2006: 83), for example, discuss how the monophthongization of /a/ to [a] in Pittsburghese, whereby about sounds like [a.b.t], enregisters the voice of working class, male speakers from the Pittsburgh area. Similarly, in fictionalized Irish English (FtE), the continuous repetition of the word roysh (‘right’) in Paul Howard’s Ross O’Carroll-Kelly (RO’CK) series enregisters the linguistic stereotype of an upper-class speaker of Dortspeak or Southside Dublin English.

The present paper takes Howard’s RO’CK series as being representative of real Dublin English (DubE) orality. These books, which began in the late nineties as a sports column in the Sunday Tribune newspaper and have since become a series of novels, are a parody of the elite and snobbish society that emerged during the Celtic Tiger in Southside Dublin. The novels, which have become incredibly popular in Ireland, have also gained critical acclaim for their natural-sounding rendition of DubE orality. However, and despite their linguistic richness, few studies have investigated the language used in the novels (see, for example, Lillo 2007, Gorman 2013, or Amador-Moreno 2016), and none have looked at the way FtE is utilized in the three unpublished RO’CK plays that have been produced to date to enregister contemporary Irish identity.

This paper, therefore, will carry out a quantitative and qualitative study of the way FtE is used in a small corpus of ca 50,000 words, comprising the plays The Last Days of the Celtic Tiger (2007), Between Foxxrock and a Hard Place (2010), and Breaking Dad (2014). Given the fact that, unlike the novels, these plays are created to be performed (and not read) in front of a live audience, the paper will examine in depth the form and use of the three most frequently occurring pragmatic features of FtE as used in the plays (namely 1) taboo word f*ucking, 2) discourse marker like, and the use of 3) high rising terminal). Finally, the paper will also determine what the use of these three linguistic features enregisters in terms of the representation of contemporary Irish (Dublin) identity through fictionalized Irish English.

References


97. Uriel Sanz, Belén. “A Linguistic Analysis of Spelling in Literary Representations of Irish English of the 17th and 18th Centuries: A Corpus-Based Study”

This proposal analyses the influence of English on literary representations of Irish English in the 17th and 18th centuries. It undertakes a corpus-based study of Irish English as represented in drama during these two centuries, focusing on a thorough linguistic analysis of spelling in four plays: Ben Jonson’s The Irish Mask at Court (1613), Thomas Shadwell’s The Lancashire Witches and Tegue O’Divelly, The Irish Priest (1682), Charles Shadwell’s The Humours of the Army (1713) and Charles Macklin’s The True-born Irishman (1767). The Early Modern English period (1500-1700) saw many changes in the language as a result of the process of standardization. As is well known, Standard English became a model to be followed, neglecting the rest of the regional dialects. At first, authors avoided the use of dialects in their literary works; however, playwrights started to introduce them for stylistic purposes, highlighting features associated with well-defined manners of speech. During the course of the 18th century, playwrights changed their attitudes towards non-standard varieties, using them with other literary connotations: such is the case of Irish English. Although the dialect was constructed drawing upon the same kind of techniques than in the previous century, playwrights, who were often outsiders with regard to Irish English, only selected a few salient traits. This goes some way to suggesting that by the end of the period examined Irish speakers, as represented on the stage, would be comparatively more fluent in English, and that the standard speaking audience would be aware of such markedly salient Irish features. In light of this, this proposal attempts to measure the reliability and consistency of the use of Irish spelling traits in these four literary writings during the 17th and 18th centuries, trying to identify diachronic variation in the spelling representation of the dialect which may point to such changing attitudes towards Irish English in the course of these centuries. Among the most remarkable features that distinguish Irish English, the corpus testifies to an important range of patterns representative of vowels, diphthongs and consonants. To conclude, interestingly, the analysis has likewise revealed that the representation of Irish English shifted from the 17th century to the 18th century, silencing Irish speech in favor of the use Standard English. The results obtained reflect the influence of English (Standard English in this case) on Irish English since the 18th century in the literary representation of this language as British playwrights only selected isolated features of Irish English that were quickly identified by a Standard English-speaking audience.

Primary References
the acquisition process: intermediate and advance levels. So, the total number of essays per level are: 436 and 321, respectively. As for the number of students, there are 13 students in level B1 and 95 in level C1.

All the essays were read and 2474 pronominal constructions tagged as being either right or wrong productions. 777 constructions were extracted from B1 essays and 1697 for level C1. If right, we further annotated the semantics of the construction (reflexive, passive, etc.) and if wrong, two possibilities are considered:

a) Students produced a pronominal sentence but it should not be pronominal at all. We annotate the type of error.
b) Students produced a non-pronominal sentence and it should be pronominal. We further annotate the kind of semantics it represents.

We would like to highlight the main contributions of this study from the methodological point of view: first, the number of participants and material analyzed; second, all pronominal meanings have been included, and, finally, most of the studies performed on this topic have been based on grammaticality judgments and not on corpora. Also, our work provides data both on omission and overgeneralization of “se”, which is not common in this area of research (Zyzik, 2006, on grammaticality judgments and not on corpora. Also, our work provides data both on omission and overgeneralization of “se”, which is not common in this area of research (Zyzik, 2006, Escrutia 2010).

As for the results, we could very briefly say that surprisingly, students with an L1 not belonging to a Romance language outperformed students with a Romance L1. Nevertheless, the difference was not statistically significant. Regarding the semantics of pronominal constructions, as expected, some constructions have proven to be more difficult than other constructions. Finally, we believe better materials for SFL can be tailored to deal with the problems we have identified in the study.

References

114. Vázquez Balonga, Delfina and Pedro Sánchez-Prieto Borja. “Variación diacrónica y diatópica en la Comunidad de Madrid a partir de un corpus de documentos archivísticos: ALDICAM-CM”

El habla de Madrid ha sido objeto de recientes trabajos que se han orientado al examen de la variación sociolingüística en la capital (Paredes García, Molina Martos, entre otros). Otros estudios trabajos sincrónicos han pretendido o pretenden recuperar usos del mundo rural de Madrid, ya en declive. Entre estos destaca en los últimos años el Atlas Lingüístico de la Comunidad de Madrid (García Mouton y Molina Martos 2015 y 2007), basado en encuestas de la periferia de la región. En cambio, los estudios dedicados a la variación histórica de la lengua madrileña son relativamente escasos, y han tenido por objeto, sobre todo, la lengua literaria de autores como Galdós (Russell 1982) o Arimches (Secco 1970). Estos textos no son, sin embargo, los que mejor se prestan a la investigación realista sobre los usos lingüísticos madrileños, tanto de la capital como de otras localidades, en todos sus niveles.

Por este motivo, creemos necesario recuperar la variación interna del habla de Madrid en el pasado a partir de las fuentes documentales que se custodian en sus archivos. Por ello hemos promovido un amplio proyecto de investigación financiado por la Comunidad de Madrid y el Fondo Social Europeo, y cuyo objetivo fundamental es la elaboración de un corpus textual electrónico de documentos archivísticos emitidos y conservados en localidades de la región. Todo este material se incorporará al Atlas Lingüístico Diacrónico y Dinámico de la Comunidad de Madrid (ALDICAM-CM), disponible próximamente en la web y consultable libremente para los usuarios. En él, los investigadores podrán realizar consultas filtradas respecto de diversos parámetros: cronología, geografía (localidades de emisión), archivos, tipología documental, temas y ámbito de elaboración. Los resultados se visualizarán de forma inmediata en mapa, en el que cada rasgo buscado se adhiere al punto de emisión del documento en que aparece. Otro objetivo es la recuperación del patrimonio documental de la Comunidad de Madrid, a menudo olvidado. Pretendemos así sacar a la luz un material de extraordinario valor para la historia lingüística regional, no solo por la diversidad geográfica interna, sino también la representatividad que todos estos documentos tienen respecto a todo el espectro sociocultural de Madrid y su tierra a lo largo del tiempo.

El marco teórico en el que se realiza este proyecto es el de la lingüística variacionista que se conecta a su vez con los desarrollos más notables de la sociolingüística histórica (Nevalainen y Raumolin-Brunberg 2003, Conde Silvestre 2007), que establece una conexión entre tipología textual y registro lingüístico. De esta manera, es posible establecer diferentes grados de formalidad inherentes a cada tipo. Algunos de estos son empleados por personas de todo el espectro de la sociedad alfabetizada, y así, las diferencias en el uso lingüístico cabe atribuirán intrínsecamente al nivel sociocultural del emisor. Es el caso de las cartas de petición que hemos examinado, sobre todo en el ámbito de la beneficiencia de Madrid capital. ALDICAM participa también de los métodos de la Dialectología, pues es posible elaborar mapas diacrónicos que muestren la diversidad de usos en un espacio geográfico determinado. Nos interesa especialmente el contraste entre los usos de la ciudad y el mundo rural para observar su interacción en diferentes épocas.

Algunos resultados llamativos se han obtenido en la documentación de archivos municipales, como los de El Escorial, Daganzo y Arganda del Rey, entre otros. Así, hemos observado que el laísimo parece arraigarse antes y con mayor intensidad en el noroeste y la capital frente al sur. En el léxico se observan algunas voces que apuntan a tener una distribución comarcal, como cacera / o advertisement (Sánchez-Prieto Borja y Vázquez Balonga, en prensa).

References
In recent years, the widespread availability of electronic corpora has created perfect opportunities for investigating the mutual association between various types of constructions and the lexemes collocating with them in naturally-occurring discourse, as shown by a number of studies (Hilpert 2008; Colleman 2010; Desagulier 2014; Wilinski 2015; Pedersen 2016). To date, however, little or no attention has been paid to both qualitative and quantitative analysis of the *be of N*-pattern, the quantification of nouns in this construction, the statistical corroboration of their occurrence in this kind of structure, or the empirical substantiation of previous hypotheses about their use. In addition, no single study exists which has attempted to postulate the occurrence of the *be of N*-construction in a constructional cline, and which has sought to determine semantic and discourse-functional properties of this construction. The primary aim of this paper is therefore twofold: first, to posit the existence of the *be of N*-construction (a pairing of form and meaning/function in its own right), and second, to carry out a qualitative and quantitative investigation of nouns occurring in this construction.

To this end, this paper employs a constructional approach to grammatical structure (Goldberg 2006), the theory of frame semantics (Fillmore, Lee-Goldman, and Rhodes 2012), and a quantitative corpus-based method that is specifically tailored for investigating the relationship between a grammatical construction and a lexical item occurring in one slot of this particular pattern. The method referred to as the attraction-reliance measure (Schmidt 2000; Schmid & Küchenhoff 2013) is used to gauge the reciprocal interaction between a noun and the *be of N*-pattern: in other words, to determine strongly attracted and repelled nouns of this construction. On the basis of the data extracted from the Corpus of Contemporary American English (COCA), the paper seeks to show that there are nouns that are more strongly attracted to the construction than others, and that the occurrence of particular nouns in this pattern is more or less significant than their use in other structures in COCA, i.e. that given nouns occur more or less frequently in this pattern than in other patterns in the corpus. In addition, the results of the analysis seem to suggest that the construction under scrutiny exhibits a strong preference for specific categories of nouns evoking different semantic frames, and that it carries multiple meanings and performs specific functions in discourse.

References


40. Yao, Gang. “El Aprendizaje de Vocabulario mediante el Data-driven Learning en un Contexto de Español como Lengua Extranjera”

Estudios antecedentes han señalado asociaciones potenciales entre los corpus y la adquisición de segunda lengua (Chambers 2010). Unos estudios, por ejemplo, intentaban investigar y demostrar el efecto del Data-driven Learning (DDL, Johns 1991) en un contexto del aprendizaje de una lengua extranjera (LE). Sin embargo, de momento son limitadas las investigaciones empíricas, especialmente las cuantitativas, sobre este enfoque, y son más escasas en un contexto de una lengua extranjera que no sea el inglés. Por tanto, los objetivos del presente estudio son, primero, investigar si el DDL, en comparación con el método tradicional, produce un efecto más positivo en el aprendizaje de vocabulario en el contexto de español como lengua extranjera (ELE); segundo, explorar actitudes de los aprendices sobre el DDL. Para alcanzar estos fines, se realizó primero un estudio experimental longitudinal con dos grupos apareados de universitarios chinos (grupo experimental, N = 16; grupo de control, N = 16) que llevaban tres o cuatro años estudiando español. El resultado del post-test inmediato indicó que el DDL era más efectivo que el método tradicional en cuanto al aprendizaje de vocabulario, (t(30) = 6.191, p < .001). En el post-test diferido, el dato cuantitativo reveló que el grupo de DDL también rendía significativamente mejor que el otro grupo (U = 60.000, z = -2.568, p = 0.01). Asimismo, una encuesta realizada a posteriori dentro del grupo experimental corrobora los buenos resultados de las pruebas, demostrando que los encuestados favorecían el DDL y adoptaban una actitud positiva sobre su futura aplicación en el aprendizaje de español, aunque el tiempo gastado y la paciencia seguían constituyendo un problema del enfoque de DDL. El presente trabajo constituye una base sólida para la teoría de DDL en el campo de la ASL, sobre todo en el área de ELE; al mismo tiempo, concluye discutiendo unas advertencias y sugiriendo posibles líneas futuras de investigación.

References
POSTERS


Although the concepts of transhumanism and posthumanism—or the belief that humanity can be enhanced and transformed by technological advances and evolve beyond its current limitations—seems recent, it is not so. In fact, as Holm (2016: 131) pointed out, the trope of a human being who is no longer human has been a recurrent one since the very inception of science fiction. This study presents an overview of how transhumanism and posthumanism appear in science fiction novels published between 1950 and 2016. Its aim is to examine some linguistic features relating to the fields of computer science and technology to assess the lexical richness of the texts, understood as the degree of variety of vocabulary used, as well as to make a diachronic distinction between the first occurrence of these terms and their respective concepts. To this end, a corpus of sixteen novels that go from Asimov’s I, Robot (1950) to Katz’s The Cybernetic Tea Shop (2016) has been created and used to see whether that manifestation—either directly or indirectly—is determined by their authors’ idiosyncrasies and context or not.

24. Ortiz, Rocío. “La alternancia de mucho/muncho y sus variantes en el español de la ciudad de Granada”

El objetivo de este trabajo es examinar la relación entre ciertas variables sociolínguísticas en la pronunciación del adverbio mu(n)chísimo (teniendo en cuenta las diferencias de género y número, así como la forma superlativa mu(n)chísimo) en español hablado en la ciudad de Granada (España) gracias a los materiales del Corpus PRESEEA (Proyecto de Estudio Sociolingüístico del español en España y América).

En este estudio se nos presentan dos posibles opciones: la pronunciación normativa y académicamente aceptada, que sería la forma sin n traba (much) o, por el contrario, la variante que presenta una n epentética al final de la primera sílaba (muncho). Aunque el DRAE marca la forma muncho como “en desuso”, parece que esto no es del todo cierto y que está bastante extendida en ciertas zonas y estratos sociales, lo cual no es sorprendente si consideramos que la adición de una -n- en ciertas secuencias fónicas es un viejo fenómeno, documentado en latín y característico del español clásico que aparece también en otras lenguas y dialectos romances.

Una vez realizados los tratamientos estadísticos estándares, verificamos, en primer lugar, que muncho y sus variantes no están en desuso en el español de la ciudad de Granada, sino que tienen un porcentaje de aparición del 11.5% que, aunque no es una cifra muy alta, no resulta para nada desestimable. En segundo lugar, verificamos que las variables que son significativas para la inserción de esta n epentética en la forma mucho son: 1) la instrucción del hablante, 2) el género del hablante, 3) la edad del hablante y 4) la categoría gramatical en la que es mu(n)chí se inscribe dependiendo de su función en la oración.

127. Cruz Ortiz, Rocío. “La alternancia de mucho/muncho y sus variantes en el español de la ciudad de Granada”

El objetivo de este trabajo es examinar la relación entre ciertas variables sociolínguísticas en la pronunciación del adverbio mu(n)chísimo (teniendo en cuenta las diferencias de género y número, así como la forma superlativa mu(n)chísimo) en español hablado en la ciudad de Granada (España) gracias a los materiales del Corpus PRESEEA (Proyecto de Estudio Sociolingüístico del español en España y América).

En este estudio se nos presentan dos posibles opciones: la pronunciación normativa y académicamente aceptada, que sería la forma sin n traba (much) o, por el contrario, la variante que presenta una n epentética al final de la primera sílaba (muncho). Aunque el DRAE marca la forma muncho como “en desuso”, parece que esto no es del todo cierto y que está bastante extendida en ciertas zonas y estratos sociales, lo cual no es sorprendente si consideramos que la adición de una -n- en ciertas secuencias fónicas es un viejo fenómeno, documentado en latín y característico del español clásico que aparece también en otras lenguas y dialectos romances.

Una vez realizados los tratamientos estadísticos estándares, verificamos, en primer lugar, que muncho y sus variantes no están en desuso en el español de la ciudad de Granada, sino que tiene un porcentaje de aparición del 11.5% que, aunque no es una cifra muy alta, no resulta para nada desestimable. En segundo lugar, verificamos que las variables que son significativas para la inserción de esta n epentética en la forma mucho son: 1) la instrucción del hablante, 2) el género del hablante, 3) la edad del hablante y 4) la categoría gramatical en la que es mu(n)chí se inscribe dependiendo de su función en la oración.


This paper aims to start a discussion on how the Global Anglicism Database (GLAD database), a contrastive database of lexical borrowings from English in world languages currently undergoing construction, can be utilized for the study of Anglicisms, specifically the Anglicisms in Japanese. The growing influence of the English language on languages around the world, or Anglicization, has been of interest to both the specialist and nonspecialist alike. The Japanese language is indubitably no exception to this trend and a considerable amount of scholarly research has been carried out on this topic. However, most of the present research has come about purely through the study of Japanese, independent from comparison with the trends in other languages. Understanding the Anglicisms of Japanese within the context of a global trend will provide a completely new perspective into the understanding of this issue. In one of the most significant contrastive studies on Anglicisms which included Japanese, Kowuer & Rosenhouse (2008) contemplated the psycholinguistic, sociolinguistic, and sociohistorical factors of the Anglicization process in twelve languages. However, due to the methodological constraints in studies like these and others, particularly the lack of a comparable list of Anglicisms, we are still left with a great deal of questions, especially regarding the differences in the quantity and types of borrowings as well as their phonological, morphological, semantic, and syntactic characteristics among world languages. The GLAD Network, established in 2015, is currently working on the creation of an Anglicism database that will contain an organized list of the Anglicisms of general use in many languages. It is expected to facilitate the efforts of researchers to conduct more effective contrastive studies to uncover the detailed aspects of Anglicisms. This paper describes the unique characteristics of the Anglicisms in Japanese according to the classified types of borrowings that will be recorded in the database: unadapted borrowing, adapted borrowing, semantic loan, loan translation, hybrid, pseudo-Anglicisms, and phonosemantic matching. Based on the descriptions, it will discuss how the GLAD database can be utilized to further uncover the characteristics of Anglicisms in Japanese. One such example Japanese that requires further examination is the large amount of pseudo-Anglicisms that have resulted from the semantic shift, clipping, and recombination of English words. Comparing the pseudo-Anglicism in Japanese with other languages both quantitatively and qualitatively, which will be made possible using the GLAD database, will reveal what is unique and what is not in this regard. Having a measurable set of data (such as overall number of direct borrowings, number of pseudo-Anglicisms, year of first attestation, etc.) from a number of languages will reveal the sociolinguistic factors that contribute to the characteristics. It is particularly important to answer how the vast geographical and genealogical distance, as well as the English level of speakers characterize the Anglicisms in Japanese. Completion of the GLAD database and contrastive studies utilizing the database will

References
help to answer these questions and to greatly advance the understanding of Anglicisms in Japanese and globally.

References

58. Monaco, Leida Mª, Luis Puente Castelo, Begoña Crespo and Isabel Moskowich. “CETeL under construction: Compiling languages for the Coruña Corpus”
The aim of this paper is to present a new subcorpus in the Coruña Corpus of English Scientific Writing (CC) which is currently under compilation. This is the case of CETeL, Corpus of English Texts on Languages.

The main idea behind the compilation of the CC has always been to gather scientific texts of the late Modern English period in several twin corpora sharing the same characteristics (Crespo and Moskowich, 2010; Moskowich 2016). We first delimited the time-span to be covered and determined the size of samples. Although there are many different proposals for the periodisation of English, we decided to set out time limits between 1700 and 1900 basing on extralinguistic factors. As for the size of the samples in the corpus, and after examining some of the texts from the period, we decided to collect 10,000-word extracts. When we made this decision, specialised corpora did not abound and not much information about them was available, except for Biber’s (1993) claim that 1,000-word samples were enough to observe variation in specialised registers. Our direct reading of the material as well as the fact that we did not intend to collect huge amounts of samples, convinced us otherwise. This, as well as our initial decision to use XML TEI for text and metadata files is now backed by other corpus compilers (VARIENG, 2016) that have partly adopted similar decisions.

Since the Coruña Corpus aims at compiling specialised texts, it was organised into different subcorpora, one for each discipline. Disciplines were considered according to the criteria of the historical moment when texts were published, that is, an inclusive one, since during the late Modern period, fields of knowledge were not as separated as they are nowadays. Nevertheless, the divisions proposed by UNESCO in 1988 were used as a starting point. This way, at the moment of writing this paper, several subcorpora are or have been published. The Corpus of English Texts on Astronomy (CETA) and the Corpus of English Philosophy Texts (CEPHIT) have been already published. The Corpus of History English Texts (CHET), Corpus of English Chemistry Texts (CECHET) and Corpus of English Life Sciences Texts (CELIST) and the Corpus of English Texts on Languages (CETeL) are all at different stages of preparation.

A metadata file was prepared in order to accompany each sample. Such metadata files provide information about both the author and the text itself, and can be used to narrow searches according to extralinguistic parameters (sex, age or geographical provenance of the author as well as date of publication, genre of the sample, etc.) with the Coruña Corpus Tool. This is part of the information which will be presented as it describes the content of CETeL.

As different disciplines require different taxonomies, the compilation process also entails the classification of samples into different genres. The categorisation of the samples in early Linguistics will be another issue to be tackled in this presentation so as to offer a complete picture of this new subcorpus.

References

60. Papadopoulou, Rania. “Idiomatic multiword expressions in Modern Greek as cases of loan translation: A web-based corpus analysis”
Over the past decades, the number of English borrowings into Modern Greek (MG) has markedly increased, because of the globalization of computer-mediated communication. English loanwords into MG can be: (a) direct loanwords (kobiúter < Engl. computer), (b) semantic loanwords (podíki < Engl. mouse), or (c) loan translations (uranokíssis < Engl. skyscraper) (Anastasiadi-Simeonidi, 1994).

Recently, however, cases of lexical borrowing have been noticed in MG, mainly in the language of the youth. They constitute instances of word by word translation or calquing. These cases, almost exclusively, belong to the category of idiomatic multiword expressions (MWEs): lexical items that cross word boundaries and present some kind of idiomaticity (Baldwin, 2015; Baldwin & Kim, 2010), as in the following examples:

(1) πέρπνo to xróνo mu
  take the time my
  ‘take my time’

(2) κλεο páno apo to xínéno yála
  cry over from the spilt milk
  ‘cry over spilt milk’

In previous works on the subject, we collected idiomatic MWEs attested in MG that appear to be word by word translations of the English equivalents, through the Internet, the television, the radio, the Press and everyday colloquial speech. We tried to check the appearance of our data through the existent MG corpora e.g. the Hellenic National Corpus (HNC; http://hnc.lispi.gr) and the Corpus of Greek Texts (CGT; http://www.sek.edu.gr). However, we could not locate our data due to the fact that these corpora are not fully up to date and they do not have a high degree of representativeness since they do not cover all genres of MG texts (e.g. sublanguage texts). In order to be able to get significant results from a corpus of MG texts we plan to build a webbased corpus of MG texts through the platform Sketch Engine.
dated from the 1980’s to date, that will include different kinds of texts and text genres, i.e. formal or informal texts such as from newspapers, blogs, webpages and social networks. The texts will cover a variety of subjects too, such as sports, fashion, lifestyle, as well as texts from different periods of time. Via Sketch Engine we will:

- Be able to build an extended web-based corpus of MG texts, by downloading a large number of elements taken from the web.
- Be able to import our own texts in the platform.
- Have access to the already existent corpus for MG on Sketch Engine, the GkWaC but also to the HNC and the CGT.
- Be able to carry out a diachronic investigation of the appearance and spread of a variety of frequent MWEs of English origin.
- Have the opportunity to obtain statistical information for the data under investigation.
- Be able to study the onomasiological competition between the new anglicized form and the original native form.

References

34. Torruella, Joan. “Diseño y elaboración de un corpus documental para el estudio de las interferencias lingüísticas entre el español y el catalán en los siglos XVIII y XIX”

El póster que queremos presentar se enmarca dentro del panel temático número 1: Diseño, elaboración y tipología de corpus. Se trata de mostrar la motivación y la estructura de un nuevo corpus diseñado para obtener datos destinados al estudio de la lengua española en Catalunya durante los siglos XVIII y XIX, especialmente de las interferencias que el estrecho contacto de esta lengua con la lengua catalana producía. Con este corpus, se pretende sentar las bases para el estudio histórico del contacto lingüístico entre el español y el catalán. El proyecto se lleva a cabo en el Grupo de Lexicografía y Diacronía de la Universidad Autónoma de Barcelona.

Al tratarse de una finalidad tan concreta, ello ha obligado a estructurar el corpus de una manera específica y a desarrollar un sistema de marcoje muy determinado. En el poster, por un lado, se dará cuenta de cómo se proyecta crear una colección de documentos inéditos escritos en español que permita estudiar la evolución de esta lengua y sus interferencias con la lengua catalana. En este apartado, es importante la relación de los tipos de documentos que se van a utilizar para que estos reflejen los diferentes niveles diastráticos (niveles socioculturales) y diafásicos (distintos registros) del uso de la lengua. Por otro lado, se mostrará cómo se estructura el programa informático que va a gestionar el corpus textual, así como el marcoje que se va a utilizar para evidenciar los distintos tipos de interferencias (fonéticas, léxicas, discursivas, etc.).